Introduction to Algebra

Rui L. Fernandes
University of Illinois at Urbana-Champaign

Manuel Ricou
Instituto Superior Técnico

April 14, 2017
## Contents

1. **Basic Algebraic Structures**  
   1.1 Introduction .......................................................... 7  
   1.2 Groups ........................................................................ 13  
   1.3 Permutations ............................................................... 19  
   1.4 Homomorphisms and Isomorphisms ................................. 24  
   1.5 Rings, Integral Domains and Fields ................................. 33  
   1.6 Homomorphisms and Isomorphisms of Rings .................... 42  
   1.7 The Quaternions ............................................................ 50  
   1.8 Symmetries .................................................................... 54

2. **The Integers**  
   2.1 The Axioms of the Integers ............................................ 65  
   2.2 Inequalities and Ordered Rings ...................................... 70  
   2.3 Principle of Induction .................................................... 77  
   2.4 Summations and Products ............................................. 83  
   2.5 Factors, Multiples and Division ..................................... 88  
   2.6 Ideals and Euclides Algorithm ...................................... 93  
   2.7 The Fundamental Theorem of Arithmetic ....................... 103  
   2.8 Congruences ................................................................. 109  
   2.9 Prime Factorizations and Cryptography ......................... 119

3. **Other Examples of Rings**  
   3.1 The Rings \(\mathbb{Z}_m\) .................................................... 125  
   3.2 Fractions and Rational Numbers .................................... 137  
   3.3 Polynomials and Power Series ..................................... 142  
   3.4 Polynomial Functions .................................................. 150  
   3.5 Division of Polynomials ............................................... 156  
   3.6 The Ideals of \(K[x]\) ....................................................... 164  
   3.7 Divisibility and Prime Factorization ............................... 168
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td>Factorization in $D[x]$</td>
<td>182</td>
</tr>
<tr>
<td>4</td>
<td>Quotients and Isomorphisms</td>
<td>189</td>
</tr>
<tr>
<td>4.1</td>
<td>Groups and Equivalence Relations</td>
<td>189</td>
</tr>
<tr>
<td>4.2</td>
<td>Quotient Group and Quotient Ring</td>
<td>197</td>
</tr>
<tr>
<td>4.3</td>
<td>Real and Complex Numbers</td>
<td>205</td>
</tr>
<tr>
<td>4.4</td>
<td>Isomorphism Theorems for Groups</td>
<td>213</td>
</tr>
<tr>
<td>4.5</td>
<td>Isomorphism Theorems for Rings</td>
<td>221</td>
</tr>
<tr>
<td>4.6</td>
<td>Free Groups, Generators and Relations</td>
<td>228</td>
</tr>
<tr>
<td>5</td>
<td>Finite Groups</td>
<td>245</td>
</tr>
<tr>
<td>5.1</td>
<td>Groups of Transformations</td>
<td>245</td>
</tr>
<tr>
<td>5.2</td>
<td>The Sylow Theorems</td>
<td>251</td>
</tr>
<tr>
<td>5.3</td>
<td>Nilpotent and Solvable Groups</td>
<td>259</td>
</tr>
<tr>
<td>5.4</td>
<td>Simple Groups</td>
<td>265</td>
</tr>
<tr>
<td>5.5</td>
<td>Symmetry Groups</td>
<td>270</td>
</tr>
<tr>
<td>6</td>
<td>Modules</td>
<td>281</td>
</tr>
<tr>
<td>6.1</td>
<td>Modules over a ring</td>
<td>281</td>
</tr>
<tr>
<td>6.2</td>
<td>Linear Independence</td>
<td>290</td>
</tr>
<tr>
<td>6.3</td>
<td>Tensor Products</td>
<td>295</td>
</tr>
<tr>
<td>6.4</td>
<td>Modules over Integral Domains</td>
<td>304</td>
</tr>
<tr>
<td>6.5</td>
<td>Finitely Generated Modules over a PID</td>
<td>310</td>
</tr>
<tr>
<td>6.6</td>
<td>Classifications</td>
<td>320</td>
</tr>
<tr>
<td>6.7</td>
<td>Categories and Functors</td>
<td>326</td>
</tr>
<tr>
<td>7</td>
<td>Galois Theory</td>
<td>333</td>
</tr>
<tr>
<td>7.1</td>
<td>Field Extensions</td>
<td>335</td>
</tr>
<tr>
<td>7.2</td>
<td>Compass-and-Straightedge Constructions</td>
<td>338</td>
</tr>
<tr>
<td>7.3</td>
<td>Splitting Fields</td>
<td>343</td>
</tr>
<tr>
<td>7.4</td>
<td>Homomorphisms of Extensions</td>
<td>350</td>
</tr>
<tr>
<td>7.5</td>
<td>Separability</td>
<td>354</td>
</tr>
<tr>
<td>7.6</td>
<td>The Galois Group</td>
<td>358</td>
</tr>
<tr>
<td>7.7</td>
<td>The Galois Correspondence</td>
<td>364</td>
</tr>
<tr>
<td>7.8</td>
<td>Applications</td>
<td>370</td>
</tr>
<tr>
<td>8</td>
<td>Commutative Algebra (not yet available)</td>
<td>381</td>
</tr>
<tr>
<td>A Set Theory</td>
<td>383</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>A.1 Relations and Functions</td>
<td>383</td>
<td></td>
</tr>
<tr>
<td>A.2 Axiom of Choice, Zorn's Lemma and Induction</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>A.3 Finite Sets</td>
<td>396</td>
<td></td>
</tr>
<tr>
<td>A.4 Infinite Sets</td>
<td>401</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

Basic Algebraic Structures

1.1 Introduction

Algebra is still today, as it always has been, the study of operations, rules and methods to solve equations. The origin of the term “Algebra” itself is specially enlightening. According to B. L. van der Waerden, one of the outstanding algebraists of the XX Century, the term was used for the first time by in the IX Century the Arab mathematician al-Khowarizmi, in the title of a monograph which aimed at presenting mathematical knowledge of practical application. The Arab word al-jabr is used in this monograph to denote two basic procedures in solving equations, namely:

1. adding the same positive quantity to both sides of an equation, in order to cancel negative quantities, and

2. multiplying both sides of the equation by the same positive quantity in order to cancel fractions.

The aforementioned monograph discusses problems of a very different nature, ranging from geometric problems to the solution of linear and quadratic equations, and includes also a variety of applications to astronomy, to commerce, to calendars, etc. With time, the term al-jabr, or Algebra, became associated with the general area of knowledge that deals with operations and numerical equations.

---

1 The title of the book was “Al-jabr wa’l muqābala”. His author, Mohammed al-Khowarizmi was an astronomer and mathematician, member of the Bait al-hikma (“House of Wisdom”) established in Baghdad, the center of scientific knowledge in those times. The name al-Khowarizmi is also the origin of another common term: algorism!
Nowadays, one sometimes distinguishes *Classic* and *Modern* Algebra. Such expressions are not very fortunate, and according the the Italian mathematician F. Severi today’s *Modern Algebra* will soon become the *Classic Algebra* of tomorrow. Actually, when one compares, e.g, the XXI Century Algebra with the XVI Century Algebra, and one eliminates differences common to all branches of scientific knowledge (rigor, formalism, notations, the amount of knowledge), the only remaining difference is the generality in which problems are stated and treated.

Actually, a tendency to generalization has always been present in Algebra. Initially, this was the case with the successive generalizations of the concept of number: first from natural to positive rational, then to negative numbers, irrational numbers and complex numbers. In the XIX Century, mathematicians realized that the so-called “algebraic ideas” could be applied to objects which are not numbers, such as vectors, matrices and transformations.

The initial slow expansion of the scope of Algebra turned into a rapid explosion when it became clear that one could study properties of any algebraic operation without specifying the nature of the objects on which the operation applies or how to calculate the result of the operation. In fact, this study can be performed simply by postulating (*i.e.*, assuming as hypothesis) a certain numbers of basic algebraic properties of the operation, such as commutativity or associativity. Algebra then finally became *axiomatic* (although with a lag of more than 2000 years relative to Geometry). This was perhaps the most significative achievement of recent times, and justifies the use of the name “Abstract Algebra”, to which nowadays we often refer to Algebra.

The axiomatization of Algebra required, first of all, precise definitions of *abstract algebraic structures*. In its simplest form, an abstract algebraic structure is formed by a non-empty set $X$, called the *support* of the structure, and a *binary operation* on $X$, which is just a map $\mu : X \times X \to X$. Distinct sets of hypothesis, or axioms, imposed on the operation $\mu$ lead to distinct abstract algebraic structures. Note, however, that the definitions do not include any assumption on the nature of elements of $X$, nor on the procedure to evaluate the operation $\mu$.

There a few conventions that everyone follows. If $\mu : X \times X \to X$ is a binary operation on $X$, one chooses a symbol such as “$+$” or “$*$” to represent it, and writes “$x + y$” or “$x * y$” instead of “$\mu(x, y)$”. Actually, one can even omit the symbol and represent the binary operation by simply juxtaposition, *i.e.*, “$xy$” instead of “$\mu(x, y)$”. The use of such notations as “$x + y$” and “$xy$” does not mean that these symbols represent or bear any
relation with the usual addition and multiplication operations on numbers. In this respect, the only generally accepted convention is that one reserves the symbol “+” to denote commutative operations, i.e., operations such that \( \mu(x, y) = \mu(y, x) \). Whenever we represent a commutative operation by the symbol “+” we will say that we use the additive notation, while in all other cases we say that we use the multiplicative notation. Also, we will make systematic use of the usual conventions with parenthesis, so that we write

\[ x \ast (y \ast z) = \mu(x, \mu(y, z)), \]

which, in general, will be distinct from

\[ (x \ast y) \ast z = \mu(\mu(x, y), z). \]

As a general rule, imposing few axioms on an algebraic structure lead to results which apply in great generality, for they will be valid for many concrete algebraic structure. However, very general results tend to be not so interesting, precisely because they rely on a small number of hypothesis. On the other hand, if one starts with a richer set of axioms, one obtains in principle more interesting results, but less general, since fewer concrete algebraic structures will satisfy all the axioms. One of the main issues in Abstract Algebra is precisely to determine a set of axioms (i.e., definitions of abstract algebraic structures) which are sufficiently general to include many concrete examples and, at the same time, sufficiently rich to lead to interesting results.

Let us consider a few important examples illustrating these remarks. Without any additional assumption on a binary operation \( \ast \), we can introduce the notion of identity element, inspired on the properties of the integers 0 and 1 relative to the usual addition and multiplication, respectively.

**Definition 1.1.1.** Let \( \ast \) be a binary operation on a set \( X \). An element \( e \in X \) is called an identity element for \( \ast \) if \( x \ast e = e \ast x = x \) for all \( x \in X \).

It is possible to prove a general result about identity elements which is valid for any binary operation.

**Proposition 1.1.2.** Every binary operation has at most one identity element.

**Proof.** Assume that \( e \) and \( e' \) are both identity elements for the operation \( \ast \). We then have:

\[ e \ast e' = e \quad \text{(since e is an identity element)}, \]
\[ e \ast e' = e \quad \text{(since e' is an identity element)}. \]

We conclude that \( e = e' \). \( \square \)
Hence, we can talk without ambiguity about the identity element, whenever this element exists. Also, it is common to use the terms “zero” and “one” (the last one usually called “identity”) to denote the identity element of *. Obviously, we should be consistent: we should reserve the use of the term “zero” (and also the symbol “0”) for the additive notation, and use the term “identity” (and also the symbols “1” or “I”) for the multiplicative notation.

Whenever a binary operation * has an identity e, we can introduce the notion of inverse element. The definition is as follows:

**Definition 1.1.3.** Given a binary operation * with identity e, an element x ∈ X is called invertible if there exists y ∈ X such that

\[ x \ast y = y \ast x = e. \]

In this case, we call y an inverse of x.

Again, to be consistent with common usage, if we use the additive notation we call an inverse element of x a symmetric of x.

Note that y is an inverse of x if and only if x is an inverse of y, i.e., in other words “being inverse of” is a symmetric relation. If one only has \( x \ast y = e \), we say that y is a right inverse of x and that x is a left inverse of y. Obviously, y is an inverse of x if and only if y is both a right and left inverse of x. However, a right inverse may fail to be a left inverse, and vice-versa. Still, when * is an associative operation we can show the following.

**Proposition 1.1.4.** Let * be an associative binary operation on X with identity e. If \( x \in X \) has a right inverse y and a left inverse z, then y = z and x is invertible.

**Proof.** Assume that y, z ∈ A satisfy \( x \ast y = z \ast x = e \). Then:

\[
\begin{align*}
  x \ast y &= e \\
  \Rightarrow z \ast (x \ast y) &= z \\
  \Rightarrow (z \ast x) \ast y &= z \\
  \Rightarrow e \ast y &= z \\
  \Rightarrow y &= z
\end{align*}
\]

(because \( z \ast e = z \)),

(because \( \ast \) is associative),

(because \( z \ast x = e \)),

(because \( e \ast y = y \)).

The previous results apply to any binary operation that satisfies the two properties used in the proofs (existence of identity and associativity). These are exactly the assumptions used as axioms in the following definition:
1.1. INTRODUCTION

Definition 1.1.5. An algebraic structure \((X,\ast)\) is called a monoid if it satisfies the following properties:

(i) The operation \(\ast\) has identity \(e\) on \(X\).

(ii) The operation \(\ast\) is associative, i.e., \((x \ast y) \ast z = x \ast (y \ast z)\), for all \(x,y,z \in X\).

If \(\ast\) is a commutative operation, i.e., if \(x \ast y = y \ast x\) for all \(x,y \in X\), we say that the monoid is abelian\(^2\). If the operation is commutative and we use the additive notation, we say that the monoid is additive.

From Proposition 1.1.4 we conclude:

Proposition 1.1.6. If \((X,\ast)\) is a monoid (with identity \(e\)), and \(x \in X\) is invertible, there exists a unique element \(y \in X\) such that \(x \ast y = y \ast x = e\).

Some well-known operations furnish examples of monoids.

Examples 1.1.7.

1. The set of all \(n \times n\) matrices with real entries with the usual product of matrices is a monoid. Hence, if \(A,B,C\) are matrices \(n \times n\), \(I\) is the identity matrix and \(AB = CA = I\), then \(B = C\) and the matrix \(A\) is invertible.

2. The set \(\mathbb{R}\) of all functions \(f : \mathbb{R} \to \mathbb{R}\) with “composition product”, defined by \((f \circ g)(x) = f(g(x))\) is a monoid. The identity is the function \(I : \mathbb{R} \to \mathbb{R}\) given by \(I(x) = x\). Hence, if there exist functions \(g,h : \mathbb{R} \to \mathbb{R}\) such that \(f \circ g = h \circ f = I\), then \(g = h\) and \(f\) is invertible (i.e., is a bijection).

3. The set \(\mathbb{R}\) of all real numbers with the usual addition is an additive monoid. In this case any element is invertible.

4. The set \(\mathbb{R}^+\) of all positive real numbers with the usual product is a monoid. Again the operation is commutative and every such element is invertible.

If an element \(x\) in a monoid \((X,\ast)\) is invertible, then the inverse of \(x\) is unique. If we use the multiplicative notation, then we denote by “\(x^{-1}\)” the inverse of \(x\), while in the additive notation we denote by “\(-x\)” the symmetric of \(x\). With these conventions, some of the well-known results for symmetrics and inverses of real numbers, actually hold for any monoid. We leave the proofs of the following results as exercises:

\(^2\) In honour of Niels Henrik Abel (1802–1829), a Norwegian mathematician who is consider to be one of the founders of Modern Algebra

\(^3\) If \(X\) and \(Y\) are sets, \(Y^X\) denotes the set of all functions \(f : X \to Y\) (See Definition A.2.4 in the Appendix).
Proposition 1.1.8. If \((X, \ast)\) is a monoid, and \(x, y \in X\) are invertible, then \(x^{-1}\) and \(xy\) are invertible, and we have:

\[
(x^{-1})^{-1} = x, \quad \text{and} \quad (xy)^{-1} = y^{-1}x^{-1}.
\]

For an additive monoid, we have:

\[
-(−x) = x, \quad \text{and} \quad -(x + y) = (-x) + (-y).
\]

Exercises.

1. Let \(X = \{x, y\}\) be a set with two elements. How many binary operations are there on \(X\)? How many of these operations are (i) commutative, (ii) associative, (iii) have an identity?

2. How many binary operations are there in a set with 10 elements?

3. In \((\mathbb{Z}, −)\) is there an identity? Inverses? Is this operation associative?

4. Let \(\mathbb{R}^R\) be the set of all functions mentioned in Example 1.1.7.2 and assume that \(f \in \mathbb{R}^R\).

   (a) Show that there exists \(g \in \mathbb{R}^R\) such that \(f \circ g = I\) if and only if \(f\) is surjective.
   
   (b) Show that there exists \(g \in \mathbb{R}^R\) such that \(g \circ f = I\) if and only if \(f\) is injective.
   
   (c) If \(f \circ g = f \circ h = I\), is true that \(g = h\)?

5. Prove Proposition 1.1.8.

6. Let \(\ast\) be binary operation on \(X\), and \(x, y \in X\). If \(n \in \mathbb{N}\) is a natural number, define the power \(x^n\) by induction as follows: \(x^1 = x\) and, for \(n \geq 1\), \(x^{n+1} = x^n \ast x\). Assuming that \(\ast\) is associative, shows that:

   (a) \(x^n \ast x^m = x^{n+m}\), and \((x^n)^m = x^{nm}\), for all \(n, m \in \mathbb{N}\).
   
   (b) \(x^n \ast y^n = (x \ast y)^n\), for all \(n \in \mathbb{N}\), if \(x \ast y = y \ast x\).

   How would you express these results in the additive notation?

7. Assume that \((X, \ast)\) is a monoid with identity \(e\), and \(x \in X\) is invertible. In this case, we define for \(n \in \mathbb{N}\) \(x^{-n} = (x^{-1})^n\), \(x^0 = e\). Show that the identities (a) and (b) in the previous problem are valid for all \(n, m \in \mathbb{Z}\).
1.2 Groups

The examples we saw in the previous section show that in an arbitrary monoid not every element is invertible. The monoids where every element has an inverse correspond to the one of the most important abstract algebraic structures.

Definition 1.2.1. A monoid \((G, \ast)\) is called a **GROUP** if all the elements in \(G\) are invertible. A group is called **ABELIAN** if the operation \(\ast\) is commutative.

The following simple examples already show how general the concept of group is.

**Examples 1.2.2.**

1. \((\mathbb{R}, +)\) is an abelian group.
2. \((\mathbb{R}^+, \cdot)\) is also an abelian group.
3. \((\mathbb{R}^n, +)\), where + denotes sum of vectors, is an abelian group.
4. The set of all \(n \times n\) invertible matrices with real entries is a non-abelian group with the usual product of matrices. This group is called the General Linear Group and denoted by \(GL(n, \mathbb{R})\).
5. The complex numbers \(z \in \mathbb{C}\) with \(|z| = 1\) (the unit circle) form an abelian group for complex multiplication, usually denoted by \(S^1\).
6. The complex numbers \(\{1, -1, i, -i\}\), with the complex multiplication form, a finite abelian group.
7. The set of all functions \(f : \mathbb{R} \to \mathbb{R}\) form an abelian group for the usual addition of functions. We can also consider special classes of functions, such as continuous functions, differentiable functions or integrable functions. All such classes give examples of abelian groups.
8. If \(X\) is any set and \((G, +)\) is an abelian group, then the set of all functions \(f : X \to G\) form an abelian group with operation “+” defined by

\[(f + g)(x) = f(x) + g(x), \quad \forall x \in X.\]

(Note that in this expression the symbol “+” is used with two different meanings!)
9. More generally, if \(X\) is a set and \((G, \ast)\) is a group, then the functions \(f : X \to G\) form a group with operation “∗” defined by

\[(f \ast g)(x) = f(x) \ast g(x), \quad \forall x \in X.\]

(Again, in this expression the symbol “∗” is used with two different meanings.)
The fourth example above illustrates a general fact: the invertible elements in an arbitrary monoid always form a group.

**Proposition 1.2.3.** Let \((X,*)\) be a monoid and \(G\) the set of invertible elements in \(X\). Then \(G\) is closed for the operation \(*\), and \((G,*)\) is a group.

**Proof.** The identity \(e\) of \(X\) is invertible with \(e^{-1} = e\), since \(e*e = e\). Hence, \(G\) is non-empty and contain the identity of \(X\).

Proposition 1.1.8 shows that \(x \in G \implies x^{-1} \in G\) (because \((x^{-1})^{-1} = x\)), and also that \(x, y \in G \implies x*y \in G\) (because \((x*y)^{-1} = y^{-1} * x^{-1}\)).

Since the operation \(*\) is associative (in the original monoid), \((G,*)\) is a group.

It is not our aim at this point to discuss in depth the theory of groups. For now, we limit ourselves to some elementary results which will be useful later in the study of many other algebraic structures.

**Proposition 1.2.4.** If \((G,*)\) is group with identity element \(e\), then:

(i) **Cancellation Law:** if \(g_1, g_2, h \in G\) and \(g_1*h = g_2*h\) or \(h*g_1 = h*g_2\), then \(g_1 = g_2\);

(ii) In particular, if \(g*g = g\) then \(g = e\);

(iii) The equation \(g*x = h\) (respectively, \(x*g = h\)) has the unique solution \(x = g^{-1} * h\) (respectively, \(x = h * g^{-1}\)).

**Proof.** We have:

\[
g_1 * h = g_2 * h \implies (g_1 * h) * h^{-1} = (g_2 * h) * h^{-1} \quad \text{because \(h\) is invertible),}
\]

\[
\implies g_1 * (h * h^{-1}) = g_2 * (h * h^{-1}) \quad \text{by associativity),}
\]

\[
\implies g_1 * e = g_2 * e \quad \text{(because \(h * h^{-1} = e\)),}
\]

\[
\implies g_1 = g_2 \quad \text{(because \(e\) is the identity).}
\]

---

Note that the results stated in this theorem are just “sophisticated” versions of the *al-jabr* operations mentioned in the introductory section.
The proof for \( h \ast g_1 = h \ast g_2 \) is similar, hence (i) holds. On the other hand,
\[
g \ast g = g \implies g \ast g = g \ast e \quad \text{(because } g \ast e = g),
\]
\[
\implies g = e \quad \text{(by the previous result)}.
\]
so (ii) holds. The proof of (iii) is left as an exercise. \( \square \)

If \((G, \ast)\) is a group and \(H \subset G\) is a non-empty set, it is possible that \(H\) is closed relative to the operation \(\ast\), i.e., that \(h_1 \ast h_2 \in H\), whenever \(h_1, h_2 \in H\). In this case, the operation \(\ast\) is a binary operation on \(H\), and we can ask under what conditions is \((H, \ast)\) a group. When this is the case, we say that \((H, \ast)\) is a SUBGROUP of \((G, \ast)\).

The next result gives simple criteria to decide if a given subset \(H\) of a group \(G\) is a subgroup.

**Proposition 1.2.5.** If \((G, \ast)\) is a group with identity element \(e\), and \(H \subset G\) is non-empty, then the following are equivalent:

(i) \((H, \ast)\) is a subgroup of \((G, \ast)\);

(ii) for all \(h_1, h_2 \in H\), \(h_1 \ast h_2 \in H\) and \(h_1^{-1} \in H\);

(iii) for all \(h_1, h_2 \in H\), \(h_1 \ast h_2^{-1} \in H\).

**Proof.** We will show the implications (i) \(\Rightarrow\) (ii) \(\Rightarrow\) (iii) \(\Rightarrow\) (i).

Assume that (i) holds. Then, by definition, \(H\) is closed for multiplication: if \(h_1, h_2 \in H\) then \(h_1 \ast h_2 \in H\). By assumption, \(H\) has an identity element \(\tilde{e}\), so that \(\tilde{e} \ast \tilde{e} = \tilde{e}\). By Proposition 1.2.4 (ii) we conclude that \(\tilde{e} = e\), so \(H\) contains the identity of \(G\). If \(h \in H\), consider the equation \(h \ast x = e\). According to Proposition 1.2.4 (iii), this equation has a unique solution \(x \in H\), which is also a solution of the same equation in \(G\). Hence, we must have \(x = h^{-1}\) (the inverse of \(h\) in the original group \(G\)). This shows that if \(h \in H\), then \(h^{-1} \in H\).

If (ii) holds, then clearly (iii) holds.

Finally, assume that (iii) holds. We must show that \((H, \ast)\) is a group. Since \(H\) is non-empty, we can choose \(h \in H\), and observe that \(h \ast h^{-1} = e \in H\). Hence, \(H\) contains the identity of \(G\). Similarly, if \(h \in H\), then \(e \ast h^{-1} = h^{-1} \in H\), so \(H\) contains the inverses (in \(G\)) of all its elements. It follows also that \(H\) is closed relative to \(\ast\), for if \(h_1, h_2 \in H\) then we already know that \(h_2^{-1} \in H\), so that \(h_1 \ast (h_2^{-1})^{-1} = h_1 \ast h_2 \in H\). Finally, \(\ast\) is associative in \(H\) because it was already associative in \(G\). \( \square \)
In any abelian group, using the additive notation, the difference $h_1 - h_2$ is defined by $h_1 - h_2 = h_1 + (-h_2)$ Hence, in additive notation, the condition “$h_1 * h_2^{-1} \in H$” is written as “$h_1 - h_2 \in H$.”

Examples 1.2.6.

1. Consider the group $(\mathbb{R}, +)$ and the subset of all integers $\mathbb{Z} \subset \mathbb{R}$. This subset is non-empty and the difference of two integers is still an integer, so $(\mathbb{Z}, +)$ is a subgroup of $(\mathbb{R}, +)$.

2. Still in the same group $(\mathbb{R}, +)$, consider the subset of natural numbers $\mathbb{N} \subset \mathbb{R}$. The difference of two natural numbers may fail to be a natural number, so $(\mathbb{N}, +)$ is not a subgroup of $(\mathbb{R}, +)$. Note that the sum of two natural numbers is still a natural number, so the sum is a binary operation on the set of natural numbers.

Let $(G, *)$ and $(H, \cdot)$ be two groups, and consider the cartesian product $G \times H = \{(g, h) : h \in G, h \in H\}$.

In $G \times H$ we define the binary operation

\[(1.2.1) \quad (g_1, h_1) \circ (g_2, h_2) = (g_1 * g_2, h_1 \cdot h_2).\]

We leave as an exercise to check that this algebraic structure is a group. It is called the direct product of the groups $G$ and $H$. Note that $G$ and $H$ can be seen as subgroups of $G \times H$ if we identify them with $G \times \{e\} = \{(g, e) : g \in G\}$ and $\{e\} \times H = \{(e, h) : h \in H\}$.

If $G$ and $H$ are abelian groups and the additive notation is used, then we will write $G \oplus H$ instead of $G \times H$, and we call this group the direct sum of $G$ and $H$.

The notion of direct product or direct sum of groups extends in a more or less straightforward way to any finite number of groups. For example, if $G$, $H$, and $K$ are groups, the direct product $G \times H \times K$ can be defined as $G \times H \times K = (G \times H) \times K$. More generally, given groups $G_1, G_2, \cdots, G_n$, we have:

$$\prod_{k=1}^{1} G_k = G_1, \text{ and } \prod_{k=1}^{n} G_k = \left(\prod_{k=1}^{n-1} G_k\right) \times G_n.$$
1.2. GROUPS

Example 1.2.7.

We leave it as an exercise to check that the direct sum of $(\mathbb{R}, +)$ with itself is $(\mathbb{R}^2, +)$. More generally, we have:

$$\mathbb{R}^n = \bigoplus_{k=1}^{n} \mathbb{R} = \mathbb{R} \oplus \cdots \oplus \mathbb{R}.$$ 

We can also consider the group $(\mathbb{Z}, +)$ and take the direct sum of this group with itself any finite number of times. The resulting group is denoted by

$$\mathbb{Z}^n = \bigoplus_{k=1}^{n} \mathbb{Z} = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}.$$ 

This group, as we shall see in Chapter 4, is the so-called free abelian group in $n$ symbols. It is a subgroup of the group $(\mathbb{R}^n, +)$.

Exercises.

1. Show that the sets $G = \{0, 1\}$ and $H = \{1, -1\}$ with the operations given by the following tables are groups.\(^6\)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\times$</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

2. Same as the previous exercise, but now with the sets $G = \{0, 1, 2\}$ and $H = \{1, x, x^2\}$, with operations given by the following tables.\(^7\)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>x</th>
<th>$x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\times$</td>
<td>1</td>
<td>x</td>
<td>$x^2$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>x</td>
<td>$x^2$</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>1</td>
<td>$x^2$</td>
</tr>
<tr>
<td>$x^2$</td>
<td>$x^2$</td>
<td>1</td>
<td>x</td>
</tr>
</tbody>
</table>

3. Complete the proof of Proposition 1.2.4 (iii).

4. Give an example showing that the cancelation law, in general, does not hold for monoids.

5. Let $(G, \ast)$ be a group. Show that the map $f : G \to G$, given by $f(x) = x^{-1}$, is a bijection of $G$ with $G$.

\(^6\)The group $G$ is usually denoted by $(\mathbb{Z}_2, +)$, for reasons to be explained later. The group $H$ is formed by the square roots of the unit.

\(^7\)The group $G$ is usually denoted by $(\mathbb{Z}_3, +)$. The group $H$ is formed by the cubic roots of the unit, since we can assume, e.g., that $x$ is the complex number $e^{\frac{2\pi i}{3}}$. 
6. How can Propositions 1.2.4 and 1.2.5 be stated in additive notation?

7. If \((G, \ast)\) is a group, its **center** is defined by \(C(G) = \{x \in G : g \ast x = x \ast g, \forall g \in G\}\). Show that \((C(G), \ast)\) is an abelian subgroup of \(G\). Determine the center of \(G = GL(n, \mathbb{R})\).

8. Let \((G, \ast)\) be a group, \(H_1, H_2\) subgroups of \(G\). Show that \(H_1 \cap H_2\) is a subgroup of \(G\).

9. Show that a group \((G, \ast)\) is abelian if and only if \((g \ast h)^2 = g^2 \ast h^2\), for all \(g, h \in G\).

10. Let \(\ast\) be an associative binary operation in a set \(G\), which satisfies:

    (i) there exists \(e \in G\) such that, for all \(g \in G\), \(g \ast e = g\) (right identity).

    (ii) For all \(g \in G\) there exists \(g' \in G\) such that \(g \ast g' = e\) (right inverses).

Show that:

(a) \((G, \ast)\) is a group.

(Hint: show first that \(g \ast g = g \Rightarrow g = e\)).

(b) Fix some set \(X\) and let \(G\) be the class of all surjective maps \(f : X \rightarrow X\), with \(\ast\) the operation of composition of maps. Is this a group? Why doesn’t this example contradicts (a)?

11. Let \(\ast\) be an associative binary operation on a non-empty set \(G\), which satisfies:

    (i) The equation \(g \ast x = h\) has a solution in \(G\) for all \(g, h \in G\).

    (ii) The equation \(x \ast g = h\) has a solution in \(G\) for all \(g, h \in G\).

Show that \((G, \ast)\) is a group.

(Hint: show first that if \(g \ast e_1 = g\) and \(e_2 \ast h = h\), for some \(g\) and \(h\), then one must have \(e_1 = e_2 = e\) and \(e\) is the identity element for \(\ast\)).

12. Let \((G, \ast)\) and \((H, \cdot)\) be two groups. Show that the binary operation on \(G \times H\) defined by \((1.2.1)\) gives this set a group structure. Verify that the direct product \((\mathbb{R}, +) \times (\mathbb{R}, +)\) coincides precisely with \((\mathbb{R}^2, +)\).

13. Determine the direct product of the groups described in Exercises 1 and 2.
14. Consider the group \((\mathbb{Z}_4, +)\) given by the following table:

\[
\begin{array}{cccc}
+ & 0 & 1 & 2 \\
0 & 0 & 1 & 2 \\
1 & 1 & 2 & 3 \\
2 & 2 & 3 & 0 \\
3 & 3 & 0 & 1 \\
\end{array}
\]

(a) Determine all its subgroups.

(b) Consider the group with support \(H = \{1, -1, i, -i\} \subset \mathbb{C}\) furnished with the product of complex numbers. Is there a bijection \(f : \mathbb{Z}_4 \rightarrow H\) such that \(f(x + y) = f(x)f(y)\), for all \(x, y \in \mathbb{Z}_4\)? If yes, how many are there?

1.3 Permutations

The bijective maps \(f : X \rightarrow X\) under composition of maps form a group \(S_X\) called the **symmetric group** on \(X\). A bijection from \(X\) onto \(X\) is called a **permutation** of \(X\), specially when \(X\) is a finite set. It should be clear that in this case it does not matter what \(X\) is, only its number of elements. We will now study the group of permutations of the set \(\{1, 2, 3, \ldots, n\}\), which is usually denoted by \(S_n\). An elementary counting argument shows that \(S_n\) is a finite group with \(n!\) elements (here \(n!\) denotes the factorial of \(n\), i.e., the product of the first \(n\) natural numbers).

**Examples 1.3.1.**

1. The group \(S_2\) has only two elements, \(I\) and \(\phi\), where \(I\) is the identity map on the set \(\{1, 2\}\), and \(\phi\) “exchanges” 1 with 2, (i.e., \(\phi(1) = 2\) and \(\phi(2) = 1\)).

2. The map \(\delta : \{1, 2, 3\} \rightarrow \{1, 2, 3\}\) defined by \(\delta(1) = 2\), \(\delta(2) = 3\), and \(\delta(3) = 1\) is one of the six permutations in \(S_3\).

3. More generally, in \(S_n\) we have the permutation \(\pi : S_n \rightarrow S_n\) which cyclic permutes all elements: \(\pi(i) = i + 1\) \((i = 1, \ldots, n-1)\) and \(\pi(n) = 1\).

One can represent a permutation \(\pi\) in \(S_n\) by a matrix with two rows, where on the first row one writes the variable \(x\) and in the second row the value \(\pi(x)\). For example, in the case of \(S_3\) its elements can be represented
by the matrices:

$$I = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \alpha = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix},$$

$$\gamma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad \delta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}.$$  

It is a routine task to find all the products (i.e., composition) of these permutations. The result is given in the following table:

<table>
<thead>
<tr>
<th></th>
<th>$I$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\delta$</th>
<th>$\varepsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I$</td>
<td>$I$</td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\gamma$</td>
<td>$\delta$</td>
<td>$\varepsilon$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$I$</td>
<td>$\delta$</td>
<td>$\varepsilon$</td>
<td>$\beta$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\varepsilon$</td>
<td>$I$</td>
<td>$\delta$</td>
<td>$\gamma$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\delta$</td>
<td>$\varepsilon$</td>
<td>$I$</td>
<td>$\alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$\delta$</td>
<td>$\gamma$</td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\varepsilon$</td>
<td>$I$</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>$\varepsilon$</td>
<td>$\beta$</td>
<td>$\gamma$</td>
<td>$\alpha$</td>
<td>$I$</td>
<td>$\delta$</td>
</tr>
</tbody>
</table>

Given any permutation $\pi$ of $X$ and element $x \in X$, the subset of $X$ consisting of all elements obtained from $x$ by applying $\pi$ repeatedly is denoted by $O_x$ and is called the orbit of the permutation $\pi$. Hence,

$$O_x = \{x, \pi(x), \pi(\pi(x)), \ldots\}.$$

**Examples 1.3.2.**

1. In the case of $S_3$, we have
   - orbits of $I$: $O_1 = \{1\}$, $O_2 = \{2\}$ and $O_3 = \{3\};$
   - orbits of $\alpha$: $O_1 = \{1\}$, $O_2 = O_3 = \{2,3\};$
   - orbits of $\varepsilon$: there is a unique orbit $O_1 = O_2 = O_3 = \{1,2,3\}.$

   The structure of the orbits of $\beta$ and $\gamma$ is similar to the orbits of $\alpha$ (can you say precisely what they are?), while $\varepsilon$, just like $\delta$, has a unique orbit (which?).

2. The permutation in $S_4$ given by

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

   has orbits $O_1 = O_2 = \{1,2\}$ and $O_3 = O_4 = \{3,4\}.$

The length of an orbit is simply the number of elements in the orbit. Note that the orbits of a given permutation $\pi$ of $X$ are disjoint subsets of
1.3. PERMUTATIONS

Let $X$, whose union is $X$, so we say that the orbits of $\pi$ form a partition of the set $X$. Note also that the identity is the unique permutation with all orbits having length 1. The permutations with at most one orbit of length greater than 1 are called cycles. For example, all permutations in $S_3$ are cycles, but the permutation $\pi$ in $S_4$ mentioned above is not a cycle: it has two orbits of length 2. Note also that $\pi(x) \neq x$, precisely when $x$ belongs to an orbit of $\pi$ of length greater than 1. A cycle with an orbit of length 2 is called a transposition.

If a permutation $\pi$ is a cycle, it is more neat to represent it by indicating which elements belong to its unique orbit $O_x$ of length greater than 1: we write $(x, \pi(x), \pi^2(x), \ldots, \pi^{k-1}(x))$, where $k$ is the length of $O_x$, i.e., the least natural number such that $\pi^k(x) = x$.

Example 1.3.3.

For $S_4$ we have:

\[
\alpha = (2, 3) = (3, 2), \quad \beta = (1, 3) = (3, 1), \quad \gamma = (1, 2) = (2, 1), \\
\delta = (1, 2, 3) = (2, 3, 1) = (3, 1, 2), \quad \varepsilon = (1, 3, 2) = (3, 2, 1) = (2, 1, 3).
\]

We can represent the identity $I$, e.g., by $I = (1)$. Note that the inverse permutation of a cycle is the cycle that is obtained by inverting the order in which the elements appear in the corresponding orbit:

\[
(i_1, i_2, \ldots, i_k)^{-1} = (i_k, \ldots, i_2, i_1).
\]

In particular, the inverse permutation of a transposition is the same transposition. In the example of $S_3$, the permutations $\alpha$, $\beta$, and $\gamma$ coincide with there inverses, and $\varepsilon$ and $\delta$ are inverses of each other.

Two cycles are said to be disjoint if the orbits of length greater than 1 are disjoint. Two disjoint cycles $\pi$ and $\rho$ commute, i.e., $\pi \rho = \rho \pi$, and any permutation is a product of disjoint cycles (one cycle for each of its orbits of length greater than 1). More precisely, in $S_n$ we have the following factorization result, which can be thought of as an analogue of the Fundamental Theorem of Arithmetic.

Proposition 1.3.4. Every permutation $\pi$ in $S_n$ is a product of disjoint cycles. This factorization is unique up to the order of the factors.

---

8 "Every natural number $n \geq 2$ is a product of prime numbers, which are unique up to the order of the factors" (see Chapter 2).
Notice that, in general, we have

\[(x_1, x_2, \ldots, x_m) = (x_1, x_m) \ldots (x_1, x_3)(x_1, x_2).\]

Hence, it is possible to factor permutations in \(S_n\) into a product of transpositions. In this case, however, the factors are not unique and the order matters, since it’s unavoidable to use non-disjoint transpositions.

**Examples 1.3.5.**

1. For the permutation \(\pi\) in \(S_4\) above, we can factor it as:

\[
\begin{pmatrix}
1 & 2 & 3 & 4 \\
2 & 1 & 4 & 3
\end{pmatrix} = \begin{pmatrix}
1 & 2 & 3 & 4 \\
2 & 1 & 3 & 4
\end{pmatrix} \begin{pmatrix}
1 & 2 & 3 & 4 \\
1 & 2 & 4 & 3
\end{pmatrix},
\]

i.e., we can write this permutation in the form \(\pi = (1,2)(3,4) = (3,4)(1,2)\).

2. Similarly, the cycle \((1,2,4,3)\) can be factor as a product of transpositions:

\[
(1,2,4,3) = (1,3)(1,4)(1,2).
\]

But this cycle also admits, e.g., the factorizations:

\[
(1,2,4,3) = (2,1)(2,3)(2,4) = (1,3)(1,4)(1,2)(2,4)(1,3)(2,4)(1,3).
\]

The previous example shows that in a factorization of a permutation as a product of transpositions, these are not uniquely determined. Note also that the number of transpositions in two distinct factorizations is not unique. In spite of this lack of uniqueness, it is possible to show that the number of factors has a fixed parity, i.e., is either always even or always odd. To see this, if \(\pi\) is a permutation with orbits \(O_1, O_2, \ldots, O_L\) having lengths \(n_1, n_2, \ldots, n_L\), we set \(P(\pi) = \sum_{i=1}^{L} (n_i - 1)\), and we show that:

**Proposition 1.3.6.** If \(\pi\) is a permutation and \(\tau\) is a transposition, then

\[P(\pi \tau) = P(\pi) \pm 1.\]

**Proof.** Let \(\tau = (a, b)\), We must consider two cases:

(a) The elements \(a\) and \(b\) belong to distinct orbits \(O_i\) and \(O_j\) of \(\pi\), and

(b) The elements \(a\) and \(b\) belong to the same orbit \(O_i\).

For each of these cases it is to check that the following statements hold:
(a) $\pi\tau$ has the same orbits as $\pi$, with the exception of $O_i$ and $O_j$, which now form a unique orbit of length $n_i + n_j$. Hence, $P(\pi\tau) = P(\pi) + 1$;

(b) $\pi\tau$ has the same orbits as $\pi$, with the exception of $O_i$, which is split into two orbits. In this case, $P(\pi\tau) = P(\pi) - 1$.

We can now show that:

**Theorem 1.3.7.** If $\tau_1, \tau_2, \ldots, \tau_m$ are transpositions such that $\pi = \tau_1\tau_2 \cdots \tau_m$, then $P(\pi) - m$ is even, and hence $P(\pi)$ and $m$ have the same parity (are both even or both odd).

**Proof.** We use induction in $m$.

If $m = 1$, then $\pi$ is a transposition and $P(\pi) = 1$, so $P(\pi) - m = 0$ is even.

If $m > 1$, we let $\alpha = \tau_1\tau_2 \cdots \tau_{m-1}$. By the induction hypothesis, we have that $P(\alpha) - (m-1)$ is even, and by the previous proposition $P(\pi) = P(\alpha) \pm 1$. We conclude that

$$P(\pi) - m = (P(\alpha) \pm 1) - m - 1 + 1 = P(\alpha) - (m - 1) - (1 \pm 1)$$

is even. 

The parity of a permutation $\pi$ is the parity of the number of transpositions in any factorization of $\pi$ into a product of transpositions or, as we have just seen, the parity of $P(\pi)$. If $P(\pi)$ is even (respectively, odd), we say that $\pi$ is an even (respectively, odd) permutation. The sign of $\pi$ is $+1$ (respectively, $-1$), if $\pi$ is even (respectively, odd), and is denoted by $\text{sgn}(\pi)$. In particular, any transposition is odd, and also the cycle $(1, 2, 4, 3)$. The identity is an even permutation since, e.g., $I = (1, 2)(1, 2)$.

The even permutations in $S_n$ form a subgroup, denoted by $A_n$, called the alternating group (in $n$ symbols). We leave for the exercises to check that $A_n$ contains $\frac{n!}{2}$ elements.

**Exercises.**

1. Find the factorization of the permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 3 & 4 & 1 & 6 & 7 & 5 \end{pmatrix}$$

into a product of disjoint cycles.
2. What is the parity of the permutation $\pi$ in the previous exercise?

3. How many transpositions are there in $S_n$?

4. How many distinct cycles of length $k$ ($1 \leq k \leq n$) are there in $S_n$?

5. Show that, if $\pi, \rho \in S_n$, then $\text{sgn}(\pi \rho) = \text{sgn}(\pi) \text{sgn}(\rho)$.

6. Prove that $A_n$ is a subgroup of $S_n$.

7. Determine all elements of the group $A_3$.

8. Determine all the subgroups of the group $S_3$.

9. Show that in $S_n$ ($n > 1$) the number of even and odd permutations coincide and conclude that $A_n$ contains $\frac{n!}{2}$ elements.

1.4 Homomorphisms and Isomorphisms

In order to compare algebraic structures which satisfy the same abstract definition, one uses a fundamental notion of Algebra, namely the notion of an isomorphism. This is a special case of the more general notion of a homomorphism, whose formal definition for monoids is as follows:

**Definition 1.4.1.** If $(X, \ast)$ and $(Y, \cdot)$ are monoids, a map $\phi : X \to Y$ is called a homomorphism if

$$
\phi(x_1 \ast x_2) = \phi(x_1) \cdot \phi(x_2), \quad \forall x_1, x_2 \in X.
$$

A homomorphism $\phi$ which is a bijection is called an isomorphism, in which case the monoids are said to be isomorphic.\footnote{The use of the following terms is also common: a monomorphism is an injective homomorphism and a epimorphism is a surjective homomorphism. Also, an endomorphism is a homomorphism from an algebraic structure into itself, while an automorphism is an isomorphism of an algebraic structure with itself.}

One suggestive way to describe the notion of an homomorphism is through the following diagram:

$$
\begin{array}{ccc}
X \times X & \xrightarrow{\text{"\ast"}} & X \\
\downarrow{\phi \times \phi} & & \downarrow{\phi} \\
Y \times Y & \xrightarrow{\text{"\cdot"}} & Y
\end{array}
$$
1.4. HOMOMORPHISMS AND ISOMORPHISMS

Such a diagram is said to be *commutative*, because one can follow two distinct paths on it without changing the final result.

**Examples 1.4.2.**

1. The logarithm function \( \phi : \mathbb{R}^+ \to \mathbb{R} \) given by \( \phi(x) = \log(x) \), is a bijection. Since \( \log(xy) = \log(x) + \log(y) \), the groups \((\mathbb{R}^+, \cdot)\) and \((\mathbb{R}, +)\) are isomorphic. Note that the inverse function, namely the (exponential) \( \psi(x) = \exp(x) \) is also an isomorphism.

2. The set of all linear transformations \( T : \mathbb{R}^n \to \mathbb{R}^n \) under composition is a monoid. Once a basis of \( \mathbb{R}^n \) has been fixed, it is possible to find for each linear transformation \( T \) its matrix representation \( M(T) \), a certain \( n \times n \) matrix. It is easy to see that the map \( T \mapsto M(T) \) is an isomorphism of monoids (composition of linear transformations corresponds to the product of their respective matrix representations).

If \((G, \ast)\) and \((H, \cdot)\) are *isomorphic* groups we will write “\((G, \ast) \simeq (H, \cdot)\)”, or even just “\(G \simeq H\)”, whenever the binary operations are obvious from the discussion. For example, we have \((\mathbb{R}^+, \cdot) \simeq (\mathbb{R}, +)\), or \(\mathbb{R}^+ \simeq \mathbb{R}\).

**Proposition 1.4.3.** Let \((G, \ast)\) and \((H, \cdot)\) be groups, with identities \(e\) and \(\tilde{e}\) respectively, and let \(\phi : G \to H\) be a homomorphism. Then

(i) Invariance of the identity: \(\phi(e) = \tilde{e}\).

(ii) Invariance of inverses: \(\phi(g^{-1}) = (\phi(g))^{-1}, \forall g \in G\).

**Proof.** (i) Since

\[
\phi(e) \cdot \phi(e) = \phi(e \ast e) = \phi(e),
\]

the cancellation law shows that \(\phi(e) = \tilde{e}\).

(ii) Since

\[
\phi(g) \cdot \phi(g^{-1}) = \phi(g \ast g^{-1}) = \phi(e) = \tilde{e},
\]

we conclude that \(\phi(g^{-1}) = (\phi(g))^{-1}\).

**Examples 1.4.4.**

1. Consider the groups \((\mathbb{R}, +)\) and \((\mathbb{C}^*, \cdot)\), where \(\mathbb{C}^*\) denotes the set of all non-zero complex numbers. Define \(\phi : \mathbb{R} \to \mathbb{C}^*\), \(\phi(x) = e^{2\pi xi} = \cos(2\pi x) + i \sin(2\pi x)\).

\[\text{We will also use the same symbol } \simeq \text{ to indicate that two objects in other contexts (e.g., monoids, rings, fields) are isomorphic.}\]
Then $\phi$ is a homomorphism $(e^z \cdot e^w = e^{z+w}$, even when $z$ and $w$ are complex numbers$)$. Obviously, $\phi$ is not surjective (because $\phi(x)$ is a complex number of absolute value 1), and is not injective (because, if $x = n$ is an integer, then $\phi(n) = 1$. Note that $\phi$ “wraps” the real line over the unit circle. According to the proposition above, the identity in the domain (the real number 0), is transformed into the identity in the target (the complex number 1), and the image of the symmetric of the real number $x$ is the inverse of the complex number $\phi(x)$.

2. Consider the groups $(\mathbb{Z}, +)$ and $(\mathbb{C}^*, \cdot)$ and define

$$\phi : \mathbb{Z} \rightarrow \mathbb{C}^*, \quad \phi(n) = i^n.$$ 

The map $\phi$ is a homomorphism which is neither surjective nor injective. The identity in the domain, the integer 0, is transformed into the identity in the target, the complex number 1, and the image of the symmetric of the integer $n$ is the inverse of the complex number $\phi(n)$.

3. The previous example can be generalized as follows: if $(G, \ast)$ is any group and $g \in G$ is a fixed element, we can define (multiplicative notation):

$$\phi : \mathbb{Z} \rightarrow G, \quad \phi(n) = g^n.$$ 

The map $\phi$ is a group homomorphism from $(\mathbb{Z}, +)$ into $(G, \ast)$.

Given a group homomorphism $\phi : G \rightarrow H$, consider now the equation

$$\phi(x) = y,$$

where we assume $y \in H$ fixed, and $x$ is the unknown to be determined. By analogy with Linear Algebra, if $y = \tilde{e}$ is the identity in the target group we say that we have an homogenous equation, and if $y \neq \tilde{e}$ we say that it is inhomogenous. The set of solutions of the homogenous equation is called the kernel of the homomorphism and will be denoted by $N(\phi)$:

$$N(\phi) = \{g \in G : \phi(g) = \tilde{e}\}.$$ 

On the other hand, the set of all $y \in H$ for which the equation $\phi(x) = y$ has a solution $x \in G$, is called the image of the homomorphism and will be denoted by $\phi(G)$ (or $\text{Im}(\phi)$):

$$\phi(G) = \{\phi(g) : g \in G\}.$$ 

\[\text{Recall that if } z = x + iy \text{ is a complex number, where } x, y \in \mathbb{R}, \text{ one defines } e^z = e^x (\cos(y) + i \sin(y)).\]
Examples 1.4.5.

1. Continuing the discussion in Examples 1.4.4, the kernel of \( \phi : \mathbb{R} \to \mathbb{C}^* \) is precisely the set \( N(\phi) = \mathbb{Z} \) of all integers and the image is the set \( \phi(\mathbb{R}) = \mathbb{S}^1 \) of all complex numbers of absolute value 1 (the unit circle).

2. Similarly, the kernel of \( \phi : \mathbb{Z} \to \mathbb{C}^* \) is precisely the set \( N(\phi) = 4\mathbb{Z} \) of all integers which are a multiple of 4 and the image is the set \( \phi(\mathbb{Z}) = \{1, -1, i, -i\} \).

The next figure illustrates the concepts of kernel and image.

![Figure 1.4.1: Kernel and image of a homomorphism.](image)

In the examples above, both the kernel and the image are subgroups of the domain and target groups, respectively. This is a general fact:

**Proposition 1.4.6.** If \((G, \ast)\) and \((H, \cdot)\) are groups and \( \phi : G \to H \) is a homomorphism, then:

(i) The kernel \( N(\phi) \) is a subgroup of \( G \);

(ii) The image \( \phi(G) \) is a subgroup of \( H \).

**Proof.** (i) The kernel \( N(\phi) \) is non-empty, since it contains at least the identity of \( G \), by Proposition 1.4.3 (i). Moreover, if \( g_1, g_2 \in N(\phi) \) we have

\[
\phi(g_1 \ast g_2^{-1}) = \phi(g_1) \cdot \phi(g_2^{-1}) = \tilde{e} \cdot (\phi(g_2))^{-1} = \tilde{e} \cdot \tilde{e}^{-1} = \tilde{e}.
\]
This shows that for any \( g_1, g_2 \in N(\phi) \) we have \( g_1 * g_2^{-1} \in N(\phi) \), so \( N(\phi) \) is a subgroup of \( G \).

(iii) \( \phi(G) \) is non-empty, because \( G \) is non-empty. If \( h_1, h_2 \in \phi(G) \), there exist \( g_1, g_2 \in G \) such that \( h_1 = \phi(g_1) \) and \( h_2 = \phi(g_2) \). Hence:

\[
h_1 \cdot h_2^{-1} = \phi(g_1) \cdot (\phi(g_2))^{-1} = \phi(g_1 * g_2^{-1}) \in \phi(G),
\]
so \( \phi(G) \) is a subgroup of \( H \).

From now on we will stop being explicit about the group operations, except if needed for clarity. Hence, if \( G \) and \( H \) are groups, \( g_1, g_2 \in G \) and \( h_1, h_2 \in H \), we will write both products as \( g_1g_2 \) and \( h_1h_2 \). From the context, it will be clear to which operation we are referring too. Similarly, we will denote by \( e \) both the identities in \( G \) and in \( H \).

A very important remark is that the kernel of a group homomorphism is not an arbitrary subgroup. It satisfies the following property:

**Definition 1.4.7.** If \( H \subset G \) is a subgroup, one says that \( H \) is a normal subgroup of \( G \) if for all \( h \in H \) and \( g \in G \), one has \( ghg^{-1} \in H \).

**Examples 1.4.8.**

1. When \( G \) is an abelian group, it is clear that \( ghg^{-1} = hgg^{-1} = h \), so all subgroups of an abelian group are normal subgroups.

2. If \( G = S_3 \) and \( H = \{e, \alpha\} \), then \( H \) is a subgroup of \( G \) which is not normal: \( \varepsilon \alpha \varepsilon^{-1} = \gamma \notin H \).

3. If \( G = S_3 \) and \( H = A_3 = \{I, \delta, \varepsilon\} \), then \( H \) is a normal subgroup. In fact, recall that \( A_3 \) is formed by the even permutations of \( S_3 \), so if \( \pi \in A_3 \) and \( \sigma \in S_3 \), then \( \sigma \pi \sigma^{-1} \) is an even permutation (why?), and hence \( \sigma \pi \sigma^{-1} \in A_3 \).

4. If \( G \times H \) is the direct product of the groups \( G \) and \( H \), then \( G \) and \( H \) (identified, respectively, with \( G \times \{e\} \) and \( \{e\} \times H \)) are normal subgroups of \( G \times H \).

**Theorem 1.4.9.** If \( \phi : G \to H \) is a group homomorphism, then its kernel \( N(\phi) \) is a normal subgroup of \( G \).

**Proof.** Let \( n \in N(\phi) \) and \( g \in G \). We must show that \( gng^{-1} \in N(\phi) \), or that \( \phi(gng^{-1}) = e \), where \( e \) is the identity of \( H \). For this, we observe that:

\[
\phi(gng^{-1}) = \phi(g)\phi(n)\phi(g^{-1}) \quad \text{(definition of homomorphism)},
\]
\[
= \phi(g)\phi(g^{-1}) \quad \text{(because } \phi(n) = e, \text{ since } n \in N(\phi)),
\]
\[
= e \quad \text{(invariance of inverses)}.
\]

\[\square\]
In Linear Algebra, the number of solutions of the non-homogeneous equation \( \phi(x) = y \), i.e., the lack of injectivity of \( \phi \), depends only on the kernel \( N(\phi) \). Similarly, we have:

**Theorem 1.4.10.** Let \( \phi : G \rightarrow H \) be a homomorphism. Then:

(i) \( \phi(g_1) = \phi(g_2) \) if and only if \( g_1g_2^{-1} \in N(\phi) \);

(ii) \( \phi \) is injective if and only if \( N(\phi) = \{ e \} \);

(iii) if \( x_0 \) is a particular solution of \( \phi(x) = y_0 \), the general solution is \( x = x_0n \), where \( n \in N(\phi) \).

**Proof.** (i) Notice that:

\[
\phi(g_1) = \phi(g_2) \iff \phi(g_1)(\phi(g_2))^{-1} = e \quad \text{(multiplication in } H \text{ by } (\phi(g_2))^{-1}),
\]

\[
\iff \phi(g_1g_2^{-1}) = e \quad \text{(because } \phi \text{ is a homomorphism),}
\]

\[
\iff g_1g_2^{-1} \in N(\phi) \quad \text{(by definition of } \phi).}
\]

(ii) \( \phi \) is injective if and only if \( \phi(g_1) = \phi(g_2) \iff g_1 = g_2 \iff g_1g_2^{-1} = e \). From (i), we conclude that this holds if and only if \( N(\phi) = \{ e \} \).

(iii) If \( \phi(x_0) = y_0 \), \( n \in N(\phi) \), and \( x = x_0n \), it is clear that

\[
\phi(x) = \phi(x_0n) = \phi(x_0)\phi(n) = \phi(x_0)e = \phi(x_0) = y_0,
\]

so \( x \) is also a solution of the inhomogenous equation. On the other hand, if \( x \) is a solution of the inhomogenous equation, then \( \phi(x) = \phi(x_0) \), so \( xx_0^{-1} \in N(\phi) \). But \( xx_0^{-1} = n \iff x = x_0n \). \( \square \)

**Examples 1.4.11.**

1. Continuing with Examples 1.4.4, we saw that the kernel of \( \phi : \mathbb{R}^+ \rightarrow \mathbb{C}^* \) is the set of all integers, i.e.,

\[
\phi(x) = 1 \iff \cos(2\pi x) = 1, \sin(2\pi x) = 0 \iff x \in \mathbb{Z}.
\]

If we consider the equation:

\[
\phi(x) = i \iff \begin{cases}
\cos(2\pi x) = 0, \\
\sin(2\pi x) = 1,
\end{cases}
\]

an obvious solution is \( x_0 = \frac{1}{4} \). The general solution of this equation is then \( x = \frac{1}{4} + n \), with \( n \in \mathbb{Z} \).
2. The kernel of \( \phi : \mathbb{Z} \to \mathbb{C}^* \) is the set of multiples of 4, i.e.,
\[
\phi(n) = 1 \iff i^n = 1 \iff n = 4k, \text{ with } k \in \mathbb{Z}.
\]
The equation
\[
\phi(n) = i \iff i^n = i,
\]
has the obvious solution \( n_0 = 1 \). The general solution of this equation is then \( n = 1 + 4k, \) with \( k \in \mathbb{Z} \).

3. The algebraic structures \((\mathbb{R}^n, +)\) and \((\mathbb{R}^m, +)\), where addition is the usual vector sum, are abelian groups. If \( T : \mathbb{R}^n \to \mathbb{R}^m \) is a linear transformation, one checks easily that \( T \) is also a group homomorphism, and the usual Linear Algebra result concerning solutions of the equation \( T(x) = y \) is just a special case of the previous theorem.

The notion of isomorphism between two algebraic structures satisfying the same set of axioms is at the basis of another fundamental problem of Algebra, namely the classification problem for algebraic structures. Loosely speaking, this problem is the following:

Given some definition of an (abstract, axiomatic) algebraic structure determine a class \( \mathcal{C} \) of concrete algebraic structures satisfying the definition, and such that any algebraic structure satisfying the definition is isomorphic to exactly one algebraic structure in the class \( \mathcal{C} \).

For example, the problem of classification of finite simple groups (a very important class of groups which we will study later in Chapter 5) was solved in the last 20 years, and is consider one of the major accomplishments of modern mathematics. We will see later some less complex classification problems. For now, in order to illustrate these ideas, we discuss a very elementary example: the classification of monoids with exactly two elements.

If \((X, \ast)\) is a monoid with two elements, we have \( X = \{I, a\} \), where \( I \) denotes the identity, and \( I \neq a \). Note that the products \( I \ast I \), \( I \ast a \) and \( a \ast I \) are completely determined by the fact that \( I \) is the identity (\( I \ast I = I \), \( I \ast a = a \ast I = a \)). It remains to determine the product \( a \ast a \), and the only possibilities are \( a \ast a = a \) or \( a \ast a = I \) (in the later case, \( a \) is invertible, and hence the monoid is a group). The multiplication tables which describe these two possibilities are:

\[
\begin{array}{c|cc}
I & a \\
\hline
I & I & a \\
a & a & I
\end{array}
\quad
\begin{array}{c|cc}
I & a \\
\hline
I & I & a \\
a & a & a
\end{array}
\]
In order to check that both cases are possible, consider the sets $M = \{0, 1\}$, and $G = \{1, -1\}$, with the usual product of integers. It is obvious that in both cases this product is a binary operation which is associative, with identity. Hence, each of these sets with the usual product is a monoid with two elements. Note also, by inspection of the main diagonals of the tables, that these monoids are not isomorphic.

\[
\begin{array}{c|cc}
 & 1 & -1 \\
\hline
1 & 1 & -1 \\
-1 & -1 & 1 \\
\end{array}
\quad
\begin{array}{c|cc}
 & 1 & 0 \\
\hline
1 & 1 & 0 \\
0 & 0 & 0 \\
\end{array}
\]

Obviously, any of the first two multiplication tables represent a monoid isomorphic to one of these monoids. Indeed, in the case $a \ast a = I$, the isomorphism is the map $\phi : X \to G$ given by $\phi(I) = 1$ and $\phi(a) = -1$. In the case $a \ast a = a$, the isomorphism is the map $\phi : X \to M$ given by $\phi(I) = 1$ and $\phi(a) = 0$. To summarize this discussion:

- $G$ and $M$ are monoids with two elements,
- $G$ and $M$ are not isomorphic, and
- If $X$ is any monoid with two elements, then either $X \simeq G$ or $X \simeq M$.

Therefore we can say that, “up to isomorphism”, there exist exactly two monoids with two elements, and the classification of monoids with two elements is the family $\{G, M\}$. Note that since any group is a monoid, and since only $G$ is a group, we can conclude also that “up to isomorphism”, there exists a unique group with two elements, which is $G$.

Exercises.

1. Show that the identity $e^{z+w} = e^z e^w$, with $z$ and $w$ complex numbers, is a consequence of the usual identities for $e^{x+y}$, $\cos(x+y)$ and $\sin(x+y)$, with $x$ and $y$ real numbers.

2. The complex solutions of the equation $x^n = 1$ are the complex numbers $e^{2\pi ki/n}$, where $1 \leq k \leq n$. These are called the $n$-th roots of the unit, and they form the vertices of a regular polygon with $n$ sides, inscribed in the unit circle, and with one vertex on the number 1.

(a) Show that $\phi : \mathbb{Z} \to \mathbb{C}^*$ given by $\phi(k) = e^{2\pi ki/n}$ is a homomorphism.

(b) Conclude that the $n$-th roots of unit form a subgroup of $\mathbb{C}^*$.

---

12. This group is obviously isomorphic to the group $(\mathbb{Z}_2, +)$ mentioned before.
(c) Determine the kernel of the homomorphism $\phi$.

3. Assume that $G$, $H$ and $K$ are groups. Show that:
   
   (a) $G \times H \simeq H \times G$, and $G \times (H \times K) \simeq (G \times H) \times K$.
   
   (b) $\phi : G \to H \times K$ is a group homomorphism if and only if $\phi(x) = (\phi_1(x), \phi_2(x))$, where $\phi_1 : G \to H$ and $\phi_2 : G \to K$ are homomorphisms.

4. Consider the algebraic structure $(\mathbb{R}^2, +)$, where $+$ is the usual sum of vectors.
   
   (a) Show that $(\mathbb{R}^2, +)$ is an abelian group.
   
   (b) Show that any linear transformation $T : \mathbb{R}^2 \to \mathbb{R}$ is a homomorphism from group $(\mathbb{R}^2, +)$ into the group $(\mathbb{R}, +)$.
   
   (c) Show that any homomorphism from the group $(\mathbb{R}^2, +)$ into the group $(\mathbb{R}, +)$ which is continuous is a linear transformation.
   
   (d) For a fixed $a \in \mathbb{R}^2$, define a linear transformation $T : \mathbb{R}^2 \to \mathbb{R}$ by $T(x) = a \cdot x$, where “$\cdot$” denotes the usual inner product. Find the kernel of $T$, and check that it is a subgroup and a linear subspace.
   
   (e) Given an example of a subgroup of $(\mathbb{R}^2, +)$ which is not a linear subspace.

5. Assume that $(A, \ast)$ and $(B, \cdot)$ are monoids with identities $e$ and $\tilde{e}$, respectively. If $\phi : A \to B$ is an isomorphism, show that:
   
   (a) $\phi(e) = \tilde{e}$.
   
   (b) $\phi(a^{-1}) = (\phi(a))^{-1}$ if $a \in A$ is invertible.
   
   (c) $\phi^{-1} : B \to A$ is also an isomorphism.
   
   (d) $(A, \ast)$ is a group if and only if $(B, \cdot)$ is a group.

6. If in the previous exercise one only assumes that $\phi$ is an injective (respectively, surjective) homomorphism, which of the statements remain valid?

7. Let $G$ be a group and denote by Aut$(G)$ the set all automorphisms $\phi : G \to G$.
   
   (a) Show that Aut$(G)$ furnished with composition of maps is a group.
   
   (b) If $G$ is the group formed by the complex solutions of $x^4 = 1$, determine Aut$(G)$.

8. Let $G$ be a group.
   
   (a) If $g \in G$ is fixed, show that $\phi_g : G \to G$ given by $\phi_g(x) = gxg^{-1}$ is an automorphism.

---

13 There are group homomorphisms which are not continuous and which are not linear transformations. Its existence can only be shown using the Axiom of Choice, which is discussed in the Appendix.
(b) Show that the map \( T : G \to \text{Aut}(G) \) defined by \( T(g) = \phi_g \) is a group homomorphism.

(c) Show that the kernel of \( T \) is the center of the group \( G \) (see Exercise 7, in the previous section).

9. Classify all the groups with three and with four elements.

10. Show that the map \( \phi : S_n \to \mathbb{Z}_2 \), which to a permutation \( \pi \) associates \( \text{sgn}(\pi) \), is a group homomorphism with kernel \( A_n \).

11. Determine all the normal subgroups of \( S_3 \).

12. Let \( G \) be a group and \( \phi : S_3 \to G \) a group homomorphism. Classify the group \( \phi(S_3) \) (i.e., list all the possibilities for \( \phi(S_3) \) up to isomorphism).

13. Let \( G \) be a group and fix \( g \in G \). Show that
   
   (a) the map \( T_g : G \to G \) given by \( T_g(x) = gx \), is a permutation on the set \( G \).
   
   (b) the map \( \phi(g) : G \to S_G \) given by \( \phi(g) = T_g \) is an injective homomorphism. In particular, \( G \) is isomorphic to a subgroup of the group of permutations of the set \( G \).

   (c) if \( G \) is a finite group with \( n \) elements, then there exists a subgroup \( H \subseteq S_n \) such that \( G \cong H \).

1.5 Rings, Integral Domains and Fields

The integers, rationals, reals and complex numbers can be added and multiplied and the result is still a number of the same type. Similarly, we can add and multiply square matrices of the same size, linear transformations of a fixed vector space, and many other objects, which are useful in Mathematics and its applications. All these are examples of algebraic structures more complex than groups or monoids, precisely because they involve simultaneously two operations. However, they share a common set of basic properties, which lies at the basis of the definition of an algebraic structure called a ring, that we will now study. We will distinguish some important classes of rings, namely the fields (rings where the product is commutative and a division by non-zero elements is always possible, such as \( \mathbb{Q}, \mathbb{R} \) and \( \mathbb{C} \)), and integral domains (rings with properties similar to \( \mathbb{Z} \)).

Let \( A \) be a non-empty set, and \( \sigma, \pi : A \times A \to A \) two binary operations in \( A \). In order to simplify the notation, we will write “\( a + b \)” instead of
“σ(a, b)” and “a · b” (or still ab) instead of “π(a, b)”. We say that \(a + b\) and \(a · b\) are respectively the addition and the product of the elements \(a, b \in A\).

**Definition 1.5.1.** The triple \((A, +, ·)\) is called a ring if:

(i) **Addition properties:** \((A, +)\) is an abelian group.

(ii) **Product properties:** The product is associative, \(i.e., \forall a, b, c \in A, (a · b) · c = a · (b · c)\).

(iii) **Mixed properties:** The addition and the product are distributive, \(i.e., \forall a, b, c \in A, a · (b + c) = a · b + a · c, \text{ and } (b + c) · a = b · a + c · a\).

Let us list more explicitly all the axioms included in this definition:

- **Associativity of addition:** \(\forall a, b, c \in A, (a + b) + c = a + (b + c)\).
- **Commutativity of addition:** \(\forall a, b \in A, a + b = b + a\).
- **Identity for addition:** \(\exists 0 \in A \forall a \in A, a + 0 = a\).
- **Symmetrics for addition:** \(\forall a \in A \exists b \in A : a + b = 0\).
- **Associativity of product:** \(\forall a, b, c \in A, (a · b) · c = a · (b · c)\).
- **Distributivity:** \(\forall a, b, c \in A, a · (b + c) = a · b + a · c, \text{ and } (b + c) · a = b · a + c · a\).

As we have already mentioned, Definition 1.5.1 is at least partially inspired by the properties of the set of integers \(A = \mathbb{Z}\), where “addition” and “product” are the usual operations on integers, and 0 is the integer zero. In particular, all the axioms in the definition correspond to well-known properties of the integers. However, this does not mean that an arbitrary ring \((A, +, ·)\) behaves just like \((\mathbb{Z}, +, ·)\). For example, in general the “product” is not commutative, nor is there any reference in Definition 1.5.1 to an identity for the product (similar to the integer 1). The examples that we will study later show that a ring may have properties dramatically distinct from the ring of integers.

If a ring \(A\) has an identity (for the product) then \((A, ·)\) is a monoid. In this case we say that \((A, +, ·)\) is a unitary ring. We shall always refer to the (unique) identity for addition has the zero of the ring, reserving the term identity without further qualifications to the (unique) identity for the product, whenever the ring has one \(i.e., \) whenever it is a unitary ring). Also, a commutative or abelian ring is a ring where \(a · b = b · a\), for all \(a, b \in A\).
Examples 1.5.2.

1. The set \( \mathbb{Z} \) of all integers with the usual operations of addition and product is an unitary abelian ring. On the other hand, the set of even integers \( 2\mathbb{Z} = \{0, \pm 2, \pm 4, \ldots\} \), still with the usual operations of addition and product, is an abelian ring without identity.

2. The sets of rational, real and complex numbers, denoted respectively by \( \mathbb{Q} \), \( \mathbb{R} \) and \( \mathbb{C} \), also with the usual addition and product are unitary abelian rings.

3. The set of \( n \times n \) square matrices with entries in either \( \mathbb{Z} \), \( \mathbb{Q} \) or \( \mathbb{C} \), which we will denote by \( M_n(A) \), where \( A = \mathbb{Z}, \mathbb{Q}, \mathbb{R} \) or \( \mathbb{C} \), still with the usual operations of addition and product of matrices, are rings (non-abelian if \( n > 1 \)) unitary (the identity is the identity matrix \( I \)). More generally, we can consider the ring of \( n \times n \) matrices \( M_n(A) \) with entries in an arbitrary ring \( A \).

4. The set of all functions \( f : \mathbb{R} \to \mathbb{R} \), with pointwise addition and product:
   \[
   (f + g)(x) = f(x) + g(x), \\
   (fg)(x) = f(x)g(x),
   \]
is a unitary abelian ring (the identity is the constant function equal to 1). Similarly, we can consider the ring of continuous functions, differentiable function, etc.

5. The set \( \mathbb{Z}_2 = \{0, 1\} \), with addition and product defined by
   \[
   0 + 0 = 1 + 1 = 0, \quad 0 + 1 = 1 + 0 = 1, \\
   0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0, \quad 1 \cdot 1 = 1,
   \]
is a unitary abelian ring. Note that the operations in this ring correspond to the boolean operations of “logical disjunction (OR)” and “logical conjunction (AND)” if we associate
   \[
   0 \to \text{False}, \quad 1 \to \text{True}.
   \]
The operations in this ring correspond also to the usual “binary” arithmetic (base 2 arithmetic), with no transport addition.

6. The set of complex numbers of the form \( z = n + mi \), with \( n, m \in \mathbb{Z} \), is a unitary abelian ring for the usual complex addition and multiplication. This ring is usually denoted by the symbol \( \mathbb{Z}[i] \) and its elements are called the GAUSSIAN INTEGERS\(^{14}\).

---

\(^{14}\)Carl Friedrich Gauss (1777-1855) was one of the great mathematicians from Göttingen. He was a prodigious child, and when he was only 19 years old he found out a method to construct a regular polygon with 17 sides using only a ruler and a compass (see Chapter 7). For more than 2000 years, since the greek geometers first attempts of such constructions,
Let us now turn to some basic properties which are shared by any ring. We start with properties which are consequences of results we have seen before in a more general context:

**Proposition 1.5.3.** Let $A$ be a ring.

(i) Cancellation law for addition: $a + c = b + c \Rightarrow a = b$, and in particular $d + d = d \Rightarrow d = 0$.

(ii) Uniqueness of symmetries: The equation $a + x = 0$ has a unique solution in $A$, given by $x = -a$.

(iii) Sign rules: $-(a) = a$, $-(a+b) = (-a) + (-b)$, and $-(a-b) = (-a)+b$.

We have mentioned above that $A$ is a unitary ring if and only if $(A, \cdot)$ is a monoid. In this case, we denote by $A^*$ the set of invertible elements of the monoid $(A, \cdot)$, also-called the invertible elements of the ring $A$. Again, recalling a result we proved in a more general context, we have:

**Proposition 1.5.4.** Let $A$ be a unitary ring. Then $(A^*, \cdot)$ is a group, and hence:

(i) $A^*$ is closed relative to the product.

(ii) If $a \in A^*$, $ax = 1$ has the unique solution $x = a^{-1}$, where $a^{-1} \in A^*$.

(iii) If $a, b \in A^*$, $(ab)^{-1} = b^{-1}a^{-1}$, and $(a^{-1})^{-1} = a$.

**Examples 1.5.5.**

1. The unique invertible integers are 1 and $-1$, i.e., $\mathbb{Z}^* = \{-1, 1\}$.

2. All the non-zero rationals, reals and complex numbers are invertible. Hence, we have, e.g., $\mathbb{Q}^* = \mathbb{Q} - \{0\}$.

3. In the ring $M_n(\mathbb{R})$, the invertible elements are the non-singular matrices, in other words the matrices with non-zero determinat.

In general, in a arbitrary ring $A$ with identity $1 \neq 0$, we can only say that 1 and $-1$ are invertible, because $1 \cdot 1 = (-1) \cdot (-1) = 1$. For the ring $\mathbb{Z}$ these the only regular polygons with a prime number of sides that were known to be constructible were the equilateral triangle and the regular pentagon. Some of the most relevant results to be discussed in this book were actually discovered by Gauss, such as the theory of congruences.

If $A$ and $B$ are sets, then $A - B = \{x \in A : x \notin B\}$.
are the only invertible elements, but at the other extreme, we have rings such as $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$, where all the non-zero elements are invertible. There are intermediate case, such as the ring $M_n(A)$ ($A$ a ring), where determining the invertible elements is more complicated.

Let us look next at properties which involve both operations in a ring, which therefore are not consequence of results we have seen so far. The first property, e.g, shows that the zero in a unitary ring is not invertible (so division by zero is not possible), except for the trivial ring $A = \{0\}$ (where $0 = 1$).

**Proposition 1.5.6.** If $A$ is a ring, for any $a, b \in A$, we have:

(i) Product by zero: $a0 = 0a = 0$.

(ii) Sign rules: $-(ab) = (-a)b = a(-b)$, and $(-a)(-b) = ab$.

**Proof.** To show that $a0 = 0$ we observe that

$$a0 + a0 = a(0 + 0) \quad \text{(distributive property)},$$

$$= a0 \quad \text{(because 0 is the identity for +),}$$

$$\Rightarrow a0 = 0 \quad \text{(by the cancellation law)}.$$

The proof of (ii) is left as an exercise.

Some of the main differences between examples of rings that we have refer so far, are related with the behavior of the product. The only issue is not just the existence of inverses, but also wether the product satisfies the "cancellation law", which can formally be defined as follows:

**Definition 1.5.7.** A ring $A$ satisfies the CANCELLATION LAW FOR THE PRODUCT if

$$\forall a, b, c \in A, \ [c \neq 0 \text{ and } (ac = bc \text{ or } ca = cb)] \Rightarrow a = b.$$

The restriction to elements $c \neq 0$ (which does not appear in the cancellation law for addition) is a consequence of the “product by zero” property. The fact that the cancellation law for the product does not hold in every ring, and hence is not a logical consequence of Definition [1.5.4](#), can already be seen in the example of $M_2(\mathbb{R})$ where we have the product

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$
so that \( ca = cb \), with \( c \neq 0 \) and \( a \neq b \).

This example also shows that there are rings where \( cd = 0 \) with \( c \neq 0 \) and \( d \neq 0 \). In this case we call \( c \) and \( d \) zero divisors. This notion should not be confused with the notion of non-invertible element. However, if \( ca = cb \) (or \( ac = bc \)) and \( c \) is invertible, it is obvious that we can multiply both sides of the equalities by the inverse of \( c \), and conclude \( a = b \). In other words, a zero divisor is always non-invertible, but a ring may have non-invertible elements which are not zero divisors: for example, in \( \mathbb{Z} \) there are no zero divisors.

On the other hand, we leave as an exercise to check that:

**Proposition 1.5.8.** The cancellation law for the product holds in a ring \( A \) if and only if \( A \) has no zero divisors.

In particular, in a ring \( A \) where all non-zero elements are invertible (i.e., \( A^* = A - \{0\} \)), the cancellation law holds. Such a ring is entitled a special name:

**Definition 1.5.9.** A **division ring** is a unitary ring \( A \) where \( A^* = A - \{0\} \). An abelian division ring is called a **field**.

The rings \( \mathbb{Q}, \mathbb{R} \) and \( \mathbb{C} \) are obviously fields. Note that \( \mathbb{Z}_2 \) is a field with two elements, but \( 2\mathbb{Z} \) or \( M_n(\mathbb{R}) \) are neither fields, nor division rings. We will describe later a division ring which is not a field.

There are also rings which are not division rings, since some non-zero elements are not invertible, but where the cancellation law for the product still holds. Examples of these are the ring of integers \( \mathbb{Z} \), the polynomial rings with coefficients in \( \mathbb{Z}, \mathbb{Q}, \mathbb{R} \) ou \( \mathbb{C} \), and the ring of Gaussian integers. This class of rings is also entitled a special name:

**Definition 1.5.10.** An **integral domain** is a unitary abelian ring where the cancellation law for the product holds.

By Proposition 1.5.8 a unitary abelian ring is an integral domain if and only if it has no zero divisors.

The following diagram shows the relationship between the various kinds of rings that we have seen so far (see also the exercises at the end of this section).

If \((A,+,\cdot)\) is a ring and \( B \subset A \), it is possible that \( B \) is closed for both the addition and the product of \( A \), i.e., it is possible that

\[
a, b \in B \Rightarrow a + b \in B \text{ and } a \cdot b \in B.
\]

If this is the case, then it is possible that \((B,+,\cdot)\) is a ring.
**Definition 1.5.11.** Let \((A, +, \cdot)\) be a ring and let \(B \subset A\) be a subset closed for the addition and the product of the ring \(A\). We say that \(B\) is a subring of \(A\) if \((B, +, \cdot)\) is a ring. In this case, we say also that the ring \(A\) is an extension of the ring \(B\).

**Examples 1.5.12.**

1. \(\mathbb{Z}\) is a subring of \(\mathbb{Q}\), and the even integers \(2\mathbb{Z}\) is a subring of \(\mathbb{Z}\).

2. The set \(\mathbb{N}\) of all natural numbers (positive integers) is closed for both the addition and product of \(\mathbb{Z}\), but it is not a subring of \(\mathbb{Z}\).

3. \(\mathbb{C}\) is an extension of \(\mathbb{R}\).

4. The ring \(M_n(\mathbb{C})\) is an extension of \(M_n(\mathbb{Z})\).

According to a result for groups proved in the previous section, if \(B \subset A\) and \(B\) is non-empty, then \((B, +)\) is a subgroup of \((A, +)\) (i.e., it verifies (i) in Definition 1.5.1) if and only if it is closed for the difference. If \(B\) is closed for the addition and product of \(A\), then it is obvious that it satisfies (ii) and (iii) in Definition 1.5.1 simply because its operations are the same as the ones of the ring \(A\). We conclude that

**Proposition 1.5.13.** Let \(A\) be a ring. A subset \(B \subset A\) is a subring of \(A\) if and only if it is non-empty and it is closed for the difference and the product.

If \(A\) and \(B\) are rings we can form the direct product \(A \times B\) of the underlying additive groups. Actually, it should be obvious that one can also introduce a product in \(A \times B\) in a similar fashion, so that:

\[
(1.5.1) \quad (a, b) + (a', b') = (a + a', b + b'), \quad (a, b)(a', b') = (aa', bb').
\]
We leave it as an exercise to check that $A \times B$ with these operations is a ring, which we call the **direct sum** of the rings $A$ and $B$, and which we denote by $A \oplus B$. It should be clear that we can form the direct sum of any **finite** number of rings (later we will consider the case of an **infinite** number of rings.

**Exercises.**

1. Check whether the following algebraic structures are rings. If the answer is yes, indicate whether the ring is commutative, has an identity, and/or verifies the cancellation law for the product. If the answer is no, indicate which axioms in Definition 1.5.1 fail:
   
   (a) The set of all integers multiple of a fixed integer $m$, with the usual addition and product of integers;
   
   (b) The set of all linear transformations $T : \mathbb{R}^n \to \mathbb{R}^n$, with the usual addition and the product being composition;
   
   (c) The set of all functions $f : \mathbb{R} \to \mathbb{R}$, with the usual addition and the product being composition;
   
   (d) The set of all non-negative integers, with the usual addition and product of integers;
   
   (e) The set of all irrational numbers, with the usual addition and product of real numbers;
   
   (f) The set $\mathbb{Z}[i]$ of Gaussian integers, with the usual addition and product of complex numbers;
   
   (g) The set $\mathbb{R}[x]$ of all polynomials in the variable $x$ with real coefficients\footnote{We will see later that if $A$ is a ring, it is possible to define the ring of polynomials “in the variable $x$” with coefficients in $A$, which is usually denoted by $A[x]$. Note that the use of the symbol $\mathbb{Z}[i]$ to denote the ring of Gaussian integers uses the same idea, since a polynomial in the imaginary unit $i$ can always be reduced to a first degree polynomial.}, with the usual addition and the product of polynomials.

2. Show that for any ring $0 = -0$.

3. Complete the proof of Theorem 1.5.6.

4. Let $A$ be a set with three distinct elements, which we denote by 0, 1, and 2. How many distinct operations of “addition” and “product” are there, turning $A$ into a unitary ring, with zero the element 0 and identity the element 1?

5. Show that, in a general ring, the difference operation is neither commutative, nor associative. Give examples of rings where this operation satisfies both properties.
6. The equation \( a + x = b \) has exactly one solution in a ring \( A \). What can you say about the solutions of the equation \( ax = b \)? What about the equation \( x = -x \)?

7. Assume that \( A, B \) and \( C \) are rings. Prove that the following statements hold:

(a) The set \( A \times B \) with the operations defined by \[1.5.1\] is a ring.

(b) If \( A \) and \( B \) both have more than one element, then \( A \oplus B \) has zero divisors.

(c) If \( A \) and \( B \) are unitary rings then \( A \oplus B \) is also a unitary ring and \((A \oplus B)^* = A^* \times B^*\).

(d) \( A \oplus B \) is isomorphic to \( B \oplus A \), and \( A \oplus (B \oplus C) \) is isomorphic to \((A \oplus B) \oplus C\).

(e) \( \phi : A \to B \oplus C \) is a homomorphism of rings if and only if \( \phi(x) = (\phi_1(x), \phi_2(x)) \), with \( \phi_1 : A \to B \) and \( \phi_2 : A \to C \) homomorphisms of rings.

8. If \( X \) is any set and \( A \) is a ring, show that the set of all maps \( f : X \to A \) is a ring with the “usual” operations of addition and product.

9. Use the result of the previous exercise with \( X = \{0, 1\} \) and \( A = \mathbb{Z}_2 \) to obtain an example of a ring with 4 elements. Show that this ring is isomorphic to \( \mathbb{Z}_2 \oplus \mathbb{Z}_2 \).

10. Let \( A = \{0, 1, 2, 3\} \) be a set with four elements. Show that there exists a field structure on \( A \), with zero 0 and identity 1.

11. Let \( A \) be a ring. Show that the set of \( n \times n \) matrices with entries in \( A \), denoted \( M_n(A) \), is a ring. Show that if \( A \) has an identity, then \( M_n(A) \) also has an identity.

12. Let \( B \) be a subring of \( A \). Give examples of rings that show that all the following cases are possible:

(a) \( A \) has identity and \( B \) does not have identity.

(b) \( A \) does not have identity and \( B \) has identity.

(c) \( A \) and \( B \) have distinct identities.

(HINT: consider subrings of rings of \( 2 \times 2 \) matrices).

13. Give an example of non-abelian, finite, ring.

14. Show that any subring of a division ring satisfies the cancellation law for the product. Conclude that a subring of a field which contains the identity of the field is an integral domain (but not necessarily a field).
15. Show that the cancellation law for the product holds in a ring $A$ if and only if $A$ has no zero divisors.

16. Let $A$ be an integral domain. Do the following statements hold in $A$?
   (a) $x^2 = 1$ implies that $x = 1$ or $x = -1$;
   (b) $-1 \neq 1$.

17. Assume that ring $A$ is an extension of the field $K$, and that $K$ contains the identity of $A$. Show that $A$ is vector space over $K$.

18. Determine the invertible elements in the ring $\mathbb{Z}[i]$ of Gaussian integers.

19. Show that a finite integral domain $A \neq \{0\}$ must be a field.

20. Show that $M_2(\mathbb{Z})^*$ is infinite.
   (Hint: Show that $M_2(\mathbb{Z})^* = \{A \in M_2(\mathbb{Z}) : \det(A) = \pm 1\}$.)

21. Let $B$ be a subring of $A$. Show that:
   (a) the zero of $B$ is the zero of $A$;
   (b) the symmetric of an element of $B$ is the same in $B$ and in $A$.

   Assume now that both $A$ and $B$ have an identity.
   (c) Is it true that $B^* \subset A^*$?
   (d) Is it true that the inverse of an element in $B^*$ is necessarily the same as the inverse in $A^*$? What if the identities of $A$ and $B$ coincide?

1.6 Homomorphisms and Isomorphisms of Rings

We consider now the notions of homomorphisms and isomorphisms associated with the concept of ring. As one could expect, these are maps whose domain and target are rings, and which moreover preserve the algebraic operations of these rings. Observe that in the following definition we use the same symbols to represent the algebraic operations of distinct rings, in spite that these can be very different operations. Note also that a homomorphism of rings is a special case of a group homomorphism.

**Definition 1.6.1.** Let $A$ and $B$ be rings and $\phi : A \rightarrow B$ a map. We say that $\phi$ is a HOMOMORPHISM OF RINGS if:
   (i) $\phi(a_1 + a_2) = \phi(a_1) + \phi(a_2), \forall a_1, a_2 \in A$;
1.6. HOMOMORPHISMS AND ISOMORPHISMS OF RINGS

(ii) \( \phi(a_1a_2) = \phi(a_1)\phi(a_2), \forall a_1, a_2 \in A. \)

An isomorphism of rings is a homomorphism which is a bijection\(^{17}\). We say that the rings \( A \) and \( B \) are isomorphic if there exists some isomorphism \( \phi : A \to B.\)

Examples 1.6.2.

1. Let us denote the complex conjugate of \( z = x + iy \) by \( \overline{z} = x - iy. \) We have

\[
\overline{z + w} = \overline{z} + \overline{w}, \quad \overline{zw} = \overline{z}\overline{w}.
\]

According to the previous definition, the map \( \phi : \mathbb{C} \to \mathbb{C} \) given by \( \phi(z) = \overline{z} \) is an automorphism of the ring \( \mathbb{C}. \)

2. Consider the map \( \phi : \mathbb{C} \to M_2(\mathbb{R}) \) defined by

\[
\phi(x + iy) = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}.
\]

Notice that

\[
\begin{pmatrix} x & -y \\ y & x \end{pmatrix} + \begin{pmatrix} x' & -y' \\ y' & x' \end{pmatrix} = \begin{pmatrix} x + x' & -y - y' \\ y + y' & x + x' \end{pmatrix},
\]

so that,

\[
\phi(x + iy) + \phi(x' + iy') = \phi((x + iy) + (x' + iy')).
\]

Similarly,

\[
\begin{pmatrix} x & -y \\ y & x \end{pmatrix} \begin{pmatrix} x' & -y' \\ y' & x' \end{pmatrix} = \begin{pmatrix} xx' - yy' & -(xy' + x' y) \\ xy' + x'y & xx' - yy' \end{pmatrix},
\]

so that,

\[
\phi(x + iy)\phi(x' + iy') = \phi((x + iy)(x' + iy')).
\]

We conclude that \( \phi \) is a homomorphism of rings. It is easy to check that \( \phi \) is an injective homomorphism (i.e., it is a monomorphism) which is not surjective.

3. Let \( \phi : \mathbb{Z} \to \mathbb{Z}_2 \) be given by \( \phi(n) = 0, \) if \( n \) is even, and \( \phi(n) = 1, \) if \( n \) is odd. It is easy to check that \( \phi \) is a surjective homomorphism of rings (i.e., it is a epimorphism), which is not injective.

4. Let \( \phi : \mathbb{R} \to M_2(\mathbb{R}) \) be given by

\[
\phi(x) = \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}.
\]

\(^{17}\)Just like the cases of monoids and groups, we will also use the terms epimorphism and monomorphism of rings, as well as endomorphism and automorphism of rings, which are defined in the obvious way.
It should be obvious that $\phi$ is a monomorphism (which is not surjective). Notice that the matrices in the image of this monomorphism form a subring of $M_2(\mathbb{R})$ with identity distinct from the identity of ring $M_2(\mathbb{R})$.

5. If $S, T : \mathbb{R}^n \to \mathbb{R}^n$ are linear transformations we define the addition $S + T$ and composition $ST$ by

$$(S + T)(x) = S(x) + T(x), \quad (ST)(x) = S(T(x)).$$

We have already observed that these operations make the set of linear transformations of $\mathbb{R}^n$ into itself a ring, which we will denote by $L(\mathbb{R}^n, \mathbb{R}^n)$. Now fix a basis for $\mathbb{R}^n$ and let $M(S)$ denote the matrix of the linear transformation $S$ relative to this basis. It is clear that $M(S)$ is an $n \times n$ matrix with real entries, and we know from Linear Algebra that the map $M : L(\mathbb{R}^n, \mathbb{R}^n) \to M_n(\mathbb{R})$ satisfies the identities

$$M(S + T) = M(S) + M(T), \quad M(ST) = M(S)M(T).$$

Hence $M$ is an isomorphism of rings.

In certain cases, when there exists an “obvious” injective homomorphism $\phi : A \to B$, we use the same symbol to denote $a$ and $\phi(a)$. Although this is not recommended practice from a logical point of view, it is unavoidable in order not to have a too heavy notation. Examples of this practice are the representation of the real number $a$ and the complex number $(a, 0)$ or the constant polynomial $p(x) = a$, by the same symbol.

Since every homomorphism of rings $\phi : A \to B$ is also a homomorphism of the additive groups $(A, +)$ and $(B, +)$, we can restate Proposition 1.4.3 as follows:

**Proposition 1.6.3.** Let $\phi : A \to B$ be a homomorphism of rings. Then:

(i) $\phi(0) = 0$;

(ii) $\phi(-a) = -\phi(a)$.

According to the definition of ring homomorphism and the previous proposition, a homomorphism $\phi : A \to B$ “preserves” additions, products, the zero, and symmetrics. However, certain notions related to the product are not preserved, at least in their most obvious form. One of the examples above shows that if $A$ has identity $1$ for the product, then $\phi(1)$ may fail to be the identity of the product in $B$. We will see in the exercises at the end of the section how to formulate a correct statement concerning “preservation” of identities.
The facts concerning the equation \( \phi(x) = y \) for a ring homomorphism is very similar to the case of group homomorphisms. The kernel of \( \phi \) is the set of solutions of the homogenous equation \( \phi(x) = 0 \), i.e.,

\[
N(\phi) = \{ x \in A : \phi(x) = 0 \},
\]

while the image \( \phi(A) \) is the set of \( y \in B \) for which the equation \( \phi(x) = y \) has solutions \( x \in A \). Obviously, \( \phi \) is surjective if and only if \( \phi(A) = B \), and \( \phi \) is injective if and only if \( N(\phi) = \{0\} \).

If we restate Theorem 1.4.10 in additive notation, we obtain:

**Theorem 1.6.4.** Let \( \phi : A \to B \) be a homomorphism of rings. Then:

(i) \( \phi(x) = \phi(x') \) if and only if \( x - x' \in N(\phi) \);

(ii) \( \phi \) is injective if and only if \( N(\phi) = \{0\} \);

(iii) if \( x_0 \) is a particular solution of \( \phi(x) = y_0 \), then the general solution is \( x = x_0 + n \), with \( n \in N(\phi) \).

Here is a simple example illustrating this result.

**Example 1.6.5.**

Consider the homomorphism \( \phi : \mathbb{Z} \to \mathbb{Z}_2 \) given by \( \phi(n) = 0 \), if \( n \) is even, and \( \phi(n) = 1 \), if \( n \) is odd. The kernel of \( \phi \) is the set of even integers. Since \( \phi(0) = 0 \) and \( \phi(1) = 1 \), we conclude that

\[
\phi(x) = 0 \iff x = 2n, \text{ with } n \in \mathbb{Z},
\]

\[
\phi(x) = 1 \iff x = 1 + 2n, \text{ with } n \in \mathbb{Z}.
\]

Given a homomorphism \( \phi : A \to B \), we know that \( N(\phi) \) and \( \phi(A) \) are subgroups of the additive groups \( (A, +) \) and \( (B, +) \). It is easy to check that these subgroups are actually subrings:

**Proposition 1.6.6.** If \( \phi : A \to B \) is a ring homomorphism, then \( N(\phi) \) is a subring of \( A \), and \( \phi(A) \) is a subring of \( B \). In particular, if \( \phi \) is injective, then \( A \) is isomorphic to \( \phi(A) \).

**Proof.** In order to show that \( N(\phi) \) and \( \phi(A) \) are subrings, we need only to show that they are closed for the respective products.

If \( b_1, b_2 \in \phi(A) \), there exist \( a_1, a_2 \in A \) such that \( b_1 = \phi(a_1) \) and \( b_2 = \phi(a_2) \). Hence, since \( \phi \) is a homomorphism, we have

\[
b_1b_2 = \phi(a_1)\phi(a_2) = \phi(a_1a_2) \in \phi(A),
\]
which shows that $\phi(A)$ is closed for the product of $B$.

If $a_1, a_2 \in N(\phi)$, then $\phi(a_1) = \phi(a_2) = 0$. Again using that $\phi$ is a homomorphism, we have

$$\phi(a_1a_2) = \phi(a_1)\phi(a_2) = 0 \cdot 0 = 0,$$

hence $a_1a_2 \in N(\phi)$, and $N(\phi)$ is closed for the product of $A$.

Finally, it is obvious that if $\phi$ is injective, then $\phi$ gives an isomorphism between $A$ and $\phi(A)$.

**Examples 1.6.7.**

1. Consider again the homomorphism $\phi : \mathbb{C} \to M_2$ defined by

$$\phi(x + iy) = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}.$$

Form the theorem above, we conclude that the set of matrices of the form

$$\begin{pmatrix} x & -y \\ y & x \end{pmatrix}$$

is a subring of $M_2(\mathbb{R})$, isomorphic to the field of complex numbers.

2. If instead we consider the homomorphism $\phi : \mathbb{R} \to M_2$ given by

$$\phi(x) = \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix},$$

we conclude that $M_2$ contains also subring isomorphic to the field of real numbers. Note, however, that $M_2$ contains various distinct subrings isomorphic to the field of real numbers: in the previous example we can restrict to the real numbers and obtain the injective homomorphism $\phi : \mathbb{R} \to M_2$, $x \mapsto xI$, where $I$ is the identity matrix.

We have seen that the kernel of a group homomorphism is a special kind of subgroup, namely a normal subgroup. Similarly, the kernel $N(\phi)$ of a ring homomorphism $\phi : A \to B$ is a subring of $A$ of a special kind: not only $N(\phi)$ is closed for the product, just like any subring, but also in order for the product $ab$ to belong to $N(\phi)$ it is enough that just one of the factors $a$ or $b$ belongs to $N(\phi)$:

**Proposition 1.6.8.** Let $\phi : A \to B$ be a ring homomorphism. If $a \in N(\phi)$ and $a'$ is any element of $A$, then both $aa'$ and $a'a$ belong to $N(\phi)$. 

Proof. Note that

\[ a_1a_2 \in N(\phi) \iff \phi(a_1a_2) = 0 \iff \phi(a_1)\phi(a_2) = 0. \]

For the product \( \phi(a_1)\phi(a_2) \) to be zero it is enough for one of the factors \( \phi(a_1) \) or \( \phi(a_2) \) to be zero, i.e., for one of the elements \( a_1 \) or \( a_2 \) to belong to the kernel \( N(\phi) \).

The subrings with this property deserve a special name.

**Definition 1.6.9.** Let \( A \) be ring and \( I \subset A \) a subring. We say that \( I \) is an **ideal** of \( A \) if for all \( a \in A \) and \( b \in I \) one has \( ab, ba \in I \).

Not all subrings of a ring are ideals. The following examples illustrate both cases.

**Examples 1.6.10.**

1. It is clear that \( \mathbb{Z} \) is a subring of \( \mathbb{R} \) (the difference and the product of integers is always an integer). However, \( \mathbb{Z} \) is not an ideal of \( \mathbb{R} \) (the product of an integer by real number may fail to be an integer).

2. Let \( A = \mathbb{Z} \) and let \( I \) be the set of even integers. Obviously, \( I \) is a subring of \( \mathbb{Z} \) (the difference and the product of even integers is an even integer) but \( I \) is also an ideal of \( \mathbb{Z} \) (the product of any integer by an even integer is always an even integer). We will see later that all the subrings of \( \mathbb{Z} \) are automatically ideals. Note also that \( I \) is a subring of \( \mathbb{R} \), but obviously not an ideal of \( \mathbb{R} \).

3. Every ring \( A \) always has the “trivial” ideals \( \{ 0 \} \) and \( A \).

4. In some cases, a ring only has the trivial ideals. For example, this always happens with a field. In fact, if \( K \) is a field, and \( I \subset K \) is an ideal which contains an element \( x \neq 0 \) (i.e., if \( I \neq \{ 0 \} \)), then \( xx^{-1} = 1 \in I \) (because \( x \in I \) and \( x^{-1} \in K \)). But then every element \( y \in K \) belongs to \( I \), because \( y = 1y \), where \( 1 \in I \) and \( y \in K \).

In a non-abelian ring \( A \) we can consider subrings for which the condition of ideal is only verified for one side: a **left ideal** of \( A \) is a subring \( B \subset A \) such that if \( a \in A \) and \( b \in B \) then \( ab \in B \). Similarly, a **right ideal** of \( A \) is a subring \( B \subset A \) such that if \( a \in A \) and \( b \in B \) then \( ba \in B \). Obviously \( I \subset A \) is an ideal in the sense of Definition 1.6.9 if and only if it is both a left and a right ideal. For an abelian ring, all these notions coincide. Due to Proposition 1.6.8 the lateral ideals play a much less important role that the bilateral ideals.
Exercises.

1. Let $A$ be a ring and $\phi, \psi : A \to A$ endomorphisms. Show that the composition $\phi \circ \psi$ is an endomorphism, but the sum $\phi + \psi$ may fail to be one. In particular, show that the set of endomorphisms of $A$, denoted by $\text{End}(A)$, with the composition operation, form a monoid.

2. Let $A$ be a ring and $\phi, \psi : A \to A$ automorphisms. Show that the composition $\phi \circ \psi$ and the inverse $\phi^{-1}$ are automorphisms. In particular, show that the set of all automorphisms of $A$, denoted by $\text{Aut}(A)$, with the composition operation, form a group.

3. Every integer $m$ is of the form $m = 3n + r$, where $n$ is the result of the division of $m$ by 3 and $r$ is the remainder. Note that $n$ and $r$ are unique, as long as $0 \leq r < 3$. Show that the map $\phi : \mathbb{Z} \to \mathbb{Z}_3$ given by $\phi(m) = r$ is a homomorphism of rings. What is the kernel of this homomorphism?

4. Show that if $A$ and $B$ are rings, $A$ has identity 1, and $\phi : A \to B$ is a homomorphism then $\phi(1)$ is the identity of $\phi(A)$, which may differ from the identity of $B$. Show also that if $a \in A^*$, then $\phi(a^{-1})$ is the inverse of $\phi(a)$ in $\phi(A)$, which may not be the inverse of $\phi(a)$ in $B$. In particular, $\phi(a)$ may fail to be invertible in $B$.

5. Show that, if $A$ and $B$ are rings, $A$ has identity 1, and $\phi : A \to B$ is an isomorphism then $\phi(1)$ is the identity of $\phi(A)$, which may differ from the identity of $B$. Show also that if $a \in A^*$, then $\phi(a^{-1})$ is the inverse of $\phi(a)$ in $B$ and $\phi(A^*) = B^*$.

6. Show that, if $A$ and $B$ are rings, $A$ has identity 1, and $\phi : A \to B$ is an isomorphism, then $B$ is a field (respectively, division ring, integral domain) if and only if $A$ is a field (respectively, division ring, integral domain).

7. Show that, if $K$ is a field, $A$ is a ring, and $\phi : K \to A$ is a homomorphism, then either $A$ contains a subring isomorphic to $K$, or $\phi$ is identically 0.

8. Are there any subrings of $\mathbb{Z} \oplus \mathbb{Z}$ which are not ideals of $\mathbb{Z} \oplus \mathbb{Z}$?

9. Determine $\text{End}(A)$ when $A = \mathbb{Z}$ and $A = \mathbb{Q}$.
   (Hint: Find $\phi(1)$ and proceed by induction.)

10. Determine $\text{End}(\mathbb{R})$.
    (Hint: Show that $\phi(x) \geq 0$ whenever $x \geq 0$, so $\phi$ is non-decreasing.)

11. Determine all the homomorphisms $\phi : \mathbb{C} \to \mathbb{C}$ which satisfy $\phi(x) \in \mathbb{R}$ whenever $x \in \mathbb{R}$.
    (Hint: Show that if $\phi(1) \neq 0$ and $x = \phi(i)$ then $x^2 = -1$.)
12. Assume that $I$ is an ideal of $M_2(\mathbb{R})$ and $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \in I$. Show that $I = M_2(\mathbb{R})$.

13. Determine all the ideals of $M_2(\mathbb{R})$.

14. Let $A$ be a commutative ring with identity and consider the ring $M_n(A)$. Show that the map $\det : M_n(A) \to A$ defined by
   \[
   \det(B) = \sum_{\pi \in S_n} \text{sgn}(\pi)a_{1\pi(1)}a_{2\pi(2)} \cdots a_{n\pi(n)},
   \]
   where $B = (a_{ij})$, preserve products (but is not a homomorphism of rings). Conclude that a matrix $B \in M_n(A)^*$ if and only if $\det(B) \in A^*$.

15. Assume that $C \subset B \subset A$ where $B$ is a subring of $A$.
   (a) If $C$ is a subring of $B$, is it true that $C$ is a subring of $A$?
   (b) If $C$ is an ideal of $B$, is it true that $C$ is an ideal of $A$?
   (c) If $C$ is an ideal of $A$, is it true that $C$ is an ideal of $B$?

16. Show that, if $A$ is a unitary abelian ring and its only ideals are $\{0\}$ and $A$, then $A$ is a field. If $A$ is non-abelian, is it true that $A$ is necessarily a division ring?

17. Let $A$ and $B$ be unitary rings.
   (a) Assume that $J$ is an ideal of $A \oplus B$, and show that $(a, b'), (a', b) \in J \Rightarrow (a, b) \in J$.
   (b) Show that $K \subset A \times B$ is an ideal of $A \oplus B$ if and only if $K = K_1 \times K_2$, where $K_1$ is an ideal of $A$, and $K_2$ is an ideal of $B$.

18. This exercise concerns the decomposition of rings into direct sums.
   (a) Assume first, that $A$ is isomorphic to $B \oplus C$, and show that there exist ideals $I$ and $J$ of $A$ such that $I \cap J = \{0\}$, and $I + J = \{i + j : i \in I, j \in J\} = A$.
   (b) Assume now that there exist ideals $I$ and $J$ of $A$ such that $I \cap J = \{0\}$, and $I + J = A$. Show that $A$ is isomorphic to $I \oplus J$ as follows:
      (i) Show that, if $i \in I$, $j \in J$ and $i + j = 0$, then $i = j = 0$.
      (ii) Show that, if $i \in I$ and $j \in J$, then $ij = 0$.
      (iii) Show that the map $\phi : I \oplus J \to A$, given by $\phi(i, j) = i + j$, is an isomorphism of rings.
1.7 The Quaternions

The field of complex numbers is an extension of the field of real numbers. This last field is an extension of the field of rationals, which in turn is an extension of the ring of integers. It is natural to ask if one can continue this sequence, if there is an extension of the field of complex numbers, or to what extend this process of determining successive extensions has a “natural” end. In the XIX Century, W. R. Hamilton \textsuperscript{18} asked and tried to solve this question.

In his first attempt (which lasted 20 years!), Hamilton looked for a field consisting of “numbers” of the form \(a + ib + cj\), where \(a, b, c \in \mathbb{R}\) and \(i\) is the imaginary unit. In other words, in modern language, he tried to look for a field structure in the set \(\mathbb{R}^3\) containing a subfield isomorphic to the field of complex numbers. After many attempts to find a “reasonable” value for the product \(ij\) (in the form \(ij = a + bi + cj\)), he realized he needed to introduce an additional “number” \(k\), so that \(ij = k\). Hamilton eventually discovered the existence not of a field, but of a division ring structure on the set \(\mathbb{R}^4\). The elements of this division ring are now called quaternions, or Hamilton’s numbers.

We denote the elements of the canonical basis of the vector space \(\mathbb{R}^4\) by

\[
1 = (1, 0, 0, 0), \quad i = (0, 1, 0, 0), \quad j = (0, 0, 1, 0), \quad k = (0, 0, 0, 1).
\]

\textsuperscript{18}William Rowan Hamilton (1805–1865), was a great Irish astronomer and mathematician. Hamilton like Gauss was very precocious: at the age of 5 he could read Greek Hebrew and Latin, and at the age of 10 he was familiar with half a dozen oriental languages!
Hence, we can write a quaternion $q = (a, b, c, d)$ as
\[ q = a1 + bi + cj + dk, \]
where $a, b, c, d$ are real numbers. We look for a ring structure on $\mathbb{R}^4$ for which the injective maps $\phi: \mathbb{R} \to \mathbb{R}^4$ and $\psi: \mathbb{C} \to \mathbb{R}^4$ defined by $\phi(x) = x1$ and $\psi(x + iy) = x1 + yi$, are ring homomorphisms, so that we can identify the real numbers $x \in \mathbb{R}$ with the quaternions $\{(x, 0, 0, 0) : x \in \mathbb{R}\}$, and the complex numbers $x + iy \in \mathbb{C}$ with the quaternions $\{(x, y, 0, 0) : x, y \in \mathbb{R}\}$.

Given a quaternion $q = a1 + bi + cj + dk$, we call $a1$ the real part of $q$, and $bi + cj + dk$ vectorial part of $q$. As in the case of complex numbers, we will usually write $q = a + bi + cj + dk$, so that the quaternion 1 is not included in the notation. The addition of quaternions is the usual vector addition in $\mathbb{R}^4$. Hence, it is obvious that, if $x$ and $y$ are reals and $z$ and $w$ are complex numbers, then
\[ \phi(x + y) = \phi(x) + \phi(y), \quad \psi(z + w) = \psi(z) + \psi(w). \]

The product of quaternions is harder to discover. Let us observe first that, if we think of the product of a real number $a$ by the quaternion $q$ as the product $(a1)q$, then this product should amount to the usual product of a scalar by a vector, and in particular the quaternions form a vector space of dimension 4 over the reals. On the other hand, we should also have
\[ (1.7.1) \quad i^2 = -1, \]
since
\[ i^2 = \psi(i)\psi(i) = \psi(i^2) = \psi(-1) = \phi(-1) = -\phi(1) = -1. \]

Hamilton found the fundamental identities:
\[ (1.7.2) \quad ij = k, \quad jk = i, \quad ki = j. \]

Using these identities, the associativity of the product, and the relation $i^2 = -1$, it is possible to determine the products $ji, kj, ik, j^2$ and $k^2$. For example,
\[ ij = k \Rightarrow i(ij) = ik \Rightarrow (ii)j = ik \Rightarrow (1)j = ik \Rightarrow -j = ik. \]

We leave the details of the other products as an exercise. The final result is:
\[ (1.7.3) \quad j^2 = k^2 = -1, \quad j i = -k, \quad k j = -i, \quad i k = -j. \]
Observe that the product of quaternions is not commutative, so the quaternions do not form a field. From the identities (1.7.1), (1.7.2), and (1.7.3), and using only the associative and distributive properties, it is possible to find the product of two arbitrary quaternions. Instead, we choose to invert the whole process and to define formally the ring of quaternions as follows.

**Theorem 1.7.1.** The set \( \mathbb{R}^4 \) with the usual vector addition and the product of \( q = a + v \) and \( r = b + w \) defined by

\[
(a + v)(b + w) = ab - (v \cdot w) + aw + bv + v \times w,
\]

is a division ring which will be denoted by \( \mathbb{H} \).

**Proof.** We start by observing that the product \((q, r) \rightarrow qr \) defined by (1.7.4) coincides with the product by scalars if \( q \in \mathbb{R} \), and also that it is \( \mathbb{R} \)-bilinear: given \( a_1, a_2 \in \mathbb{R}, q_1, q_2 \in \mathbb{R}^4 \) and \( r_1, r_2, r \in \mathbb{R}^4 \) we have

\[
(a_1q_1 + a_2q_2)r = a_1(q_1r) + a_2(q_2r), \quad q(a_1r_1 + a_2r_2) = a_1(qr_1) + a_2(qr_2).
\]

We check also that, with the notation \( i, j \) and \( k \) for the canonical basis of \( \mathbb{R}^3 \), the identities (1.7.1), (1.7.2), and (1.7.3) hold.

In order to check that associativity holds, we now use the bilinearity and the identities (1.7.1), (1.7.2), and (1.7.3), to find the products

\[
(a_0 + a_1i + a_2j + a_3k)(((b_0 + b_1i + b_2j + b_3k)(c_0 + c_1i + c_2j + c_3k))
\]

and

\[
((a_0 + a_1i + a_2j + a_3k)(b_0 + b_1i + b_2j + b_3k))(c_0 + c_1i + c_2j + c_3k),
\]

and verify that they coincide.

Finally, it is clear that

\[
q1 = 1q = q,
\]

and we leave as an exercise to check that for any non-zero quaternion \( q = (a, b, c, d) \) the following identities hold:

\[
qq' = q'q = 1, \quad \text{where } q' = \frac{a - bi - cj - dk}{a^2 + b^2 + c^2 + d^2}.
\]

\[19\]In this expression, \((v \cdot w)\) and \(v \times w\) denote the usual inner and cross products. Actually, these operations and the notation \( i, j \) and \( k \) for the canonical basis of \( \mathbb{R}^3 \) are traces that remain from Hamilton’s original work.
An interesting fact is that the ring of quaternions $\mathbb{H}$ is isomorphic to a subring of the ring $M_4(\mathbb{R})$ of $4 \times 4$ matrices with real entries. This gives concrete realization of this division ring and a different proof of Theorem 1.7.1. For that, consider the $2 \times 2$ matrices:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad 0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$  

These allow to define a linear transformation $\rho : \mathbb{R}^4 \rightarrow M_4(\mathbb{R})$ by setting on the canonical basis:

$$\rho(i) = \begin{pmatrix} N & 0 \\ 0 & -N \end{pmatrix}, \quad \rho(j) = \begin{pmatrix} 0 & M \\ -M & 0 \end{pmatrix}, \quad \rho(k) = \begin{pmatrix} 0 & N \\ N & 0 \end{pmatrix},$$

and

$$\rho(1) = \begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix}.$$  

Then, we have:

**Proposition 1.7.2.** The map $\rho : \mathbb{H} \rightarrow M_4(\mathbb{R})$ is an injective homomorphism, so the ring $M_4(\mathbb{R})$ has a subring isomorphic to the division ring $\mathbb{H}$.

The proof of this proposition is a simple computation. We will see later that as a consequence of the Fundamental Theorem of Algebra ("Every polynomial with complex coefficients of positive degree has one complex root") there is no field which is an extension of $\mathbb{C}$ and at the same time a vector space of finite dimension over $\mathbb{R}$ or $\mathbb{C}$. In other words, we now know that the original problem of Hamilton does not have a solution.

**Exercises.**

1. Using formula (1.7.4) for the product of quaternions show that relations (1.7.1), (1.7.2) and (1.7.3) hold.

2. Using only (1.7.4), bilinearity and the identities (1.7.1), (1.7.2), and (1.7.3), prove formula (1.7.5) for the inverse of a quaternion $q \neq 0$.

3. For any quaternion $q = a1 + bi + cj + dk$ one defines its conjugate to be the quaternion $\overline{q} = a1 - bi - cj - dk$. Show that:

   (a) The map $\phi : \mathbb{H} \rightarrow \mathbb{H}$ defined by $\phi(q) = \overline{q}$ is an automorphism of $(\mathbb{H}, +)$. What can you say about $\phi(q_1q_2)$?

   (b) $qq = ||q||^2$ where $||q|| = \sqrt{a^2 + b^2 + c^2 + d^2}$ denotes the norm of the quaternion $q = a1 + bi + cj + dk$. 

(c) The inverse of the quaternion $q$ is the quaternion $q^{-1} = \frac{\overline{q}}{|q|^2}$.

4. Check that the set formed by all linear combinations of the $4 \times 4$ matrices:
$$
\begin{pmatrix}
M & 0 \\
0 & M
\end{pmatrix},
\begin{pmatrix}
N & 0 \\
0 & -N
\end{pmatrix},
\begin{pmatrix}
0 & M \\
-M & 0
\end{pmatrix},
\begin{pmatrix}
0 & N \\
N & 0
\end{pmatrix},
$$
is a subring of $M_4(\mathbb{R})$.

5. Prove Proposition 1.7.2 i.e., show that the map $\rho : \mathbb{H} \to M_4(\mathbb{R})$ is an injective homomorphism.

6. Find all solutions of the equation $x^2 = -1$ in the ring $\mathbb{H}$ of quaternions.

7. Assume that $\{1\} \subset K \subset L \subset M$ are fields, with $M$ an extension of $L$, and $L$ an extension of $K$. Recall that $M$ is a vector space over $K$ and over $L$, and that, in turn, $L$ is a vector space over $K$. Assume that the dimension of $M$ over $L$ is $m$, and that the dimension of $L$ over $K$ is $n$. Show that the dimension of $M$ over $K$ is $mn$. Conclude that a non-trivial extension of $\mathbb{C}$ has at least dimension 4 over $\mathbb{R}$.

8. Consider the quaternions of the form $a + bi + cj + dk$, where $a, b, c, d \in \mathbb{Z}$. Verify that these quaternions form a non-abelian ring, which is not a division ring, but where the cancellation law holds.

9. Verify that the set formed by all the invertible elements of the ring in the previous exercise is a non-abelian group with 8 elements, which will be denoted by $\mathbb{H}_8$. Determine all the subgroups of $\mathbb{H}_8$, and identify the normal subgroups.

1.8 Symmetries

In this section we will see how group theory can be used to formalize the notion of symmetry. We will focus mainly on symmetries of planar figures. For that, we start by observing that a “plane figure” is formally a set $\Omega \subset \mathbb{R}^2$, and we call a SYMMETRY of $\Omega$ any map $f : \mathbb{R}^2 \to \mathbb{R}^2$ which preserves distances between points of $\mathbb{R}^2$, i.e., such that:
$$
||f(x) - f(y)|| = ||x - y||, \quad \forall x, y \in \mathbb{R}^2,
$$
and which transforms the set $\Omega$ in itself, i.e., such that:
$$
f(\Omega) = \Omega.$$
Example 1.8.1.

If \( \Omega \) is the unit circle centered at the origin, any planar rotation around the origin is a symmetry of \( \Omega \). Similarly, any reflection of the plane relative to a line through the origin is also a symmetry of \( \Omega \).

The symmetries of the plane or, more generally, the symmetries of \( \mathbb{R}^n \), are the maps \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) which preserve distances, and which for that reason are called isometries.

Examples 1.8.2.

1. Any translation is an isometry of the plane.
2. Any rotation is an isometry of the plane.
3. Any reflection (relative to a line or a point) is an isometry of the plane.

Our next aim is to classify all the isometries of \( \mathbb{R}^n \). For that, we look first at the isometries \( f \) which fix the origin, i.e., such that \( f(0) = 0 \).

Proposition 1.8.3. For a map \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) such that \( f(0) = 0 \), the following statements are equivalent:

(i) \( f \) is an isometry, i.e.,
\[
||f(x) - f(y)|| = ||x - y||, \forall x, y \in \mathbb{R}^n;
\]

(ii) \( f \) preserves inner products, i.e.,
\[
f(x) \cdot f(y) = x \cdot y, \forall x, y \in \mathbb{R}^n.
\]

Proof. We assume first that \( f \) is an isometry, and we observe that
\[
||x|| = ||x - 0|| = ||f(x) - f(0)|| = ||f(x)||.
\]

Moreover, we note that
\[
||f(x) - f(y)||^2 = ||f(x)||^2 + ||f(y)||^2 - 2 f(x) \cdot f(y),
\]
\[
||x - y||^2 = ||x||^2 + ||y||^2 - 2 x \cdot y.
\]

Since by assumption \( ||f(x) - f(y)|| = ||x - y|| \), and we have already shown that \( ||x|| = ||f(x)|| \), it follows immediately that \( f(x) \cdot f(y) = x \cdot y \), for all \( x, y \in \mathbb{R}^n \). We conclude that (i) implies (ii).

We leave as an exercise the proof that (ii) implies (i). \( \square \)

---

20In this section we denote by \( \cdot \) the usual inner product of \( \mathbb{R}^n \).
Still considering only isometries fixing the origin, we show next that such an isometry is necessarily a linear transformation.

**Proposition 1.8.4.** If \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is an isometry and \( f(0) = 0 \), then \( f \) is a linear transformation.

**Proof.** Denote by \( \{e_1, \ldots, e_n\} \) the canonical basis of \( \mathbb{R}^n \) and let \( v_k = f(e_k) \).

The vectors \( v_k \) are unitary (because \( ||v_k|| = ||f(e_k)|| = ||e_k|| = 1 \)) and orthogonal (because \( v_i \cdot v_j = f(e_i) \cdot f(e_j) = e_i \cdot e_j \)). Hence, the vectors \( \{v_1, \ldots, v_n\} \) also form a basis of \( \mathbb{R}^n \) (why?).

Let now \( x, y \in \mathbb{R}^n \), where \( y = f(x) \). Since \( \{e_1, \ldots, e_n\} \) and \( \{v_1, \ldots, v_n\} \) are both bases of \( \mathbb{R}^n \), there exist scalars \( x_1, \ldots, x_n \) and \( y_1, \ldots, y_n \) such that

\[
x = \sum_{k=1}^{n} x_k e_k, \quad y = \sum_{k=1}^{n} y_k v_k.
\]

Notice that \( x_k = x \cdot e_k \) and \( y_k = y \cdot v_k \), and since

\[
y \cdot v_k = f(x) \cdot f(e_k) = x \cdot e_k,
\]

we must have \( x_k = y_k \). Therefore:

\[
f(x) = f(\sum_{k=1}^{n} x_k e_k) = \sum_{k=1}^{n} y_k v_k = \sum_{k=1}^{n} x_k f(e_k),
\]

so \( f \) is a linear transformation. \( \square \)

Since an isometry satisfying \( f(0) = 0 \) is a linear transformation, it is natural to characterize an isometry in terms of its matrix representation. For that, we recall that a \( n \times n \) matrix is called orthogonal if \( A^T A = I \), or equivalently \( A^{-1} = A^T \). Since \( \det(A^T) = \det(A) \), for any orthogonal matrix \( A \) we have

\[
[\det A]^2 = \det A^T \det A = \det(A^T A) = \det I = 1,
\]

so that \( \det A = \pm 1 \).

**Proposition 1.8.5.** If \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is any map, the following statements are equivalent:

(i) \( f \) is an isometry and \( f(0) = 0 \);

(ii) \( f \) is a linear transformation and the matrix of \( f \) relative to the canonical basis is orthogonal: \( f(x) = Ax \), with \( A^T A = I \).
1.8. SYMMETRIES

Proof. (i) ⇒ (ii). If $A$ is the matrix of $f$ relative to the canonical basis, then

$$A = (a_{ij})$$

where $v_j = \sum_{i=1}^{n} a_{ij} e_i$. Therefore the column $j$ of $A$ is formed

by the components of the vector $v_j$ in the canonical basis. Since the vectors

$v_j$ are unitary and orthogonal, we conclude that

$$v_j \cdot v_k = \sum_{i=1}^{n} a_{ij} a_{ik} = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k, \end{cases}$$

which means that $A^T A = I$. This shows that the matrix $A$ is orthogonal.

(ii) ⇒ (i). Exercise. \qed

The linear transformations which are isometries are usually called orthogonal transformations. We can now characterize an arbitrary isometry

of $\mathbb{R}^n$ as an orthogonal transformation followed by a translation, as follows:

**Theorem 1.8.6.** If $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, the following statements are equivalent:

(i) $f$ is an isometry,

(ii) there exists an orthogonal matrix $A$ and a vector $a \in \mathbb{R}^n$ such that

$$f(x) = Ax + a.$$  

Proof. (i) ⇒ (ii). Let $f$ be an isometry, and $a = f(0)$. The map $g(x) = f(x) - a$ satisfies $g(0) = 0$ and is an isometry:

$$||g(x) - g(y)|| = ||f(x) - f(y)|| = ||x - y||.$$  

According to Proposition 1.8.5, $g$ is an orthogonal transformation.

(ii) ⇒ (i). An orthogonal matrix $A$ defines an orthogonal transformation $g(x) = Ax$, which we know is an isometry. It is immediate to check that if $a \in \mathbb{R}^n$, then $f(x) = Ax + a$ is an isometry. \qed

The set of all isometries $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, with binary operation composition of isometries, form a group $E(n, \mathbb{R})$, called the symmetry group of $\mathbb{R}^n$ or Euclidean group (why is this a group?). Theorem 1.8.6 shows that an isometry $f$ can be identified with a pair $(A, a)$, formed by an orthogonal matrix $A$ and a vector $a \in \mathbb{R}^n$, so one often identifies the euclidean group with the set of all such pairs:

$$E(n, \mathbb{R}) = \{(A, a) : A \in M_n(\mathbb{R}), A^T A = I, a \in \mathbb{R}^n\}. $$
We leave it as an exercise to check that the composition of isometries corresponds to the following binary operations on pairs:

\[(A_1, a_1) \cdot (A_2, a_2) = (A_1 A_2, A_1 a_2 + a_1).\]

The euclidean group has various interesting subgroups. For example, the translations are the isometries of the form \(f(x) = x + a\) and they form a subgroup of \(E(n, \mathbb{R})\) isomorphic to \((\mathbb{R}^n, +)\). Under the identification of isometries with pairs, we have that \(\mathbb{R}^n \subset E(n, \mathbb{R})\) corresponds to the injective homomorphism: \(a \mapsto (I, a)\).

The orthogonal transformations \(f : \mathbb{R}^n \to \mathbb{R}^n\) also form a subgroup \(O(n, \mathbb{R}) \subset E(n, \mathbb{R})\), called the ORTHOGONAL GROUP. We can view this group as the group of orthogonal matrices:

\[O(n, \mathbb{R}) = \{A \in M_n(\mathbb{R}) : A^T A = I\}.\]

As we have observed above, the determinant of an orthogonal transformation \(f\) must be 1 or \(-1\), and the orthogonal transformations \(f\) of determinant 1 form a subgroup \(SO(n, \mathbb{R}) \subset O(n, \mathbb{R})\) called the SPECIAL ORTHOGONAL GROUP. The elements of \(SO(n, \mathbb{R})\) are called PROPER ROTATIONS or simply ROTATIONS. In terms of matrices:

\[SO(n, \mathbb{R}) = \{A \in M_n(\mathbb{R}) : A^T A = I, \ det A = 1\}.\]

An isometry \(f : \mathbb{R}^n \to \mathbb{R}^n\) which coincides with its own inverse, i.e., such that \(f \circ f = I\) is called a REFLECTION. An isometry \(f(x) = A x + a\) is a reflection if and only if \(A^2 = I\) and \(A a = -a\). Composition of reflections may fail to be a reflection, so reflections do not form a group.

More generally, if \(\Omega \subset \mathbb{R}^n\), then the isometries of \(\mathbb{R}^n\) which preserve \(\Omega\) form a subgroup \(G_\Omega \subset E(n, \mathbb{R})\), called the SYMMETRY GROUP of \(\Omega\):

\[G_\Omega = \{f : \mathbb{R}^n \to \mathbb{R}^n : f \text{ is an isometry and } f(\Omega) = \Omega\}.
\]

We can talk of the symmetries of \(\Omega\) which are translations, orthogonal transformations, rotations, reflections, etc.

**Examples 1.8.7.**

1. Let \(\Omega \subset \mathbb{R}^2\) be a rectangle centered at origin, with sides (of different lengths) parallel to the coordinate axes. The corresponding symmetry group \(G_\Omega\) has 4 elements: the identity, the reflections in the axes Ox and Oy, and the rotation of 180° around the origin (which is also a reflection relative to the origin). The symmetry group of the rectangle, often called the KLEIN GROUP, is isomorphic to the direct product \(\mathbb{Z}_2 \times \mathbb{Z}_2\).
1.8. SYMMETRIES

2. If $\Omega \subset \mathbb{R}^2$ is a regular polygon with $n$ sides centered at origin, the corresponding symmetry group $G_\Omega$ is called the dihedral group, and has $2n$ elements: the rotations of $2k\pi/n$ around the origin, the reflections relative to lines through the origin and the vertices, and the reflections relative to lines through the origin and which bisect the sides of the polygon. It is common to denote this group by the symbol $D_n$.

For example, for an equilateral triangle ($n = 3$), there are three rotational symmetries \{R, R^2, R^3 = I\} generated by a rotation $R$ of $2\pi/3$ around the origin. Their matrix representations relative to the canonical basis of $\mathbb{R}^2$ are:

$$R = \begin{pmatrix} -\frac{1}{\sqrt{3}} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad R^2 = \begin{pmatrix} -\frac{1}{\sqrt{3}} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}. $$

We have also three axes of symmetry giving rise to 3 reflections \{\sigma_1, \sigma_2, \sigma_3\}. If we choose the triangle so that its vertices are $1, e^{\frac{2\pi i}{3}},$ and $e^{\frac{4\pi i}{3}},$ then the matrix representations of the reflections relative to the canonical basis of $\mathbb{R}^2$ are:

$$\sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} -\frac{1}{\sqrt{3}} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} -\frac{1}{\sqrt{3}} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. $$
We leave as an exercise to check that the resulting symmetry group $D_3$ is isomorphic to the symmetric group $S_3$.

In the previous examples, all the figures are bounded. It is also interesting to study groups of symmetry of unbounded figures. Consider, for example, the following subset of the plane:

$$\Omega = \{ na + mb : n, m \in \mathbb{Z} \},$$

where $a, b \in \mathbb{R}^2$ are fixed, linearly independent vectors of the plane. Then $\Omega$ is a discrete set of points, which can be thought of as a model for a planar lattice of atoms that extend indefinite in the plane (e.g., a crystal).

![Figure 1.8.3: A planar lattice $\Omega$.](image)

The unbounded planar sets which exhibit symmetry are used extensively in the decoration of surfaces (tilings). Looking at various examples of decorations, one is lead to conjecture that such symmetric tilings are obtained by repetition of one the following figures: equilateral triangle, square, rectangle or hexagon\(^{21}\).

We will not determine the full symmetry group $G_{\Omega}$, but will focus instead on the simpler problem of determining which rotations can be symmetries of $\Omega$. The facts mentioned above suggest that if some rotation is a symmetry of $\Omega$, then it must be a rotation by $60^\circ$, $90^\circ$, $120^\circ$ or $180^\circ$ (besides, of course, the identity, which is also a rotation). To see that this is indeed the case, let $f$ be some rotation which is a symmetry of $\Omega$, and denote by $A$ its matrix

---

1.8. SYMMETRIES

Figure 1.8.4: Symmetries of unbounded planar figures.

representation relative to the canonical basis. We can always write $A$ in the form:

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. $$

Also, let $B$ be the matrix representation of $f$ relative to the basis $\{a, b\}$. In this basis all the elements of $\Omega$ have integer coordinates (indeed, the points of $\Omega$ are precisely the vectors of $\mathbb{R}^2$ whose components in the basis $\{a, b\}$ are integers). Hence, the matrix $B$ must have integer entries, since these entries represent the vectors $f(a)$ and $f(b)$, which are necessarily points of $\Omega$. We can then write:

$$B = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}, $$

where the $n_{ij}$ are integers.

Since $A$ and $B$ represent the same linear transformation relative to distinct basis, we see that $A$ and $B$ are similar matrices i.e., there exists a non-singular matrix $S$ such that $S^{-1}AS = B$. Now observe that $A$ and $B$ have the same trace (the sum of the diagonal elements):

$$2 \cos \theta = \text{tr}(A) = \text{tr}(ASS^{-1}) = \text{tr}(S^{-1}AS) = \text{tr}(B) = n_{11} + n_{22}, $$

where we used the fact that $\text{tr}(CD) = \text{tr}(DC)$, for any square matrices $C$ and $D$. In particular, $2 \cos \theta$ must be an integer. Since $-1 \leq \cos \theta \leq 1$, we
conclude that the possible values of \( \cos \theta \) are \(-1, -\frac{1}{2}, 0, \frac{1}{2}\) or \(1\), i.e., that \( \theta = 180^\circ, 120^\circ, 90^\circ, 60^\circ \) or \(0^\circ\), as claimed.

There are many other examples, with concrete important applications, where group theory plays fundamental role in understanding issues related with the idea of symmetry.

**Examples 1.8.8.**

1. By exploring the symmetries of certain unbounded planar regions one can determine the so-called crystallographic groups, which are used in the classification of the crystals that occur in Nature.

2. We will see in Chapter 7 that one can associate to each polynomial a group of symmetries, consisting of certain permutations of its roots, the so-called Galois group of the polynomial. The structure of the Galois group of the polynomial determines whether the roots of the polynomials can be obtained from its coefficients by expressions which only involve fractions and radicals. The application of group theory to Galois theory allows one to explain why there are no formulas to express the solutions of polynomial equations for degree greater or equal to 5.

3. One of the most basic and fruitful ideas in Physics in the principle of “material objectivity” or “material frame-indifference”. This principle expresses the idea that different observers, using different frames of reference, should describe the same physical phenomenon using the same laws of Physics. From a mathematical point of view, this forces the laws of Nature to have as symmetry groups the groups of transformations that relate observers in different frames of reference. For example, according to this principle, the laws of Classical Mechanics and the laws of Electromagnetism must have the same symmetry group. It was the careful and systematic use of this idea that led Albert Einstein to the discover the Theory of Relativity, for sure one of the greatest achievements of Science.

**Exercises.**

1. Assume that \( f : \mathbb{R}^2 \to \mathbb{R}^2 \) is a non-trivial isometry of the plane. Show that:
   
   (a) if \( f \) fixes two points \( a \) and \( b \) of the plane, then \( f \) is a reflection relative to the line determined by \( a \) and \( b \);
   
   (b) if \( f \) fixes a unique point \( a \), then \( f \) is a rotation around \( a \);
   
   (c) if \( f \) does not fix any point, then \( f \) is a translation, followed possibly by a rotation and/or a reflection.

2. Conclude the proof of Proposition 1.8.3

3. Conclude the proof of Proposition 1.8.5
4. Show that the Klein group is isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

5. Show that the following sets of transformations are groups:
   (a) The orthogonal transformations and proper rotations $f : \mathbb{R}^n \to \mathbb{R}^n$.
   (b) The isometries $f : \mathbb{R}^n \to \mathbb{R}^n$.
   (c) The symmetries of a figure $\Omega \subset \mathbb{R}^n$.

6. Show that the special orthogonal group $SO(n, \mathbb{R})$ is a normal subgroup of $O(n, \mathbb{R})$, but it is not a normal subgroup of $E(n, \mathbb{R})$.

7. Show that $D_3$ (the symmetry group of the equilateral triangle) is isomorphic to the symmetric group $S_3$.

8. Describe the the group $D_4$ (the symmetry group of a square).

9. The honeycomb in Figure 1.8.4 has rotational symmetries consisting of rotations of $60^\circ$ around the centers of the faces, as well as rotations of $120^\circ$ around the vertices. How can one modify this figure so that the only rotational symmetries are rotations of $120^\circ$?

10. Let $f : \mathbb{R}^3 \to \mathbb{R}^3$ be a rotation. Show that there exists a basis $\{v_1, v_2, v_3\}$ of $\mathbb{R}^3$ relative to which the matrix representation of $f$ is

$$
\begin{pmatrix}
1 & 0 & 0 \\
0 & \cos \theta & -\sin \theta \\
0 & \sin \theta & \cos \theta
\end{pmatrix}.
$$

One calls $v_1$ the axis of rotation and $\theta$ the angle of rotation of $f$. Can you generalize this result to dimensions $n > 3$?

11. Let $f : \mathbb{R}^3 \to \mathbb{R}^3$ be a reflection that fixes the origin. Show that there exists a basis $\{v_1, v_2, v_3\}$ of $\mathbb{R}^3$ relative to which the matrix representation of $f$ is one of the following:

$$
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & -1
\end{pmatrix}, \quad \begin{pmatrix}
1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & -1
\end{pmatrix}, \quad \begin{pmatrix}
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & -1
\end{pmatrix}
$$

(reflection through a plane, through a line, through the origin, respectively). What if $f(0) \neq 0$? Can you generalize this result to dimensions $n > 3$?

12. Let $f : \mathbb{R}^3 \to \mathbb{R}^3$ be an isometry. Show that $f$ can be factored as the composition of a rotation, followed by a reflection, followed by a translation. Is this factorization unique?
Chapter 2

The Integers

2.1 The Axioms of the Integers

We have mentioned (always informally) the ring of integers and some of the properties that these numbers satisfy. So far we have assumed that these properties are well-known and “obvious”. However, it is not possible to develop any mathematical theory in a precise way without a careful examination of its foundations. In particular, one must make a careful distinction between what are the axioms of the theory and what are the consequences of those axioms, i.e., the propositions and theorems.

We wish now to discuss which properties of the integers should be considered axioms and which are simply consequences of those axioms. Note, however, that our discussion will not be completely formal: we will continue to use notions and results from set theory in an informal way, without a rigorous formulation. The reason is that they fall outside the domain of Algebra. However, the reader may wish to look at the Appendix in this respect.

The choice of axioms that form the base of some theory is, up to some point, arbitrary. It is always possible to chose distinct, but logically equivalent axioms. Hence, the final choice is necessarily a consequence of subjective criteria related to such aesthetical aspects as elegance, brevity, or efficiency of presentation.

We choose to start with an axiom which fits easily in the previous discussion:

**Axiom I.** There exists an integral domain \( \mathbb{Z} \), whose elements are called integers.
The zero and the identity of \( \mathbb{Z} \) will be denoted by 0 and 1, respectively. It follows from the results in the previous chapter that a large number of elementary properties of the integers are a direct consequence of Axiom [1]. In particular, the cancellation laws for addition and product and the sign rules are valid in \( \mathbb{Z} \), as well as the property \( 0 = -0 \) (proved in the exercises). On the other hand, it should be clear that the this axiom does not characterize completely the integers; for example, it is impossible to decide based solely on this axiom if the property \( 1 \neq -1 \) holds or not, or to decide if the integers form a finite or a infinite set (why?). In order to complete this with other axioms, we will now look carefully at the properties of the natural numbers.

In an informal way, we may think of the natural numbers as integers which are obtained from 1 by “successive addition” of 1, or in other words, the numbers of the form:

\[
1, \ 2 = 1 + 1, \ 3 = 2 + 1, \ 4 = 3 + 1, \ldots
\]

In particular, denoting by \( \mathbb{N} \) the set of natural numbers, we should have:

\[
1 \in \mathbb{N}, \ \text{and} \ n \in \mathbb{N} \Rightarrow n + 1 \in \mathbb{N}.
\]

Taking a more general perspective, which will be useful later, we introduce the following definition.

**Definition 2.1.1.** If \( A \) is a unitary ring with identity 1, a subset \( B \subset A \) is called inductive if:

(i) \( 1 \in B \);

(ii) \( n \in B \Rightarrow n + 1 \in B \).

The ring \( A \) itself is obviously an inductive subset of \( A \). Hence, \( \mathbb{Z} \) is an inductive subset of itself, so \( \mathbb{N} \) is not the only inductive subset of \( \mathbb{Z} \). However, the heuristic description of \( \mathbb{N} \) above suggests another property of this set, namely that any inductive subset of \( \mathbb{Z} \) necessarily contains all the natural numbers. In other words, \( \mathbb{N} \) is the smallest inductive subset of \( \mathbb{Z} \).

Coming back to the general case of an arbitrary ring \( A \) with identity 1, we formalize these ideas as follows. Since the ring \( A \) itself is always an inductive set, the family of all inductive subsets de \( A \) is necessarily non-empty. Denoting by \( N(A) \) the intersection of all inductive subsets of \( A \), it is clear that \( N(A) \) is contained in any inductive subset of \( A \).

This remark deserves a name:
Theorem 2.1.2 (Principle of Finite Induction). Let A be a unitary ring. Then:

(i) if \( B \subset A \) is inductive, then \( N(A) \subset B \), and

(ii) if \( B \subset N(A) \) is inductive, then \( B = N(A) \).

The previous result becomes even more interesting when one observes that:

Proposition 2.1.3. If \( A \) is a unitary ring, then \( N(A) \) is an inductive subset of \( A \).

Proof. Since 1 belongs to all inductive subsets of \( A \), it is obvious that \( 1 \in N(A) \).

Assume now that \( a \in N(A) \) and \( B \) is any inductive subset of \( A \). Then \( a \in B \) (because \( N(A) \subset B \)) and \( a + 1 \in B \) (because \( B \) is inductive). Since \( B \) is arbitrary, it follows that \( a + 1 \) belongs to all inductive subsets of \( A \), i.e., \( a + 1 \in N(A) \). Hence, we conclude that \( N(A) \) is inductive.

The previous results show that \( N(A) \) is an inductive subset which is contained in any inductive subset of \( A \). For this reason we introduce:

Definition 2.1.4. For a unitary ring \( A \), the set \( N(A) \) is called the smallest inductive subset of \( A \). If \( A = \mathbb{Z} \) we denote this subset by \( \mathbb{N} \) instead of \( N(\mathbb{Z}) \), and call \( \mathbb{N} \) the set of natural numbers.

We will see later that the usual principle of mathematical induction is precisely Theorem 2.1.2 applied to the ring of integers, and we will identify all the possible sets \( N(A) \) (up to isomorphism). Note also that our previous (heuristic) description of \( \mathbb{N} \) also applies to the set \( N(A) \), i.e., this set consists of the elements of \( A \) obtained from the identity 1 by “successive addition” of 1 (this will be made more precise later).

Examples 2.1.5.

1. If \( A = \mathbb{C} \), then \( N(A) = \mathbb{N} \).
2. If \( A = M_n(\mathbb{C}) \), then \( N(A) = \{ mI : m \in \mathbb{N} \} \).
3. If \( A = \mathbb{Z}_2 \), then \( N(A) = \mathbb{Z}_2 \).

We know that the addition and the product of natural numbers are still natural numbers. We can now prove this statement as a special instance of a general statement valid for any ring:
Proposition 2.1.6. The subset $N(A) \subset A$ is closed relative to the addition and the product, i.e.,
\[ \forall a, b \in N(A), a + b \in N(A) \text{ and } ab \in N(A). \]

Proof. We prove the statement for addition, leaving the case of the product as an exercise. For that, fix $a \in N(A)$ and define $B_a \subset N(A)$ as the set of all elements $b \in N(A)$ such that $a + b \in N(A)$. We must show that $B_a = N(A)$, which will follow from Theorem 2.1.2 (ii) by showing that $B_a$ is inductive:

1. Since $N(A)$ is inductive, it is clear that $a + 1 \in N(A)$, and hence $1 \in B_a$.
2. If $b \in B_a$, then $a + b \in N(A)$ and so $(a + b) + 1 \in N(A)$, because $N(A)$ is inductive. Since $(a + b) + 1 = a + (b + 1)$, it follows that $b + 1 \in B_a$.

We conclude that $B_a$ is inductive. \qed

After this long digression, we return to the question of how to characterize axiomatically the ring of integers. Recall that the set of integers is usually described, informally, as consisting of the natural numbers, the symmetric of the natural numbers, and zero:

\[ Z = \{0, 1, -1, 2, -2, 3, -3, \ldots \}, \]

being understood that in this list there are no repetitions. Our next axiom makes this precise.

**Axiom II.** Every $m \in Z$ satisfies exactly one of the following:

\[ m = 0 \text{ or } m \in \mathbb{N} \text{ or } -m \in \mathbb{N}. \]

Note that for all the other examples of rings that we have seen before if we replace in the previous statement $Z$ by $A$ and $\mathbb{N}$ by $N(A)$, we obtained a false statement. It is convenient (and standard) to denote the set $\mathbb{N} \cup \{0\}$ by $\mathbb{N}_0$.

Axioms II and III will be the basis of our study of the integers. Eventually, we will show that the usual axioms for the rational and real numbers are logical consequences of these axioms for the integers. The question of knowing if this set of axioms is complete, i.e., if they allow to decide if any “reasonable” statement about the integers is either true or false, and consistent, in the sense that some statement cannot be proved to be both true and false, is a deep and delicate problem studied in Mathematical Logic, so it
2.1. THE AXIOMS OF THE INTEGERS

will not be discussed here. We mentioned only that an equivalent question was the subject of Hilbert’s second problem. The solution given by Kurt Gödel in 1930 was one of the most surprising and amazing achievements of 20th Century Mathematics. Gödel showed that these two attributes of the axiomatization of the integers (complete and consistent) are themselves inconsistent: any system of axioms for $\mathbb{Z}$ which is consistent must contain statements whose logical value (true or false) cannot be decided based solely on those axioms.

It may sound strange at first sight that someone would even ask such questions. Actually, the amazing fact is that their study had profound consequences in our lives. Gödel’s works were followed up by the English mathematician Alan Turing, who in 1936 come up with its own theory of a Universal Computing Machine, now known as a Turing Machine. His theoretical findings inspired the Hungarian mathematician John von Neumann, after he moved to the USA, in his efforts to improve the ENIAC (Electronic Numerical Integrator and Computer), one of the first digital computers. This machine and other similar machines, which were constructed in the years after the publication of Turing’s work, were the precursors of modern day computers and digital machines.

The efforts of these two mathematicians also had profound consequences in other ways. The Second World War started in 1939 and Turing played a fundamental role in the English war efforts, helping to decipher the German secret transmission codes with the help of another digital machine, called ENIGMA. Von Neumann on the other hand, with the help of the ENIAC,

---

1 David Hilbert (1862-1943), German mathematician, who was a professor in Göttingen. Hilbert’s communication to the International Congress of Mathematicians held in Paris (1900) included a list of 23 problems which he thought should be consider by mathematicians throughout the XX Century (see Bull. Am. Math. Soc., 2nd ser., vol. 8 (1901-02), pp. 437-79).

2 Kurt Gödel (1906-1978) was born in Austria and emigrated at a young age to the USA, where he became a member of the Institute for Advanced Study, in Princeton. When he solved Hilbert’s second problem he was only 24 years old.

3 Alan Turing (1912-1954) was an English mathematician who towards the end of his life also worked in theoretical biology. He was interested in the development of patterns and shapes in biological organisms (morphogenesis) and the role of the Fibonacci numbers in plant structures. His life is mixed with tragedy, since he was prosecuted and convicted for his sexual orientation, resulting in chemical castration and the removal of his security clearance. These events eventually led him to commit suicide.

4 John von Neumann (1903-1957), was born in Hungary, and lectured in Berlin and Hamburg, before emigrating to the USA. Together with Einstein, he was one of the first members of the Institute for Advanced Study, in Princeton. Von Neumann gave the first axiomatization of the notion of a Hilbert space, a basic notion of Functional Analysis which plays a fundamental role in Quantum Mechanics.
collaborated on the Manhattan project, which resulted in the construction of the first atomic bomb, and the destruction of the Japanese cities of Hiroshima and Nagasaki, leading to the end of the war.

Exercises.

1. Complete the proof of Theorem 2.1.2 and of Proposition 2.1.6.

2. Determine the smallest inductive subset of each of the rings with identity mentioned in Exercise 1.5.1.

3. For which of the rings in the previous exercise does an axiom similar to Axiom II hold if the expression “exactly one” is replaced by “one”?

4. What are the consequences of replacing in Axiom II the expression “exactly one” by “one”, and adding one of the following condition:
   
   (a) $0 \notin \mathbb{N}$;
   
   (b) $n \in \mathbb{N} \Rightarrow -n \notin \mathbb{N}$;
   
   (c) $n \notin \mathbb{N} \Rightarrow -n \in \mathbb{N}$.

5. A set $X$ is said to be infinite if there exists an injective map $\Psi : X \to X$ which is not surjective. Show that $\mathbb{N}$ is infinite.

6. Show that, if there exists a bijection $\Psi : X \to Y$, then $X$ is infinite if and only if $Y$ is infinite.

7. Show that, if $Y \subset X$ and $Y$ is infinite, then $X$ is also infinite. In particular, show that $\mathbb{Z}$ is infinite.

2.2 Inequalities and Ordered Rings

Some of the elementary properties of the integer, rational and real numbers, concern the manipulation of inequalities, i.e., the order relation which exists in these rings. In these section we will study order relations in a ring, in an abstract manner, looking for properties which are common to all ordered rings. We will understand why some rings can be ordered, while other cannot, and whether orders are unique or not. In the case of the integers, we will see that there are properties of the order which are specific to this ring,

\footnote{The notion of number of elements (or cardinal) of a set is discussed in the Appendix.}
and moreover, that they are all consequence of the axioms in the previous section.

The abstract notion of an order relation can be defined on any set without any reference whatsoever to some algebraic structure (see the Appendix). Here we are interested in a special kind of an order relation:

**Definition 2.2.1.** A binary relation “>” on a set $X$ is called a **strict total order relation** if:

(i) *Transitivity*: $\forall x, y, z \in X$, $x > y$ and $y > z \Rightarrow x > z$.

(ii) *Trichotomy*: $\forall x, y \in X$, exactly one of the following three cases hold: $x > y$ or $y > x$ or $x = y$.

We call the pair $(X, >)$ an **ordered set**.

Note that condition (ii) states that any two elements of an ordered set $(X, >)$ can be compared, which is not the case for a general partial order relation.

Given an ordered set $X$ with an order relation “>”, which one reads “greater than”, one can define, just like for the usual numbers, the relations “<”, “≥” and “≤” (read as usual):

- $a < b$ if and only if $b > a$;
- $a \geq b$ if and only if $a > b$ or $a = b$;
- $a \leq b$ if and only if $a < b$ or $a = b$.

We recall next some elementary concepts valid for any ordered set $(X, >)$.

**Definition 2.2.2.** Let $(X, >)$ be an ordered set. If $Y \subset X$ and $x \in X$, then:

(i) $x$ is called an **upper bound** (respectively **lower bound**) of $Y$ if $x \geq y$ (respectively $x \leq y$), for all $y \in Y$;

(ii) $Y$ is called **bounded above** (respectively **bounded below**) in $X$ if $Y$ has at least one upper bound (respectively lower bound) in $X$;

(iii) $Y$ is called **bounded** in $X$ if it is bounded above and below in $X$.

Using these notions, we also have the usual notions of maximum, minimum, supremum and infimum, for any ordered set $(X, >)$, which we also state for future reference.
Definition 2.2.3. Let \((X, >)\) be an ordered set. If \(Y \subset X\) and \(x \in X\), then:

(i) if \(x\) is an upper bound (respectively lower bound) of \(Y\) and \(x\) belongs to \(Y\), \(x\) is called the maximum (respectively minimum) of \(Y\), and we denote \(x\) by \(\max(Y)\) (respectively \(\min(Y)\));

(ii) if the minimum (respectively maximum) of the set of upper bounds of \(Y\) exists, it is called the supremum or least upper bound (respectively infimum or largest lower bound) of \(Y\), and it is denoted by \(\sup(Y)\) (respectively \(\inf(Y)\)).

It should be clear from the definition of an ordered set that the maximum/minimum of a subset (and hence also the supremum/infimum) if they exist they are unique. Hence, these definitions and notations make sense.

The usual notions of intervals, familiar for real numbers, can be defined for any ordered set \((X, >)\). For example, if \(a, b \in X\), then

\[ [a, +\infty] = \{ y \in X : y > a \}, \quad ]a, b] = \{ y \in X : a < y \leq b \}, \quad \ldots \]

These are also commonly denoted by \((a, +\infty)\) and \((a, b)\).

Examples 2.2.4.

1. Let \(X = \mathbb{R}\) with the usual order relation “\(>\)” and \(Y = ] - \infty, 0[\). In this case, \(Y\) has no lower bounds in \(\mathbb{R}\), and hence cannot have an infimum or a minimum. The set of upper bounds is the interval \([0, +\infty[\), which has the minimum \(0\) which does not belong to \(Y\). Obviously, \(\sup(Y) = 0\) and \(Y\) has no maximum.

2. Let \(X = \mathbb{R}^2\) and define a strict, total order “\(>\)” as follows:

\[(x_1, y_1) > (x_2, y_2) \quad \text{if and only if} \quad \begin{cases} y_1 > y_2, \text{ or} \\ y_1 = y_2 \text{ and } x_1 > x_2, \end{cases}\]

where \(>\) on the right side denotes the usual order in \(\mathbb{R}\). One sometimes call this ordered set \((X, >)\) the long line. Now, the interval \(Y = ](1, 0), (0, 2)[\) (draw a picture of this interval) has \(\max(Y) = (0, 2) = \sup(Y)\) and \(\inf(Y) = (1, 0)\), while \(Y\) has no minimum.

When the set \(X\) has some algebraic structure it is natural to consider order relations on \(X\) which, in some sense, respect the algebraic structure. In the case of a ring \(A\) we will demand that

\[ a > b \iff a - b > 0, \]
i.e., that \( a > b \) if and only if \( a - b \) is positive, and that addition and product of positive elements be positive elements. Hence, for a ring it is somewhat more convenient to describe an order relation in terms of the set of positive elements.

**Definition 2.2.5.** A ring \( A \) is called an **ordered ring** if there exists a subset \( A^+ \subset A \) satisfying:

(i) *Closeness relative to addition and product:* \( a + b, ab \in A^+, \forall a, b \in A^+ \).

(ii) *Trichotomy:* For all \( a \in A \) exactly one of the following three cases hold: \( a \in A^+ \) or \( a = 0 \) or \( -a \in A^+ \).

If \( A \) is an ordered ring, we define in \( A \) an order relation "\( > \)" by

\[
    a > b \iff a - b \in A^+
\]

Notice that this is indeed an order relation in \( A \): transitivity of "\( > \)" follows from the fact that \( A^+ \) is closed for addition, and the trichotomy of "\( > \)" is an immediate consequence of (ii).

Notice also that one has \( a > 0 \) if and only if \( a \in A^+ \), and therefore we say that \( A^+ \) is the set positive elements of the ordered ring \( A \). It also clear that \( a < 0 \) if and only if \( -a \in A^+ \), and hence we call the elements of the set \( A^- = \{ a \in A : -a \in A^+ \} \) the negative elements.

The following proposition lists the basic rules to handle inequalities and are valid in any ordered ring. It shows that a large number of properties of the order relation of the integers, rationals and reals coincide, because they are shared by all ordered rings.

**Proposition 2.2.6.** Let \( A \) be an ordered ring. For all \( a, b, c \in A \), the following properties hold:

(i) \( a > b \iff a + c > b + c \);

(ii) \( a > b \iff -a < -b \);

(iii) \( ab > 0 \iff (a > 0 \ and \ b > 0) \ or \ (a < 0 \ and \ b < 0) \);

(iv) \( ab < 0 \iff (a > 0 \ and \ b < 0) \ or \ (a < 0 \ and \ b > 0) \);

(v) \( ac > bc \iff (a > b \ and \ c > 0) \ or \ (a < b \ and \ c < 0) \);

(vi) \( a \neq 0 \iff a^2 = aa > 0 \).
Proof. The proof of these properties is almost immediate from the definition. We illustrate them with the proof of property (i), leaving the rest for the exercises. By Definition 2.2.5,

\[ a + c > b + c \iff (a + c) - (b + c) \in A^+ \]
\[ \iff a - b \in A^+ \iff a > b. \]

It should now be clear that the ring of integers can be ordered by setting \( Z^+ = \mathbb{N} \), and this corresponds to the usual order of the integers. The trichotomy property of \( \mathbb{N} \) is exactly Axiom II for the integers, and we have already checked that in any ring with identity the set \( N(A) \) is closed relative to addition and the product. Perhaps, more interesting is the fact that the usual order of the integers is the unique order in this ring. In order to show it, we will need one more result valid for any ordered ring \( A \) where \( 1 \neq 0 \).

**Theorem 2.2.7.** If \( A \) is an ordered ring with identity \( 1 \neq 0 \), then one must have \( N(A) \subset A^+ \).

**Proof.** We only need to check that \( A^+ \) is an inductive subset and apply the definition of \( N(A) \):

(a) Since \( 1 \neq 0 \) and \( 1 = (1)(1) = 1^2 \), it follows from property (vi) in Proposition 2.2.6 that \( 1 > 0 \), i.e., that \( 1 \in A^+ \);

(b) Since \( A^+ \) is closed relative to the addition,

\[ a \in A^+ \Rightarrow a + 1 \in A^+. \]

Hence, \( A^+ \) is inductive and we conclude that \( N(A) \subset A^+ \).

Obviously there exist rings \( A \) such that \( A^+ \neq N(A) \), i.e. such that \( N(A) \subsetneq A^+ \). This is the case, e.g., with the rings \( \mathbb{Q} \) and \( \mathbb{R} \). However, as we mentioned above, if \( A \) is the ring of integers, then we must have \( N(A) = A^+ \).

**Theorem 2.2.8.** The ring of integers \( \mathbb{Z} \) has only one order, namely the one obtained by setting \( Z^+ = \mathbb{N} \).

**Proof.** Assume that \( \mathbb{Z} \) is ordered, possibly differently from the usual order, and let \( Z^+ \) be the corresponding set of “positive” integers. By the previous theorem, we have that \( \mathbb{N} \subset Z^+ \). We claim that \( Z^+ \subset \mathbb{N} \), from which the result follows.

To prove the claim, if \( m \in Z^+ \), then \( m \neq 0 \) and \( -m \notin Z^+ \), according to trichotomy property in Definition 2.2.5. Since \( \mathbb{N} \subset Z^+ \), it follows that \( -m \notin \mathbb{N} \). Finally, by Axiom II we must have \( m \in \mathbb{N} \).
2.2. INEQUALITIES AND ORDERED RINGS

One should not conclude from the previous results that the order of an arbitrary ring is always unique! For example, we shall indicate in the exercises of this section a subring of \( \mathbb{R} \) that can be ordered in several distinct ways. On the other hand, there are many rings which cannot be ordered.

**Example 2.2.9.**

*If \( A = \mathbb{Z}_2 \) is the field with two elements, then \(-1 = 1\), so we see that trichotomy cannot be satisfied and \( \mathbb{Z}_2 \) cannot be ordered.*

As we have already pointed out, the results at the beginning of this section show that a large number of properties of the order of the integers is common to all other ordered rings. The previous result suggests that the properties specific to the order of \( \mathbb{Z} \) result from the fact that \( A^+ = N(A) \) for this ring. We illustrate this fact by showing that the positive integers are the integers which are larger than or equal to 1, a statement that is obviously false if we replace the word “integers” by “rationals” or “reals”.

**Proposition 2.2.10.** \( \mathbb{Z}^+ = \{m \in \mathbb{Z} : m \geq 1\} \).

**Proof.** Let \( S = \{m \in \mathbb{Z} : m \geq 1\} \). If \( m \in S \), then \( m \geq 1 \), and since \( 1 > 0 \), we have that \( m > 0 \). Hence, \( S \subseteq \mathbb{N} \).

It is obvious that \( 1 \in S \) and that

\[
1 > 0 \Rightarrow 1 + 1 > 1 + 0 = 1.
\]

Hence, if \( m \in S \), then \( m + 1 \geq 1 + 1 > 1 \), and so \( m + 1 \in S \), showing that \( S \) is inductive. By the Principle of Induction and by Theorem 2.2.8, we conclude that \( S = \mathbb{N} = \mathbb{Z}^+ \). \( \square \)

The statement in the previous proposition is equivalent to the fact that in the ring of integers one has \([0, +\infty[ = [1, +\infty[\), or also \([0, 1[ = \emptyset\). Note that in \( \mathbb{Q} \) and in \( \mathbb{R} \) the interval \([0, 1[\) is an infinite set. Other properties specific to the order of the integers are discussed in the exercises at the end of this section and the next one.

The notion of absolute value or modulus can be introduced without much difficulty in any ordered ring.

**Definition 2.2.11.** Let \( A \) be an ordered ring and \( a \in A \). The **absolute value or modulus** of \( a \) is the element \( |a| \in A \) defined by

\[
|a| = \max\{-a, a\}.
\]
CHAPTER 2. THE INTEGERS

It should be clear from this definition that the absolute value of a non-zero number $a \in A$ is a positive number $|a| \in A^+$. We list next some properties of the absolute value, which may be shown directly from this definition.

**Proposition 2.2.12.** For any $a, b \in A$, we have:

(i) $-|a| \leq a \leq |a|;$

(ii) $|a + b| \leq |a| + |b|;$

(iii) $|ab| = |a||b|.$

The proofs are left to the exercises.

**Exercises.**

1. Complete the proof of Proposition 2.2.6.

2. Show that, if $A$ is a ring with identity $1 \neq 0$ and there exists in $A$ an element $i$ such that $i^2 = -1$, then $A$ cannot be ordered.

3. Show that, if $m \in \mathbb{Z}$, then $[m, m + 1[= \emptyset$ in $\mathbb{Z}$.

4. Show that in $\mathbb{Z}$:

   (a) $m > n \iff m \geq n + 1$;
   
   (b) $mn = 1 \iff (m = n = 1)$ or $(m = n = -1)$;
   
   (c) $m = \sup S$ if and only if $m = \max S$;
   
   (d) $m = \inf S$ if and only if $m = \min S$.

5. Prove Proposition 2.2.12.

6. Show that in any ordered ring:

   (a) $|a| \leq |b| \iff -|b| \leq a \leq |b| \iff a^2 \leq b^2$;
   
   (b) $||a| - |b|| \leq |a - b|$.

7. Let $B$ be an ordered ring and $\phi : A \to B$ an isomorphism of rings. Show that $A$ can be ordered with $A^+ = \{a \in A : \phi(a) \in B^+\}$.

8. Let $A = \{m + n\sqrt{2} : m, n \in \mathbb{Z}\}$. Show that $A$ can be ordered with an order distinct from the one induced from $\mathbb{R}$.

   (HINT: Note that $\phi(m + n\sqrt{2}) = m - n\sqrt{2}$ is an automorphism of $A$ and use the previous exercise.)
2.3. **PRINCIPLE OF INDUCTION**

9. Show that, if an ordered ring $A$ is bounded above or bounded below, then $A = \{0\}$.

10. Show that any ordered ring $A \neq \{0\}$ is infinite.

11. Show that the cancellation law for the product holds in any ordered ring.

12. An ordered ring $A$ is called **archimedean** if for any $a, b \in A$ with $a \neq 0$ there exists $x \in A$ such that $ax > b$. Show that the ring of integers is archimedean.

---

**2.3 Principle of Induction**

According to the definition of $\mathbb{N}$, discussed in the previous section, the set of natural numbers satisfies the **Principle of (finite) Induction**:

**Theorem 2.3.1 (Principle of Induction).** If $S \subset \mathbb{Z}$ is an inductive set, then $\mathbb{N} \subset S$. In particular, if $S \subset \mathbb{N}$, then $S = \mathbb{N}$.

Traditionally, a proof “by induction” obeys the following scheme: given a proposition $P(n)$, depending on some natural $n \in \mathbb{N}$, we must (i) show that $P(1)$ is true, and (ii) show that if $P(n)$ is true, then $P(n + 1)$ is also true. The relationship between this procedure and the Principle of Induction can be easily explained by introducing the set

$$S = \{n \in \mathbb{N} : P(n) \text{ is true}\}.$$

Then we have that:

- $(P(1) \text{ is true}) \iff (1 \in S)$;
- $(P(n) \Rightarrow P(n + 1)) \iff (n \in S \Rightarrow n + 1 \in S)$.

Hence, we see that the usual induction argument amounts to show that the set $S$ of natural numbers for which the proposition is true is an inductive subset of $\mathbb{N}$. Often, one does not mention explicitly the set $S$, but that should not be a source of any misunderstanding.

**Example 2.3.2.**

We say that $n \in \mathbb{N}$ is even (respectively, odd) if there exists $k \in \mathbb{Z}$ such that $n = 2k$ (respectively, $n = 2k + 1$). In order to prove the statement “every natural number is even or odd”, we consider $P(n) =$ “$n$ even or odd”, and observe:
(i) If \( n = 1 \), then \( 1 = 2 \cdot 0 + 1 \), hence 1 is odd, so that \( P(1) \) is true.

(ii) If \( n \) is a natural number such that \( P(n) \) is true, we have \( n = 2k \) or \( n = 2k+1 \). It follows that that \( n+1 = 2k+1 \) or \( n+1 = (2k+1)+1 = 2(k+1) \), and hence \( P(n+1) \) is true.

We conclude that \( P(n) \) is true for any natural number \( n \).

We will often use induction in our proofs. We start illustrating its use by proving a few more properties of the order of the natural numbers and of the integers. Our first application is to prove the Well-ordering Principle.

**Theorem 2.3.3** (Well-ordering Principle). Any non-empty set of natural numbers has a minimum.

**Proof.** Let \( S \) be any non-empty set of natural numbers and consider the statement

\[
P(n) = \text{"If } S \text{ contains a natural } k \leq n, \text{ then } S \text{ has minimum".}
\]

Our theorem will follow (why?) if we can show that:

"\( P(n) \) is true for every \( n \in \mathbb{N} \)."

Therefore, we can try to apply the Principle of Induction:

(i) \( P(1) \) is true, because, if \( S \) contains a natural \( k \leq 1 \), then according to Proposition (2.2.10) we have \( k = 1 \), and 1 is obviously the minimum of \( S \), because it is the minimum of \( \mathbb{N} \).

(ii) Assume now that \( P(n) \) is true for some natural \( n \), and assume that \( S \) contains a natural \( k \leq n + 1 \), We must show that \( S \) has a minimum. If \( S \) contains a natural \( k \leq n \), then it follows from \( P(n) \) that \( S \) has a minimum. Otherwise, \( S \) contains a natural in the interval \([1, n + 1]\), but none in the interval \([1, n]\). Since we know that the interval \([n, n+1]\) is empty, we conclude that \( n + 1 \) is the minimum of \( S \).

On the other hand, when \( S \) is a set of integers, we have

**Theorem 2.3.4.** Any non-empty set of integers which is bounded below (respectively, bounded above) has a minimum (respectively, a maximum).
2.3. PRINCIPLE OF INDUCTION

We leave the proof of this theorem an exercise, and instead we prove the following theorem, which can be used in proofs by induction that “start” at any integer \( k \neq 1 \). One should notice that the proof uses Theorems 2.3.3 and 2.3.4 which occasionally are more practical to apply than the Principle of Induction.

**Proposition 2.3.5.** If a statement \( \mathcal{P}(n) \) is true for some integer \( n = k \) and if \( \mathcal{P}(n) \Rightarrow \mathcal{P}(n+1) \) for every integer \( n \geq k \), then \( \mathcal{P}(n) \) is true for every integer \( n \geq k \).

**Proof.** Let
\[
S = \{n \in \mathbb{Z} : n \geq k \text{ and } \mathcal{P}(n) \text{ is false}\}.
\]
We claim that \( S \) is empty, i.e., that \( \mathcal{P}(n) \) is true for every \( n \geq k \).

Note that \( S \) is, by definition, bounded below by \( k \). Hence, according to Theorem 2.3.3, if \( S \) is non-empty then it must have a minimum \( m \). Moreover, since \( m \in S \), it follows that \( m \geq k \).

By assumption \( \mathcal{P}(k) \) is true, and hence \( k \not\in S \), so that \( m > k \). Consider now the integer \( m - 1 \). Since \( m > k \), we have \( m - 1 \geq k \). Since \( m - 1 \) is less than the minimum of \( S \), we have \( m - 1 \not\in S \), i.e., \( \mathcal{P}(m - 1) \) is true. But by assumption \( \mathcal{P}(n) \Rightarrow \mathcal{P}(n + 1) \) for all \( n \geq k \) so that \( \mathcal{P}(m) \) is true. This contradicts the fact that \( m \) belongs to \( S \). We conclude that \( S \) cannot have minimum, and so it must be empty, proving our claim.

In certain circumstances it is more convenient to use yet another form of the Principle of Induction. For example, consider the sequence \( a_n \) consisting of the Fibonacci numbers:

\[
1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots
\]

This sequence is usually defined as the sequence \( a_n \) satisfying:

\[
(2.3.1) \quad a_0 = a_1 = 1, \text{ and } a_{n+1} = a_n + a_{n-1} \text{ for } n \geq 1.
\]

It is possible to determine an explicit expression for this sequence. We will not discuss now how to deduce it, but the expression is the following:

\[
(2.3.2) \quad a_n = \frac{5 + \sqrt{5}}{10} \left( \frac{1 + \sqrt{5}}{2} \right)^n + \frac{5 - \sqrt{5}}{10} \left( \frac{1 - \sqrt{5}}{2} \right)^n.
\]

---

^6Fibonacci, also known as Leonardo of Pisa (1180-1250), found this sequence by considering the following problem about rabbit breeding: Beginning with a single pair of rabbits, how many pairs of rabbits are there at the end of one year if each pair of rabbits gives rise to a new pair of rabbits every month, and each new pair starts giving birth at the end of 2 months?
It looks obvious that the sequence $a_n$ defined by this last formula is the only sequence that satisfies (2.3.1). However, the proof of this statement is not as easy as it looks at first sight. First of all it is convenient to formalize the notion of sequence in any set $X$.

**Definition 2.3.6.** A sequence of elements in a set $X$ is a map $f : \mathbb{N} \to X$.

When dealing with sequences it is common to write $f_n$ instead of $f(n)$, and to speak of the “sequence $\{f_1, f_2, \ldots \}$”, “$\{f_n\}$”, or even “$f_n$”, instead of saying “the sequence $f$”. We will also commit this abuse of language (why is this an abuse of language?).

Note that, if $k$ is an integer, then to a map $g : [k, \infty[ \cap \mathbb{Z} \to X$ corresponds the sequence $f : \mathbb{N} \to X$ given by $f(n) = g(n + k - 1)$. For this reason, we say that $g$ is a sequence defined in $[k, \infty[$. Hence, the “Fibonacci sequence” above is a sequence defined in $[0, \infty[.$

The result that we wish to prove is the following:

**Proposition 2.3.7.** If $f$ is a sequence of natural numbers defined in $\mathbb{N}_0$, such that $f_0 = f_1 = 1$ and $f_{n+1} = f_n + f_{n-1}$ for $n \geq 1$, then $f(n) = a(n)$, where $a(n) = a_n$ is the sequence defined by (2.3.2).

If we let $P(n)$ be the statement “$f(n) = a(n)$”, we must show $P(n)$ for all integers $n \geq 0$. The difficult in applying the usual induction method is that, for obvious reasons, we cannot show that $P(n) \Rightarrow P(n+1)$, but only that $(P(n) \land P(n-1)) \Rightarrow P(n+1)$. In order to facilitate the solution of the to this and similar problems, we introduce the following form of the Principle of Induction:

**Theorem 2.3.8.** Let $S$ be a set of integers such that

(i) $k \in S$, and

(ii) for all $n \geq k$, $[k, n] \subset S \Rightarrow [k, n+1] \subset S$.

Then $[k, +\infty[ \subset S$.

In terms of the statement $P(n) = \text{“} n \in S \text{”}$, the previous result says that, if $P(k)$ is true and if $P(n)$ is true whenever $P(m)$ is true for all $k \leq m < n$, (and not just for $m = n - 1$), then $P(n)$ is true for all $n \geq k$. The proof of this statement is not hard if we use the Well-ordering Principle and proceed as in the proof of Proposition 2.3.5.
2.3. PRINCIPLE OF INDUCTION

Proof of Theorem 2.3.8. Let

\[ F = \{n \geq k : n \not\in S \} \]

We claim that \( F \) is empty, so that \( [k, +\infty) \subset S \).

If \( F \neq \emptyset \), since \( F \) is bounded below by definition, we conclude that \( F \) has a minimum \( m \geq k \). Since \( k \in S \), it is clear that \( m > k \). Hence, \( m - 1 \geq k \), and the interval \([k, m - 1]\) does not contain any element of \( F \), because all its elements are less than the minimum of \( F \). In other words, we have \([k, m - 1] \subset S \). It follows now from (ii) that we must have \([k, m] \subset S \), and hence \( m \in S \), i.e., \( m \not\in F \), a contradiction.

We conclude that \( F \) cannot have minimum, hence \( F \) is empty, as claimed.

\[
\square
\]

The previous theorem allows for an immediate proof of Proposition 2.3.7, which we leave for the exercises.

Note, however, that this result does not eliminate all the logical difficulties with definitions like (2.3.1). This definition is an example of a recursive definition, where the object to be defined appears in the description of the object. As we know since at least back to Epimenides of Crete (circa 600 BC) who became famous with the phrase “All Cretans are liars”, it is possible to create logical paradoxes, or propositions whose logical value cannot be decided, by making statements which in some way refer to themselves. A classical example is Bertrand Russell’s paradox\footnote{Bertrand Russell (1872-1970) co-authored together with Alfred N. Whitehead (1861-1947) the famous treaty \textit{Principia Mathematica} (3 vols., 1910-13), where they attempted to formalize in an axiomatic manner all the fundamental notions of Arithmetics. This monumental work was the ultimate result of the “logicist” program to formalize Mathematics, which consisted in constructing all Mathematical knowledge through logic deduction from a small number of concepts and principles. Although this program was doomed to fail, due to the later work of Gödel, it was fundamental for the advance of Mathematical Logic.} suggested by the attempts to define the “set of all sets”.

Observe that a definition like “\( U \) is the set of all sets” is recursive, because \( U \), being a set, is one of the elements which enters its description. In other words, \( U \) has the strange property of being an element of itself, which is not at all usual in the sets we are familiar with! If this property of \( U \) is unpleasant, we can consider instead the set \( N \) of “normal” sets, i.e., the sets which are not elements of themselves. In symbols,

\[
N = \{C \in U : C \not\in C \}.
\]
The question is now: is \( N \) a “normal” set? Unfortunately, if we assume that \( N \) is “normal” \((i.e., \ N \notin N)\) then \( N \) belongs to the set of normal sets \((i.e., \ N \in N)!\). If we assume that \( N \) is not “normal”, we have that \( N \in N \). But then \( N \) is an element of the set of normal sets, and hence \( N \) is itself normal \((N \notin N)!\). In other words, we cannot give a logical value to the statement “\( N \in N \)”.

At a superficial level, the lesson to learn from this example is simply that one must be careful with recursive definitions, something that the Greek philosophers were already aware of. A similar difficulty occurs when one uses a spreadsheet and creates a closed loop between cells in the spreadsheet, or when one states a “theorem” such as:

**Theorem 2.3.9.** This statement is false…

It is more or less clear that such difficulties do not occur with definitions like the one for the Fibonacci sequence, and we will often use such recursive definitions (see the examples in the next section and the Appendix for a formalization).

However, at a deeper level, the logical difficulties with recursive definitions, or more generally with statements that refer to themselves, are unavoidable and are related with some of the deepest questions tackled by mathematicians and philosophers nowadays: Is it possible to give a (rigorous) definition of “rigorous definition”? Can we understand how our own intelligence works? How can we reconcile the mechanical aspect of logical deductions, exhibit by a computer program, with the apparent infinite adaptive nature of what we call “intelligent behavior”? At the end of the day, this is the central problem of Artificial Intelligence.

**Exercises.**

1. Determine all the inductive subsets of \( \mathbb{Z} \).

2. Prove Theorem 2.3.4.

3. Prove Proposition 2.3.7.

4. Consider the statement (obviously false!) “All blonde women have blue eyes”. What is the mistake in the following “proof” by induction? Denote by \( \mathcal{P}(n) \) the statement “If in a set of \( n \) blonde women there exists one with blue eyes, then all have blue eyes”. Then:

   (a) \( \mathcal{P}(1) \) is obviously true.
2.4. SUMMATIONS AND PRODUCTS

(b) Assume that \( P(n) \) is true, and consider a set with \( n + 1 \) blonde women \( L = \{M_1, \ldots, M_{n+1}\} \). Assume that \( M_1 \) has blue eyes. Define \( L_{n+1} = \{M_k : k \neq n+1\} \) and \( L_n = \{M_k : k \neq n\} \). Since \( P(n) \) is true, all women in \( L_n \) and \( L_{n+1} \) have blue eyes. Since \( L = L_n \cup L_{n+1} \), all women in \( L \) have blue eyes.

(c) Since there exists at least one blonde women with blue eyes, we conclude that all blonde women must have blue eyes.

5. Here is another variant of the “blonde women with blue eyes” problem. Consider the statement (obviously false!) “All finite groups are abelian”. We have the following “proof” by induction: Let \( G \) be a group and denote by \( P(n) \) the statement “In a subset of \( G \) with \( n \) elements, all elements commute”.

(a) \( P(1) \) is obviously true.

(b) Assume that \( P(n) \) is true, and consider a subset of \( G \) with \( n + 1 \) elements \( L = \{g_1, \ldots, g_{n+1}\} \). Denote by \( L_i = \{g_k : k \neq i\} \) the subset formed by all elements of \( L \) with the exception of the element \( i \). Since \( P(n) \) is true, each \( L_i \) is commutative. Since the \( L_i \) exhaust all the elements of \( L \), we conclude that \( L \) is commutative.

(c) Since \( G \) is finite, we conclude that \( G \) is commutative.

6. Formula (2.3.2) for the Fibonacci sequence can be obtained by determining the sequences of the form \( \beta^n \) which satisfy equation (2.3.1). What are the sequences of integers which satisfy \( b_{n+1} = b_n + 6b_{n-1} \) and \( b_0 = b_1 = 1 \)?

### 2.4 Summations and Products

Often one is interested in adding and multiplying more than two elements. For this reason, it is convenient to generalize these algebraic operations to an arbitrary (finite) number of factors. We start by looking at the definition of products of more than two factors, since the results for summations are obtained by a simple change of notation.

In this section, \( S \) will denote a set with an associative binary operation. Given a sequence \( a_1, a_2, \ldots \) of elements of \( S \), the corresponding sequence of partial products is \( \pi_1 = a_1, \pi_2 = a_1a_2, \pi_3 = (a_1a_2)a_3, \ldots \). Formally:

**Definition 2.4.1.** Given a sequence \( a : \mathbb{N} \to S \) its SEQUENCE OF PARTIAL PRODUCTS is the sequence \( \pi : \mathbb{N} \to S \) defined inductively by:

\[
\pi_1 = a_1 \quad \text{and} \quad \pi_n = \pi_{n-1}a_n \quad \text{if} \quad n > 1.
\]

We write \( \pi_n = \prod_{k=1}^n a_k \) and called it the **PRODUCT** of the \( a_k \)'s, with \( k \) from 1 to \( n \).
In terms of these sequences the associativity property reads:

**Proposition 2.4.2.** If \( a, b : \mathbb{N} \to S \) are sequences in \( S \), we have:

(i) for any \( n > m \):
\[
(\prod_{k=1}^{n} a_k) (\prod_{k=m+1}^{n} a_k) = \prod_{k=1}^{n} a_k;
\]

(ii) if the operation is commutative:
\[
(\prod_{k=1}^{n} a_k) (\prod_{k=1}^{n} b_k) = \prod_{k=1}^{n} (a_k b_k).
\]

The proof (by induction) is left for the exercises. Also, note that the definition above can be easily adapted to products that start at some \( k > 1 \).

A particular interesting case of Definition 2.4.1 happens when the sequence \( a : \mathbb{N} \to S \) is constant (i.e., when \( a_n = c \), for all \( n \in \mathbb{N} \)). The product of the first \( n \) terms of the sequence \( a \) corresponds then to the notion of power with base \( c \) and exponent \( n \).

**Definition 2.4.3.** The power of base \( c \) and exponent \( n \in \mathbb{N} \) is defined by:
\[
c^n = \prod_{k=1}^{n} c.
\]

The power can be thought of as map \( \Phi : \mathbb{N} \times S \to S, (n, c) \to c^n \). If we fix \( c \), we have the exponential map \( \phi : \mathbb{N} \to S \), while if we fix \( n \), we obtain a map \( \psi : S \to S \). According to Definitions 2.4.1 and 2.4.3 we have:

(2.4.1)
\[
c^1 = c \text{ and } c^n = (c^{n-1})^n \text{ for } n > 1.
\]

When \( S \) is a monoid with identity \( I \) and \( c \) is invertible, we define:

(2.4.2)
\[
c^0 = I \text{ and } c^{-n} = (c^{-1})^n \text{ for } n > 1.
\]

Hence, for \( c \) is invertible, the power \( c^m \) is well-defined for any integer \( m \). In particular, when \( S \) is a group, the map \( \Phi : \mathbb{N} \times S \to S \) can be extended to a map \( \Phi : \mathbb{Z} \times S \to S \).

The usual rules for powers are valid in this more general context:

**Proposition 2.4.4.** If \( S \) has an associative operation, then for any \( a, b \in S \), one has:

(i) \( a^n a^m = a^{n+m} \) and \( (a^n)^m = a^{nm} \), for \( n, m \in \mathbb{N} \);

(ii) If \( ab = ba \), then \( (ab)^n = a^n b^n \), for \( n \in \mathbb{N} \);

If \( S \) is a monoid, then for any \( a, b \in S \) invertible, one has:

(iii) \( a^n a^m = a^{n+m} \) and \( (a^n)^m = a^{nm} \), for \( n, m \in \mathbb{Z} \);

(iv) If \( ab = ba \), then \( (ab)^n = a^n b^n \), for \( n \in \mathbb{Z} \).
2.4. SUMMATIONS AND PRODUCTS

Proof. As one could guess the proof of these results uses the Principle of Induction. We illustrate it in the case of (i), as an example where the argument involves two natural numbers. The proofs of the other cases are left for the exercises.

We prove that \(a^n a^m = a^{n+m}\) by induction in the natural number \(m\). Let \(\mathcal{P}(m)\) be the statement:

\[\mathcal{P}(m) = \text{"}a^n a^m = a^{n+m}\text{ for any } a \in S \text{ and natural number } n\text{".}\]

\(\mathcal{P}(1)\) follows from Definition 2.4.1, and to prove \(\mathcal{P}(m) \Rightarrow \mathcal{P}(m + 1)\), we observe that

\[
\begin{align*}
a^n a^{m+1} &= (a^n)(a^m a) & \text{(by 2.4.1)}, \\
&= ((a^n)(a^m))a & \text{(by associativity)}, \\
&= (a^{n+m})a & \text{(by the induction hypothesis)}, \\
&= a^{n+m+1} & \text{(by 2.4.1)}.
\end{align*}
\]

Notice that according to (iii) in the previous proposition, and assuming that \(S\) is a group, the map \(f : \mathbb{Z} \to S, n \mapsto a^n\), for some fixed \(a \in S\), is a homomorphism of groups. If \(S\) is just a monoid, the restriction of this function to \(\mathbb{N}_0\) is still a homomorphism of monoids.

Let us now formulate the results above in the additive notation. If “+” denotes a commutative binary operation on the set \(S\), one should restate Definition 2.4.1 as follows:

**Definition 2.4.5.** Let + be a commutative, associative, binary operation on the set \(S\). Given a sequence \(a : \mathbb{N} \to S\) its sequence of partial sums is the sequence \(\sigma : \mathbb{N} \to S\) defined inductively by:

\[
\sigma_1 = a_1 \text{ and } \sigma_n = \sigma_{n-1} + a_n \text{ if } n > 1.
\]

One calls \(\sigma_n\) the summation of the \(a_k\)’s, with \(k\) from 1 to \(n\), and denotes it by \(\sum_{k=1}^n a_k\).

When \(a : \mathbb{N} \to S\) is a constant sequence with \(a_n = c\), then we write \(nc = \sum_{k=1}^n c\). Moreover, if \(S\) has identity \(0\) and the element \(c\) has a symmetric \(-c\), we define

\[
0c = 0, \quad (-n)c = n(-c).
\]

The operation \(\Phi : \mathbb{N} \times S \to S, (n,c) \mapsto nc\) is called the “product of a natural \(n\) by an element \(c\) of \(S\)”. Also when \(c\) has a symmetric, we call
the operation $\Psi : \mathbb{Z} \times S \to S$, $(n,c) \mapsto nc$, the “product of an integer $n$ by an element $c$ of $S$”. This terminology is slightly ambiguous when $S = \mathbb{Z}$: apparently, we obtain in $\mathbb{Z}$ two distinct product operations, namely the operation mentioned in Axiom 2.4.3 and the operation we have just introduced in Definition 2.4.5. We leave as simple exercise to check that these two operations actually coincide. In fact, this duplication suggests that all the references to the product in the axiomatization of the integers are superfluous and redundant. Though we have not chosen this path, this is indeed the case: it is possible to present a set of axioms for the integers without mentioning the product operation, and to deduce all the properties of the product as theorems.

If $(G,+)$ is an abelian group, the basic algebraic properties of the product of elements of $G$ by integers can be summarized as follows:

**Proposition 2.4.6.** Given an abelian group $(G,+)$, if $g,h \in G$ and $m,n \in \mathbb{Z}$, then we have:

(i) Identity: $1g = g$.

(ii) Distributivity: $(n + m)g = ng + mg$ and $n(g + h) = ng + nh$.

(iii) Associativity: $n(mg) = (nm)g$.

One may notice that these properties are formally similar to the properties in the definition of a linear vector space. More precisely, if we replace the elements of the group $G$ by vectors of a linear vector space $V$ over some field and the integers by scalars of the field, then the properties in Proposition 2.4.6 are exactly the conditions on the operation “product of scalar by a vector” in the definition of a vector space. In fact, we will see in Chapter 6 the more general concept of a module over a ring, which allows to treat at the same level both the notion of an abelian group and the notion of a vector space over a field.

If $(A,+,\cdot)$ is a ring, we can also look at some “mixed” properties combining the addition and the product:

**Proposition 2.4.7.** Let $(A,+,\cdot)$ be a ring. If $a,a_1,\ldots,a_n,b,c \in A$ and $n \in \mathbb{N}$, then we have:

(i) $c(\sum_{k=1}^{n} a_k) = \sum_{k=1}^{n} (ca_k)$;

(ii) $(\sum_{k=1}^{n} a_k) c = \sum_{k=1}^{n} (a_k c)$;

(iii) $n(ab) = (na)b = a(nb)$. 

2.4. SUMMATIONS AND PRODUCTS

We mentioned before that, when $G$ is a group and $g \in G$, then the map
\[ \psi : \mathbb{Z} \to G, \quad n \mapsto g^n \]
is a homomorphism of groups. Obviously, if $G$ is an abelian group, the map $\psi : \mathbb{Z} \to G, \quad n \mapsto ng$ is also a homomorphism of groups. Finally, if $(A, +, \cdot)$ is a ring and $a \in A$, then $\psi : \mathbb{Z} \to A, \quad n \mapsto na$ is always a homomorphism of groups between $(\mathbb{Z}, +)$ and $(A, +)$, which in general fails to be a ring homomorphism, except when $a^2 = a$ (why?).

Exercises.

1. What is the sequence defined in $\mathbb{Z}$ by $a_1 = 1, a_{n+1} = \sum_{k=1}^{n} a_k$?

2. Complete the proofs of the results stated in this section.

3. Show that, if $S$ is a monoid where the cancellation law holds, then the identity
\[ \left( \prod_{k=1}^{n} a_k \right) \left( \prod_{k=1}^{n} b_k \right) = \prod_{k=1}^{n} (a_k b_k) \]
holds if and only if $S$ is abelian.

4. Let $G$ be an abelian group, and assume that $n \in \mathbb{Z}$, and $g_1, g_2 \in G$. Is it true that one always has
\[ n \neq 0 \text{ and } ng_1 = ng_2 \Rightarrow g_1 = g_2? \]
(HINT: Consider the group $\mathbb{Z}_2$.)

5. Show that, if $B$ is a subset of the ring $A$ which is closed relative to the difference in $A$, then $B$ is also closed relative to the product by integers, i.e.,
that:
\[ \text{If } [a, b \in B \Rightarrow a - b \in B] \text{ then } [(n \in \mathbb{Z} \text{ and } b \in B) \Rightarrow nb \in B]. \]

6. Use the previous result to prove that in the ring of integers, if $B \subset \mathbb{Z}$ is non-empty, the following statements are equivalent:

   (a) $B$ is closed relative to the difference;
   (b) $B$ is a subring of $\mathbb{Z}$;
   (c) $B$ is an ideal of $\mathbb{Z}$.

7. Show that:
   (a) if $\phi : G \to H$ is a homomorphism of additive groups, then $\phi(ng) = n\phi(g)$, for all $n \in \mathbb{Z}, g \in G$;
(b) if \( \phi : \mathbb{Z} \to G \) is a homomorphism of additive groups, then \( \phi(n) = ng \), for some \( g \in G \).

How can one generalize these results for non-additive groups?

8. Show that, if \( G \) is a group and \( g \in G \), then \( H = \{ g^n : n \in \mathbb{Z} \} \) is the smallest subgroup of \( G \) that contains \( g \).

9. Let \( A \) be a ring with identity \( I \), and let \( \phi : \mathbb{Z} \to A \) be given by \( \phi(n) = nI \). Show that:

(a) \( \phi \) is a ring homomorphism, and \( \phi(\mathbb{Z}) = \{ nI : n \in \mathbb{Z} \} \) is the smallest subring of \( A \) that contains \( I \);

(b) For each \( a \in A \), \( \{ n \in \mathbb{Z} : na = 0 \} \) is an ideal of \( \mathbb{Z} \) that contains the kernel \( N(\phi) \);

(c) \( \phi(\mathbb{N}) = \{ nI : n \in \mathbb{N} \} = \{ \sum_{k=1}^{n} I : n \in \mathbb{N} \} \) coincides with the set \( N(A) \).

10. Let \( A \neq \{ 0 \} \) be a ring with identity \( I \), and let \( \phi : \mathbb{Z} \to A \) be given by \( \phi(n) = nI \). Show that, if \( A \) is well-ordered (i.e., if \( A \) is ordered and every non-empty subset of \( A^+ \) has a minimum), then \( A \) is isomorphic to \( \mathbb{Z} \).

(HINT: Show the following:

(a) The set \( \{ a \in A : 0 < a < I \} \) is empty;

(b) \( A^+ = \phi(\mathbb{N}) \);

(c) \( A = \phi(\mathbb{Z}) \);

(d) \( \phi \) is injective.)

### 2.5 Factors, Multiples and Division

In an arbitrary ring \( A \) the equation \( ax = b \) may fail to have a solution, even if \( a \neq 0 \) (if \( a = 0 \) the equation can only have solutions if \( b = 0 \)). In order to avoid having to distinguish the equation \( ax = b \) from the equation \( xa = b \), we will assume in this section that \( A \) is a commutative ring.

**Definition 2.5.1.** If \( a, b \in A \), we say that \( a \) is **factor** (or **divisor**) of \( b \), or that \( b \) is **multiple** of \( a \), and we write “\( a \mid b \)”, if the equation \( ax = b \) has some solution \( x \in A \).

\(^8\)This is the most rigorous way to formulate the idea that the elements of \( N(A) \) are obtained by adding the identity \( I \) to itself an arbitrary (but finite) number of times.

\(^9\)Note that the term **divisor** is used here in with a slightly different meaning from the one in the term **zero divisor**. Recall that \( a \neq 0 \) is called a zero divisor if the equation \( ax = 0 \) has a solution \( x \neq 0 \).
Examples 2.5.2.

1. In a ring with identity 1, any element $b$ has at least the factors $1, -1, b, -b$, since $b = 1b = (-1)(-b)$ (but note that is possible that $1 = -1 = b = -b$).

2. If $K$ is a field and $0 \neq k \in K$, then every $r \in K$ is a multiple of $k$.

3. If $a \in \mathbb{Z}$ and $a^2 \neq 1$, the set of multiples of $a$ is smaller than $\mathbb{Z}$.

4. The multiples of $(x - 1)$ in the ring $\mathbb{R}[x]$ of polynomials with real coefficients in the variable $x$ are precisely $\{p(x) \in \mathbb{R}[x] : p(1) = 0\}$.

It is obvious that that the relation “is a factor of” is transitive (if $a|b$ and $b|c$, then $a|c$), and if $c \neq 0$ is not a zero divisor, we have that $ac|bc$ if and only if $a|b$. On the other hand, if $A$ is an ordered ring, it is also clear that $a|b$ if and only if $|a| \leq |b|$.

In this chapter we are mainly interested in the case $A = \mathbb{Z}$. The study of factorization and divisibility in more general rings will be done in the next chapter. In the case of the integers, the implication $n > 0 \Rightarrow n \geq 1$ gives us the following:

Lemma 2.5.3. If $m, n \in \mathbb{Z}$, then:

(i) $m|n \Rightarrow (|m| \leq |n| \text{ or } n = 0)$;

(ii) $(m|n \text{ and } n|m) \iff |m| = |n|$.

According to the previous lemma, if $n$ and $k$ are natural numbers and $k|n$, then $k < n$. As we have already observed, 1 is a factor of any natural number $n$. Hence, if $n$ and $m$ are natural numbers, the set of common factors (or divisors) of $n$ and $m$ is non-empty and bounded above, so it must have a maximum.

Similarly, the set of natural numbers common multiples of $n$ and $m$, i.e., the set $\{k \in \mathbb{N} : n|k \text{ and } m|k\}$, is non-empty since $nm > 0$ is a common multiple of $n$ and $m$. Hence, according to the the Well-ordering Principle, this set must have a minimum element.

Definition 2.5.4. If $n, m \in \mathbb{N}$, then:

(i) $\gcd(n, m) = \max\{k \in \mathbb{N} : n|k \text{ and } m|k\}$ is called the greatest common divisor of $n$ and $m$;

(ii) $\text{lcm}(n, m) = \min\{k \in \mathbb{N} : n|k \text{ and } m|k\}$ is called the least common multiple of $n$ and $m$. 
Example 2.5.5.

If \( n = 12 \) and \( m = 16 \), the natural numbers which are divisors of \( n \) and \( m \) form the sets

\[
\{ k \in \mathbb{N} : k | 12 \} = \{ 1, 2, 3, 4, 6, 12 \},
\]
\[
\{ k \in \mathbb{N} : k | 16 \} = \{ 1, 2, 4, 8, 16 \}.
\]

Therefore, the natural numbers which are common divisors of \( 12 \) and \( 16 \) form the set

\[
\{ k \in \mathbb{N} : k | 12 \text{ and } k | 16 \} = \{ 1, 2, 4 \}.
\]

We conclude that the corresponding greatest common divisor is \( \gcd(12, 16) = 4 \).

The natural numbers which are multiples of \( 12 \) and \( 16 \) are

\[
\{ k \in \mathbb{N} : 12 | k \} = \{ 12, 24, 36, 48, \ldots \},
\]
\[
\{ k \in \mathbb{N} : 16 | k \} = \{ 16, 32, 48, \ldots \},
\]

so the natural numbers which are common multiples of \( 12 \) and \( 16 \) form the set

\[
\{ k \in \mathbb{N} : 12 | k \text{ and } 16 | k \} = \{ 48, 96, \ldots \}
\]

We conclude that the corresponding least common multiple is \( \text{lcm}(12, 16) = 48 \).

Notice that in this example \( \gcd(m, n) \) is a multiple of all the common divisors of \( n \) and \( m \), while \( \text{lcm}(n, m) \) is a factor of all the common multiples of \( n \) and \( m \). In fact, we have:

Proposition 2.5.6. Let \( n, m, d, l \in \mathbb{N} \). Then:

(i) \( d = \gcd(n, m) \) if and only if:

(a) \( d | n \) and \( d | m \);
(b) for any \( k \in \mathbb{N} \), \( (k | n \text{ and } k | m) \Rightarrow k | d \).

(ii) \( d = \text{lcm}(n, m) \) if and only if:

(a) \( n | d \) and \( m | d \);
(b) for any \( k \in \mathbb{N} \), \( (n | k \text{ and } m | k) \Rightarrow d | k \).

We will prove these statements in the next section. For now, we introduce two more definitions for future reference. Note that in our conventions the natural number 1 is not prime.
Definition 2.5.7. If $p, m, n \in \mathbb{N}$.

(i) We say that $p$ is prime if $p > 1$ and if, for every $k \in \mathbb{N}$ such that $k | p$, we have that $k = 1$ or $k = p$.

(ii) We say that $n$ and $m$ are relatively prime if $\gcd(n, m) = 1$.

Example 2.5.8.

It is easy to check by listing all the possibilities that 4 and 9 are relatively prime, i.e., $\gcd(4, 9) = 1$, and that 13 is a prime number. In order to show that $p$ is prime it is not necessary to test all the numbers $k$ with $1 < k < p$: it is enough to stop testing at the greatest integer $k$ such that $k^2 < p$. For example, in the case $p = 13$, it is enough to check that 13 is not a multiple of 2 or 3.

We saw in the previous section (by using induction!) that any natural number $n$ is either even or odd, i.e., that given $n$, there exist integers $q$ and $r$ such that $n = 2q + r$, with $0 \leq r < 2$. This result is not specific to the natural number 2: for example, we can always write $n = 3q' + r'$, with $0 \leq r' < 3$, or $n = 4q'' + r''$, with $0 \leq r'' < 4$, etc. A moment of thought shows that these statements are a consequence of the Division Algorithm learned in Elementary School!

Theorem 2.5.9 (Division Algorithm). If $n, m \in \mathbb{Z}$ and $n \neq 0$, there exist unique integers $q, r$, such that $m = nq + r$, and $0 \leq r < |n|$.

Proof. We will prove only the case $n, m \in \mathbb{N}$, leaving as an exercise the generalization to any integers. Note that the argument to prove existence amounts to the usual method to perform the division of two numbers.

(i) Existence: Consider the set $Q = \{x \in \mathbb{N}_0 : nx \leq m\}$. Note that $Q$ is non-empty (because $0 \in Q$) and bounded above (because $x \leq nx \leq m$). Therefore $Q$ has a maximum $x = q$. It is clear that $nq \leq m < n(q+1)$, because $q \in Q$ and $(q+1) \notin Q$. Subtracting $nq$ to both sides of the inequality, we obtain that $0 \leq r \leq n$, since $r = m - nq$.

(ii) Uniqueness: Assume that $m = nq + r = nq' + r'$, with $0 \leq r, r' < n$. We have $-n < r - r' < n$, or equivalently $|r - r'| < |n|$, and also that $n(|q - q'|) = |r - r'|$. If $q \neq q'$, then $|q - q'| \geq 1$ and $|r - r'| > n$, so we must have $q = q'$. But from $n(|q - q'|) = |r - r'|$, we conclude also that $r = r'$. \qed

Definition 2.5.10. If $n, m \in \mathbb{Z}$, $n \neq 0$, and $m = nq + r$ with $0 \leq r < |n|$, we say that $q$ and $r$ are, respectively, the quotient and the remainder of the division of $m$ by $n$. 
CHAPTER 2. THE INTEGERS

The way the quotient and the remainder depend on the algebraic signs of \( m \) and \( n \) is not entirely obvious. This is illustrated in the following table by various examples:

\[
\begin{array}{c|c|c|c}
 m & n & q & r \\
\hline
 5 & 3 & 1 & 2 \\
 5 & -3 & -1 & 2 \\
 -5 & 3 & -2 & 1 \\
 -5 & -3 & 2 & 1 \\
\end{array}
\]

Assuming that \( n \neq 0 \) is fixed, consider the map \( \rho : \mathbb{Z} \to \{0, 1, \ldots, n-1\} \), where \( \rho(m) \) is the remainder of the division of \( m \) by \( n \). We leave as an exercise to show that the following result holds true:

**Proposition 2.5.11.** If \( x, y \) are arbitrary integers, we have

(i) \( \rho(x) = \rho(y) \) if and only if \( n|(x - y) \);

(ii) \( \rho(x \pm y) = \rho(\rho(x) \pm \rho(y)) \);

(iii) \( \rho(xy) = \rho(\rho(x)\rho(y)) \).

The Division Algorithm will be used many times in the sequel. For example, we will use in the next section to describe all the ideals of \( \mathbb{Z} \). Moreover, in the next Chapter we will show how it can be generalized to other rings, including some polynomial rings. Actually, all the notions that we have introduced in this section, such as prime number, gcd, lcm, etc., will eventually be generalized to a large class of rings.

**Exercises.**

1. Let \( A \) be a commutative ring, and \( a, b, c \in A \). Show that, if \( a|b \) and \( b|c \), then \( a|c \), and that, if \( c \neq 0 \) is not a zero divisor, we have \( ac|bc \) if and only if \( a|b \).

2. Prove Lemma 2.5.3.

3. Conclude the proof of Theorem 2.5.9.

4. Show that, if \( m, n, k \in \mathbb{N}, mn = k \) and \( m^2 > k \), then \( n^2 < k \).

5. List all the natural numbers between 100 and 200. Observe that \( 17^2 = 289 \), and remove from the list all the multiples of 2, 3, 5, 7, 11 and 13. Which numbers are left in the list?\(^{10}\)

---

\(^{10}\)This procedure is called **Eratosthenes filter**. Eratosthenes (276 B.C.-194 B.C.) was born where is now Libya, and was the third librarian of the famous Library of Alexandria. Among other achievements he established the sphericity of the Earth and found its diameter with great accuracy.
6. Determine the prime numbers between 1950 and 2050.
(HINT: Determine first the primes \( p \leq \sqrt{2050} \)).

7. If \( m, n \in \mathbb{Z} \), \( n \neq 0 \) and \( \rho : \mathbb{Z} \to \{0, 1, \ldots, n - 1\} \) is the remainder of the division by \( n \), when is it true that \( \rho(m) = \rho(-m) \)?

8. Prove Proposition 2.6.11. What is the relationship between this theorem and the “casting out nines” test of elementary arithmetics?

9. Is Theorem 2.5.9 still valid if one replaces in its statement the ring of integers by the ring formed by the multiples of 2. What if one replaces the ring of integers by the ring real numbers?

10. State and prove the analogue of Theorem 2.5.9 for the ring of Gaussian integers.

### 2.6 Ideals and Euclid's Algorithm

Assuming that one has two sand watches (hourglasses), one measuring intervals of 21 minutes and the other measuring intervals of 30 minutes, what intervals of time can one measure using the two watches? Certain intervals are obviously possible, by using \textit{successively} each watch, such as 30, 60, 90, \ldots, 21, 42, 63, \ldots, or additions of these numbers, such as 51, 81, 111, \ldots, 102, 123, \ldots, 132, 153, \ldots, etc. By using \textit{simultaneously} both watches we can also measure the differences of these numbers, such as 9 = 30 − 21, 3 = 63 − 60, etc.

Looking carefully at the resulting numbers, one reaches the following conclusions:

- It is possible to obtain any natural number of the form \( x21 + y30 \) with \( x, y \in \mathbb{Z} \).
- All the numbers of the form \( x21 + y30 \) are multiples of 3 (since 3 is the greatest common divisor of 21 and 30).

On the other hand, there exist integers \( x' \) and \( y' \) (\textit{e.g.}, \( x' = 3, y' = -2 \)) such that \( 3 = x'21 + y'30 \). In particular, if \( m = k3 \) is any multiple of 3, then \( m = k(x'21 + y'30) = k(x'21 + y'30) = x''21 + y''30 \). In other words:

- the numbers of the form \( x21 + y30 \) are precisely the multiples of 3;
- \( 3 = \gcd(21, 30) \) is the smallest natural number of the form \( x21 + y30 \).
Our aim now is to explore these remarks, generalizing them to an arbitrary ring. As a byproduct of this work we will be able to determine all the ideals of the ring \(\mathbb{Z}\) and we will find a method to compute \(\gcd(n,m)\), the so-called *Euclidean Algorithm*. This algorithm does not require the knowledge of prime factors of \(n\) and \(m\), so it is specially useful for computational purposes, and it is also easily generalized to polynomials.

If \(n\) and \(m\) are arbitrary fixed integers, we will consider the set

\[ I = \{ xn + ym : x, y \in \mathbb{Z} \}. \]

It is clear that \(I\) is non-empty and closed relative to the difference and the product by arbitrary integers, i.e., if \(x, y, x', y', z \in \mathbb{Z}\), then

\[
(xn + ym) \pm (x'n + y'm) = (x \pm x')n + (y \pm y')m \in I, \\
z(xn + ym) = (xn + ym)z = (zx)n + (zy)m \in I.
\]

In other words, \(I\) is an *ideal* of \(\mathbb{Z}\). On the other hand, if \(J\) is an ideal such that \(n, m \in J\), it is obvious that \(xn, ym \in J\) for all \(x, y \in \mathbb{Z}\), hence also \(xn + ym \in J\), so that \(I \subset J\). For this reason we say that \(I\) is the *smallest* ideal of \(\mathbb{Z}\) that contains \(n\) and \(m\), or still that \(I\) is the ideal *generated* by \(n\) and \(m\).

More generally, consider any ring \(A\) instead of the ring of integers, and replace the set \(\{n, m\}\) by an arbitrary (non-empty) subset \(S \subset A\). Note that the ring \(A\) itself is an ideal of \(A\) that contains \(S\). Hence, the family of ideals of \(A\) which contains \(S\) is non-empty. We will denote by \(\langle S \rangle\) the *intersection* of all the ideals in this family. It is clear that \(S \subset \langle S \rangle\). Also, one verifies directly from the definition of ideal that the intersection of a family of ideals of \(A\) is still an ideal of \(A\). Hence, \(\langle S \rangle\) is an ideal of \(A\) that contains \(S\). Finally, observe that if \(S \subset I \subset A\) and \(I\) is an ideal, then \(\langle S \rangle \subset I\). We leave as an exercise to verify that all these statements hold true.

**Definition 2.6.1.** If \(A\) is a ring and \(S \subset A\), one calls \(\langle S \rangle\) the *ideal generated* by \(S\), or the *smallest ideal* of \(A\) that contains \(S\). The elements of \(S\) are said to be *generators* of the ideal \(\langle S \rangle\), and \(S\) is called a *generating set* of \(\langle S \rangle\).

When \(S = \{a_1, a_2, \ldots, a_n\}\) is a finite subset of a ring \(A\), we will often write \(\langle a_1, a_2, \ldots, a_n \rangle\) instead of \(\langle S \rangle\). In the case of the integers, we saw above that \(\langle n, m \rangle = \{xn + ym : x, y \in \mathbb{Z} \}\), and it is immediate to show that \(\langle n \rangle = \{xn : x \in \mathbb{Z} \}\). However, there are rings where determining the ideal generated by a given set of elements is not so easy.
2.6. IDEALS AND EUCLIDES ALGORITHM

Examples 2.6.2.

1. Let $A$ be an abelian ring and $a \in A$. Then \{xa : x \in A\} is a subring of $A$, for if $x, y \in A$, we have

$$xa - ya = (x - y)a, \quad (xa)(ya) = (xay)a,$$

showing that \{xa : x \in A\} is closed for the difference and the product. Also, if $x, b \in A$, since $A$ is abelian, we have

$$b(xa) = (xa)b = (bx)a,$$

so \{xa : x \in A\} is an ideal. Finally \langle a \rangle contains necessarily the “multiples” of $a$, and we conclude that \langle a \rangle = \{xa : x \in A\}.

2. Let $A = M_2(\mathbb{Z})$ be the ring of $2 \times 2$ matrices with entries in $\mathbb{Z}$, and

$$a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The set of all “multiples” of $a$ is

$$\{ax : x \in M_2(\mathbb{Z})\} = \left\{ \begin{pmatrix} n & m \\ 0 & 0 \end{pmatrix} : n, m \in \mathbb{Z} \right\},$$

but \langle a \rangle = M_2(\mathbb{Z}) as it will be clear in the exercises.

The ideal of $\mathbb{Z}$ generated by 21 and 30 is also generated by 3. This is true for all ideals in the integers, which are always generated by only one of its elements (but it is by no means a property valid for arbitrary rings):

Theorem 2.6.3. \( I \) is an ideal of $\mathbb{Z}$ if and only if there exists $d \in \mathbb{Z}$ such that $I = \langle d \rangle$.

Proof. We verify both implications:

(a) If $I = \langle d \rangle$, then $I$ is obviously an ideal of $\mathbb{Z}$.

(b) Let $I$ be an ideal of $\mathbb{Z}$. If $I = \{0\}$ it obvious that $I = \langle 0 \rangle$, so we can assume that $I \neq \{0\}$. In this case, we observe that the ideal $I$ contains positive integers, since if $n \in I$ then $-n \in I$. Set

$$I^+ = \{n \in I : n > 0\},$$

and let $d$ be the minimum of $I^+$, which exists by the Well-ordering Principle.

Since $d \in I$ it follows that $\langle d \rangle \subseteq I$ (why?). Hence, it remains to show that $I \subseteq \langle d \rangle$, i.e., that if $m \in I$, then $m$ is a multiple of $d$. Let $m \in I$,
and let \( q \) and \( r \) be the quotient and the remainder of the division of \( m \) by \( d \) (recall that \( d > 0 \), so \( d \neq 0 \)). We have then that \( m = qd + r \), or \( r = m - qd \). Since \( qd \in \langle d \rangle \subseteq I \), this means that \( qd \in I \). But then \( r = m - qd \) is the difference of two elements in the ideal \( I \), so we must have \( r \in I \). Finally, since \( 0 \leq r < d \) and \( d \) is by definition the smallest positive element of \( I \), we must have \( r = 0 \), so \( m \) is a multiple of \( d \), i.e., \( m \in \langle d \rangle \). Hence, \( I \subseteq \langle d \rangle \).

In an arbitrary ring we have a special designation for those ideals which are generated by one of its elements:

**Definition 2.6.4.** An ideal \( I \) in a ring \( A \) is called a **principal ideal** if there exists \( a \in A \) such that \( I = \langle a \rangle \).

According to Theorem 2.6.3, all the ideals in the ring of integers are principal, but we will see later many examples of rings which have ideals that are not principal. However, the rings where all ideals are principal form an important class of rings.

The example of the ideal \( \langle 21, 30 \rangle \) suggests that, if \( I = \langle d \rangle = \langle n, m \rangle \), then \( |d| \) is the greatest common divisor of \( n \) and \( m \). The previous result allows us to prove this result and also part (i) of Proposition 2.5.6.

**Corollary 2.6.5.** If \( n, m \in \mathbb{N} \), then \( \langle n, m \rangle = \langle d \rangle \) where \( d = \gcd(n, m) \). In particular, we have that:

(i) the equation \( xn + ym = d \) has solutions \( x, y \in \mathbb{Z} \);

(ii) if \( k \) is a common divisor of \( n \) and \( m \), then \( k \) is also divisor of \( d \).

**Proof.** (i) The theorem shows that there exists a natural number \( d \) such that \( \langle n, m \rangle = \langle d \rangle \). Obviously, \( d \in \langle d \rangle = \langle n, m \rangle \). Since \( \langle n, m \rangle \) is the set of integers of the form \( xn + ym \), there are integers \( x', y' \) such that \( d = x'n + y'm \).

(ii) It is also obvious that \( n, m \in \langle n, m \rangle = \langle d \rangle \). Hence, \( n \) and \( m \) are multiples of \( d \), which is a common divisor of \( n \) and \( m \). On the other hand, if \( k \in \mathbb{N} \) is any common divisor of \( n \) and \( m \), then we have \( n = kn' \) and \( m = km' \), so that \( d = x'n + y'm = k(x'n' + y'm') \), or \( k | d \). In particular, \( k \leq d \) and \( d \) is the greatest common divisor of \( n \) and \( m \).

This corollary suggest that one can compute \( \gcd(n, m) \) by searching for the smallest natural number in the ideal \( \langle n, m \rangle \). The Division Algorithm makes this search possible by using the following lemma.

**Lemma 2.6.6.** If \( n, m \in \mathbb{N} \) and \( m = qn + r \), then \( \langle n, m \rangle = \langle n, r \rangle \).
Proof. On the one hand,

\[ k \in \langle n, m \rangle \implies k = xm + yn \]
\[ \implies k = x(qn + r) + yn \]
\[ \implies k = (xq + y)n + xr \]
\[ \implies k \in \langle n, r \rangle, \]

so \( \langle n, m \rangle \subset \langle n, r \rangle \). On the other hand,

\[ k \in \langle n, r \rangle \implies k = xn + yr \]
\[ \implies k = xn + y(m - qn) \]
\[ \implies k = (x - yq)n + ym \]
\[ \implies k \in \langle n, m \rangle, \]

so \( \langle n, r \rangle \subset \langle n, m \rangle \).

The Euclides Algorithm consists in applying repeatedly the previous lemma until one obtains an exact division (i.e., \( r = 0 \)). This is a very simple procedure, easy to program in a computer, which we illustrate now for the example that we started this section.

Example 2.6.7.

If \( n = 21 \) and \( m = 30 \), then

\[ 30 = 1 \cdot 21 + 9 \implies \langle 30, 21 \rangle = \langle 21, 9 \rangle, \]
\[ 21 = 2 \cdot 9 + 3 \implies \langle 21, 9 \rangle = \langle 9, 3 \rangle, \]
\[ 9 = 3 \cdot 3 + 0 \implies \langle 9, 3 \rangle = \langle 3 \rangle. \]

Hence

\[ \langle 30, 21 \rangle = \langle 3 \rangle, \]

and by the previous corollary we have that \( 3 = \gcd(21, 30) \).

In general, if we start with two natural numbers \( n \) and \( m \), and assuming that \( n < m \), the general procedure should be clear, and it corresponds to an iterative process which is very easy to program. Observe that it is also possible to determine integers \( x \) and \( y \) such that \( \gcd(n, m) = xn + ym \): from the equations above it follows immediately that

\[ 3 = 21 + (-2)9 \text{ and } 9 = 30 + (-1)21, \]

so:

\[ 3 = 21 + (-2) \cdot 30 + (-1)21 = (3)21 + (-2)30. \]
The next result takes advantage of the fact that \( \text{gcd}(n, m) \) is a linear combination of \( n \) and \( m \).

**Proposition 2.6.8.** Let \( m, n, p, k \in \mathbb{N} \) and assume that \( mn \mid kp \). If \( m \) and \( p \) are relatively prime, then \( m \) is factor of \( k \).

**Proof.** By assumption, \( \text{gcd}(m, p) = 1 \), so there exist integers \( x', y' \) such that \( 1 = x'm + y'p \). Hence, \( k = k(x'm + y'p) \). Moreover, since \( mn \mid kp \), there exists an integer \( z' \) such that \( kp = z'mn \). Hence,

\[
k = k(x'm + y'p) \\
= kx'm + y'kp \\
= kx'm + y'z'mn = (kx' + y'z'n)m,
\]

which shows that \( m \mid k \).

The least common multiple of two natural numbers maybe found also by applying Theorem 2.6.3. Given natural numbers \( n \) and \( m \), we observe that \( \langle n \rangle \cap \langle m \rangle \) is the set of common multiples of \( n \) and \( m \). Since the intersection of two ideals is an ideal, we conclude from Theorem 2.6.3 that \( \langle n \rangle \cap \langle m \rangle = \langle l \rangle \), where \( l \) is a natural number. Of course, \( l \) is a common multiple of \( n \) and \( m \), and any common multiple \( k \) of \( n \) and \( m \) is a multiple of \( l \), so that \( l \leq |k| \). Using this, and similarly to the case of the greatest common divisor, one can prove part (ii) of Proposition 2.5.6.

These remarks also suggest a way to define the greatest common divisor and the least common multiple of two integers:

**Definition 2.6.9.** If \( m, n, d, l \) are integers, and \( d, l \geq 0 \), then:

(i) \( d = \text{gcd}(n, m) \) if \( \langle d \rangle = \langle n, m \rangle \);

(ii) \( l = \text{lcm}(n, m) \) if \( \langle l \rangle = \langle n \rangle \cap \langle m \rangle \).

This definition is consistent with Theorem 2.6.3 and lead to the “natural” result that \( \text{gcd}(n, m) = \text{gcd}(|n|, |m|) \) and \( \text{lcm}(n, m) = \text{lcm}(|n|, |m|) \). Notice, by the way, that \( \text{gcd}(n, 0) = |n| \) and \( \text{lcm}(n, 0) = 0 \). Moreover, we leave to the exercises to show that for any integers \( m, n \in \mathbb{Z} \) one has:

\[
\text{lcm}(m, n) \text{gcd}(m, n) = |m||n|.
\]

Hence, the Euclides Algorithm also leads to the computation of \( \text{lcm}(m, n) \). We will further explore these results later to extend the notions of greatest common divisor and least common multiple to certain classes of rings.
If \( \langle m \rangle \) and \( \langle n \rangle \) are ideals of \( \mathbb{Z} \), it is clear that \( \langle n \rangle \subset \langle m \rangle \) if and only if \( m \mid n \).

In other words, to determine all the ideals that contain \( \langle n \rangle \) is equivalent to determine all the divisors of \( n \).

This is illustrated in the next figure, when \( n = 12 \), where each rectangle represents an ideal of \( \mathbb{Z} \) that contains \( \langle 12 \rangle \). Note that if a rectangle is contained in another, then the corresponding ideals also are, and that the ideal generated by 1 is obviously the whole ring \( \mathbb{Z} \).

Figure 2.6.1: The ideals of \( \mathbb{Z} \) that contain \( \langle 12 \rangle \).

Alternatively, we can represent the ideals that contain \( \langle 12 \rangle \) as in the following diagram (notice the subdiagram in the right, expressing the general property used here).

Figure 2.6.2: Os divisors of 12.

One should note that a diagram of this sort can be extended indefinitely below, but not above. In particular, given an ideal \( \langle m \rangle \subset \mathbb{Z} \) it is possible that the only ideal strictly containing it is \( \mathbb{Z} \) itself, and that happens precisely when \( n \) is a prime number. The same may happen with an ideal in an arbitrary ring, so we introduce:
**Definition 2.6.10.** An ideal $I \subset A$ is **maximal** if for every ideal $J \subset A$:

$$I \subset J \implies J = I \text{ or } J = A.$$  

Maximal ideals play an important role in many rings, not just in the ring of integers. For example, we will see later that the maximal ideals of an integral domain $D$ can be used to associate to $D$ certain fields. As we mentioned above, it is possible to identify the maximal ideals of $\mathbb{Z}$:

**Theorem 2.6.11.** An ideal $\langle p \rangle$ in $\mathbb{Z}$ is **maximal** if and only if $p = 1$ or $|p|$ is a prime number.

**Proof.** If $p, q \in \mathbb{N}$ and $\langle p \rangle \subset \langle q \rangle$, then $q|p$. If $p = 1$ or $p$ is prime, we have that $q = 1$ or $q = p$, so that $\langle q \rangle = \mathbb{Z}$ or $\langle q \rangle = \langle p \rangle$, and hence $\langle p \rangle$ is maximal.

On the other hand, if $p > 1$ is not prime, then there exists $q \in \mathbb{N}$ such that $1 < q < p$ and $q|p$. It follows that $\langle p \rangle \subsetneq \langle q \rangle \subseteq \mathbb{Z}$, so $\langle p \rangle$ is not a maximal ideal. $\square$

In the next section we will look more in detail into the properties of prime numbers.

**Examples 2.6.12.**

1. The ideal $\langle 0 \rangle$ is maximal in the ring $\mathbb{R}$, but not in the ring $\mathbb{Z}$.

2. The ideal $\langle x^2 - 2 \rangle$ is not maximal in $\mathbb{R}[x]$, because $\langle x^2 - 2 \rangle \subsetneq \langle x + \sqrt{2} \rangle$, and $\langle x + \sqrt{2} \rangle \neq \mathbb{R}[x]$. On the other hand, $\langle x^2 - 2 \rangle$ is maximal in $\mathbb{Z}[x]$ (why?).

3. In the Appendix, using Zorn’s Lemma, we show that in an arbitrary ring $A$ any proper ideal $I \subset A$ is contained in a maximal ideal.

If $A$ is a ring with identity $I$, then the map $\phi : \mathbb{Z} \rightarrow A$, $n \mapsto nI$, is a ring homomorphism. Hence, the set of solutions of the homogenous equation $nI = 0$ ($n \in \mathbb{Z}$), i.e., the kernel of $\phi$, is an ideal of $\mathbb{Z}$. Since every ideal of $\mathbb{Z}$ is principal, we see that there exists an unique integer $m \geq 0$ such that

$$\{n \in \mathbb{Z} : nI = 0\} = \langle m \rangle.$$  

Actually, according to the proof of Theorem 2.6.3 if $m > 0$, then $m$ is simply the smallest positive solution of the equation $nI = 0$. In any case, the integer $m$ deserves a special name.

**Definition 2.6.13.** If $A$ is a ring with identity $I$, its **characteristic** is the unique number $m \in \mathbb{N}_0$ such that

$$\{n \in \mathbb{Z} : nI = 0\} = \langle m \rangle.$$
2.6. IDEALS AND EUCLIDES ALGORITHM

Examples 2.6.14.

1. The rings \( \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C} \) and \( \mathbb{H} \) all have characteristic 0.

2. The ring \( \mathbb{Z}_2 \) has characteristic 2.

Exercises.

1. Give examples of natural numbers \( a, b, n \) such that \( ab \) is not a factor of \( n \), but \( a|n \) and \( b|n \).

2. Determine integers \( x \) and \( y \) such that \( \gcd(135, 1987) = x135 + y1987 \).

3. How can one measure 1 pint of water, if one only has two jars of sizes, respectively, 15 and 23 pints?

4. Let \( a, b, d, m, x, y, s, t \in \mathbb{Z} \). Show that:
   
   (a) \( d = \gcd(a, b), a = dx, b = dy \Rightarrow \gcd(x, y) = 1 \);
   
   (b) \( as + bt = 1 \Rightarrow \gcd(a, b) = \gcd(a, t) = \gcd(s, b) = \gcd(s, t) = 1 \);
   
   (c) \( \gcd(ma, mb) = |m| \gcd(a, b) \);
   
   (d) \( \gcd(a, m) = \gcd(b, m) = 1 \Leftrightarrow \gcd(ab, m) = 1 \);
   
   (e) \( a|m \) and \( b|m \Rightarrow ab|m \) \( \gcd(a, b) \);
   
   (f) \( |ab| = \gcd(a, b) \text{lcm}(a, b) \).

5. Let \( \mathcal{I} = \{I_\beta\}_{\beta \in B} \) be a family of ideals of a ring \( A \) indexed by some set \( B \). Consider the intersection of all sets in this family:

   \[ I = \bigcap_{\beta \in B} I_\beta = \{a \in A : a \in I_\beta, \text{ for all } \beta \in B\}. \]

   (a) Show that \( I \) is an ideal of \( A \).

   (b) Let \( S \subset A \) and let \( \mathcal{I} \) be the family of all ideals of \( A \) that contains \( S \). Show that \( I \) is the smallest ideal of \( A \) that contains \( S \).

6. Show that, if \( S_1 \subset S_2 \subset A \), then \( \langle S_1 \rangle \subset \langle S_2 \rangle \).

7. Assume that \( m, n \in \mathbb{Z} \), and show that:

   (a) \( \langle n \rangle \subset \langle m \rangle \Leftrightarrow m|n \);

   (b) \( \langle n \rangle = \langle m \rangle \Leftrightarrow m = \pm n \).
8. Show that, if \( A \) is an abelian ring with identity and \( a_1, a_2, \ldots, a_n \in A \), then
\[
\langle a_1, a_2, \ldots, a_n \rangle = \left\{ \sum_{k=1}^{n} x_k a_k : x_k \in A, 1 \leq k \leq n \right\}.
\]
How would you describe \( \langle a_1, a_2, \ldots, a_n \rangle \) if \( A \) was abelian without identity?

9. Let \( \{a_n\}_{n \in \mathbb{N}} \) be a sequence of integers.
   (a) Define \( d_n = \gcd(a_1, a_2, \ldots, a_n) \) and \( l_n = \text{lcm}(a_1, a_2, \ldots, a_n) \), and show that the equation \( d_n = \sum_{k=1}^{n} x_k a_k \) has solutions \( x_k \in \mathbb{Z} \).
   (b) Show that \( d_{n+1} = \gcd(d_n, a_{n+1}) \) and \( l_{n+1} = \text{lcm}(l_n, a_{n+1}) \).
   (c) For which values of \( n \) does the equation \( 30x + 105y + 42z = n \) have integer solutions?

10. Construct a diagram similar to the one for the divisors of 12 for \( n = 18 \).

11. Assume that \( A \) and \( B \) are unitary rings of characteristic, respectively, \( n \) and \( m \). Show that the characteristic of \( A \oplus B \) is \( \text{lcm}(n, m) \).

12. Let \( A \) be an abelian ring with identity 1, and let \( J \) be an ideal of \( A \). Show that \( J \) is maximal if and only if, for all \( a \not\in J \), the equation \( xa + y = 1 \) has solutions \( x \in A \) and \( y \in J \).
   (HINT: Check that \( \{xa + y : x \in A, y \in J\} \) is the ideal generated by \( J \cup \{a\} \).)

13. Let \( A \) be a ring with identity \( I \) and characteristic \( m > 0 \). Show that:
   \[
   \{nI : n \in \mathbb{Z}\} = \{I, 2I, \ldots, mI\}
   \]
is a ring with \( m \) elements. Show also that, if \( a \in A \), then the smallest positive solution of \( na = 0 \) is a factor of \( m \).
   (HINT: If \( \phi : \mathbb{Z} \to A, n \to nI, \) and \( n = mq + r \), then \( \phi(n) = \phi(r) \).)

14. Let \( A \) be a ring of characteristic 4. Determine the addition and multiplication tables of the subring \( \{nI : n \in \mathbb{Z}\} \subset A \). Decide whether or not this ring is isomorphic to the field with 4 elements mentioned in Chapter 1.

15. Determine the ideals of each of the following rings:
   (a) The ring \( 2\mathbb{Z} \) of even integers.
   (b) The ring \( \mathbb{Z}[i] \) of Gaussian integers.
   (c) The ring \( \mathbb{Z} \oplus \mathbb{Z} \).
   (d) The ring \( M_n(\mathbb{Z}) \).

16. For each of the rings in the previous exercise, determine whether or not its ideals are all principal.
2.7 The Fundamental Theorem of Arithmetic

One of the most important properties of the prime numbers is the fact that they generate, via multiplication, all the natural numbers \( n \geq 2 \). Our aim now is to make this fact precise and to prove it. Its correct formulation is known as the *Fundamental Theorem of Arithmetic*.

We start by showing that:

**Proposition 2.7.1.** Any natural \( n \geq 2 \) has at least one prime divisor \( p \).

*Proof.* The set
\[
D = \{ m \in \mathbb{N} : m > 1 \text{ and } m \mid n \}
\]
is non-empty, since it contains \( n \). Let \( p \) be a minimum of \( D \). If \( p \) is not a prime, then \( p = mk \), where \( 1 < m < p \). Since \( m \) is obviously a factor of \( n \), \( p \) cannot be a minimum of \( D \). We conclude that \( p \) must be a prime. \( \square \)

A consequence of this proposition is the *existence* of factorizations into prime numbers for any natural number \( n \geq 2 \).

**Corollary 2.7.2.** If \( n \geq 2 \) is a natural number there exist prime numbers \( p_1 \leq p_2 \leq \cdots \leq p_k \) such that
\[
n = \prod_{i=1}^{k} p_i.
\]

*Proof.* We give a proof by induction.

If \( n = 2 \), it is evident that \( n \) has a prime factorization with \( k = 1 \) and \( p_1 = 2 \). Assume now that any natural number \( m \) with \( 2 \leq m < n \) has a prime factorization. We wish to show that \( n \) also has a prime factorization.

Let \( P = \{ p \in \mathbb{N} : p \mid n, p \text{ prime} \} \) be the set of prime factors of \( n \). We now that \( P \) is bounded above (\( n \) is an upper bound) and non-empty (according to Proposition 2.7.1). Hence, \( P \) has a maximum \( q \). If \( q = n \), then \( n \) is a prime, so we set \( k = 1 \) and \( p_1 = q = n \). Otherwise, \( q < n \) so \( n = mq \), where \( 2 \leq m < n \). By the induction assumption, there exist prime numbers \( p_1 \leq p_2 \leq \cdots \leq p_{k'} \) such that \( m = \prod_{i=1}^{k'} p_i \), all of them bounded above by \( q \). In this case, \( k = k' + 1 \), and \( p_k = q \). \( \square \)

Now that we know the *existence* of prime factorizations for all natural numbers \( n > 2 \), we look into the question of *uniqueness* of these factorizations (up to the order of the factors). This is a problem of a rather different nature, as the following example illustrates.
Example 2.7.3.

Let us call an element in the ring $2\mathbb{Z}$ of even integers a “prime” if it cannot be expressed as a product of other even integers. For example, according to this definition the elements 2, 6 and 18 are all “primes”. It is left to the exercises to prove the analogues of the previous results for this ring which, just like the integers, is well-ordered. On the other hand, since $36 = 2 \cdot 18 = 6 \cdot 6$, it is clear that the “prime” factorizations in this ring are not unique.

The fundamental result about the integers that yields the uniqueness of prime factorizations is the following:

**Lemma 2.7.4 (Euclidean)**. Let $m,n,p \in \mathbb{Z}$ and assume that $p$ is a prime number. If $p$ is a factor of the product $mn$, then $p$ is a factor of $m$ or a factor of $n$.

**Proof.** Let $d = \gcd(m,p)$. Since $d$ is a factor of $p$ and $p$ is prime, we must have $d = 1$ or $d = p$.

Obviously, if $d = p$, then $p$ is factor of a $m$. On the other hand, If $d = 1$, there exist integers $x$ and $y$ such that $1 = xm + yp$. Hence, $n = nxm + nyp$ and since $p$ is a factor of $mn$, there exists also an integer $z$ such that $mn = zp$. We conclude that $n = xzp + nyp = (zx + ny)p$, which shows that $p$ is a factor of $n$.

The example above shows that Euclidean Lemma is not valid in the ring of even integers, for the definition of “prime” that we indicated.

Example 2.7.5.

When the Greek mathematicians found out about the existence of what we now call irrational numbers they were very intrigued and surprised, and even tried to hide its existence! We can now show without much difficulty that $\sqrt{2}$ is irrational, i.e., that there are no integers $n$ and $m$ such that $\left(\frac{n}{m}\right)^2 = 2$, or equivalently $n^2 = 2m^2$. For this we will argue by contradiction.

Assuming that $n$ and $m$ exist, we can choose them to be relatively prime (why?). Now observe that by Euclidean Lemma we must have:

\[ n^2 = 2m^2 \Rightarrow 2|m^2 \Rightarrow 2|n, \]

We conclude that $n = 2k$, for some integer $k$. Hence, $n^2 = 4k^2$, so $4k^2 = 2m^2$, or still $2k^2 = m^2$. Since $2|m^2$, it follows again from Euclidean Lemma that $2|m$. But $2$ being a divisor of both $m$ and $n$ contradicts our assumption that they are relatively prime. We conclude that the equation $n^2 = 2m^2$ has no integer solutions, and hence that $\sqrt{2}$ is not rational.
2.7. THE FUNDAMENTAL THEOREM OF ARITHMETIC

We leave it for the exercises to generalize this example to the case $\sqrt{n}$, where $n \in \mathbb{N}$ is not a perfect square. In other words, if $n$ is a natural number, its square root is either another natural number (case in which $n$ is a perfect square) or an irrational number.

It is not hard to generalize Euclid’s Lemma to any finite product of integers. The proof is by induction in the number of factors and is left to the exercises.

**Corollary 2.7.6.** If $p \in \mathbb{N}$ is prime and $p \mid \prod_{i=1}^{k} m_i$, then:

(i) $p \mid m_j$ for some $j$, with $1 \leq j \leq k$.

(ii) If the integers $m_i$ are primes, then $p = m_j$, for some $j$, with $1 \leq j \leq k$.

Finally, we can state and prove the

**Theorem 2.7.7 (Fundamental Theorem of Arithmetic).** Any natural number $n > 2$ has a prime factorization which is unique up to the order of the factors.

**Proof.** The existence of prime factorizations follows from Corollary 2.7.2. It remains to prove the uniqueness.

Assume that $k$ and $m$ are natural numbers, $p_1 \leq p_2 \leq \cdots \leq p_k$ and $q_1 \leq q_2 \cdots \leq q_m$ are primes, and

$$\prod_{i=1}^{k} p_i = \prod_{j=1}^{m} q_j.$$ 

We claim that $k = m$ and $p_i = q_i$. For that, we argue by induction in $k$:

- If $k = 1$, the result is obvious from the definition of a prime number.

- Assume that the result holds for the natural $k - 1$, and that

$$\prod_{i=1}^{k} p_i = \prod_{j=1}^{m} q_j.$$ 

Let $P = \{p_i : 1 \leq i \leq k \}$ and $Q = \{q_j : 1 \leq j \leq m \}$. Again by the definition of prime number we must have $m > 1$, since $k > 1$. By Corollary 2.7.6 it is clear that $p_k \in Q$, so that $p_k \leq q_m$. Similarly $q_m \in P$, so that $q_m \leq p_k$. We conclude that $p_k = q_m$, and it follows from the cancellation law that

$$\prod_{i=1}^{k-1} p_i = \prod_{j=1}^{m-1} q_j.$$
By the induction hypothesis, \( k - 1 = m - 1 \) and \( p_i = q_i \), for \( i < k \), so we conclude that \( k = m \) and \( p_i = q_i \), for \( i \leq k \).

A prime factorization of \( n \) can, of course, contain repeated factors. For that reason, one often writes it in the form:

\[ n = \prod_{i=1}^{m} p_i^{e_i} \quad (e_i \geq 1). \]

This is called the prime powers factorization of \( n \). Again, this expression is unique, up to the order of the factors.

The Fundamental Theorem of Arithmetic does not imply directly the existence of an infinite number of primes. This last fact was also discovered by Euclid.

**Theorem 2.7.8 (Euclid).** The set of prime numbers is unbounded.

**Proof.** We show that for any natural number \( n \) there exists a prime number \( p > n \). Given a natural number \( n \), consider the natural number \( m = n! + 1 \), where \( n! \) is the factorial of \( n \). It is obvious from the Division Algorithm that the remainder of the division of \( m \) by any natural number between 2 and \( n \) is 1. In particular, all the factors of \( m \), including its prime factors (which exist according to Proposition 2.7.1), are larger than \( n \). We conclude that there exist prime numbers greater than \( n \).

Hence, the prime numbers form an infinite sequence

\[ 2, 3, 5, 7, 11, 13, \ldots, p_n, \ldots \]

and this raises immediately some obvious questions. For example, is it possible to determine an explicit formula involving the natural number \( n \), which gives the prime \( p_n \)? How many prime numbers are there in the interval of 1 to \( n \)? How hard is it to find the prime factors of a given natural \( n \)? How large can it be the gap between two successive primes?

The answer to the first question above seems to be negative. In particular, no explicit formula is known that produces only prime numbers. One famous attempt is due to Fermat\(^1\) and involved the numbers of the form

\[ F_n = 2^{2^n} + 1, \]

\(^1\)Pierre of Fermat (1601-1665), French mathematician. Fermat, was a lawyer and one of the most fascinating characters of the History of Mathematics. He was one of the founders of Calculus and he discovered independently from Descartes (of whom he was a friend) the principles of Analytic Geometry. However, his most important contributions were, without a doubt, to Number Theory.
2.7. THE FUNDAMENTAL THEOREM OF ARITHMETIC

now known as Fermat numbers. It is easy to compute the Fermat numbers corresponding to \( n = 0, 1, 2 \) and 3: namely, 3, 5, 17, and 257, which are all primes. If \( n = 4 \) one finds the Fermat number 65537, which is still a prime number. There are however Fermat numbers that are not primes, has Euler find out in 1732 by taking \( n = 5 \). If this looks like a simple observation, note that \( n = 5 \) corresponds to the number:

\[
2^{2^5} + 1 = 4 294 967 297,
\]

and Euler found that its prime factorization is:

\[
2^{2^5} + 1 = 641 \cdot (6 700 417).
\]

The choice of the exponent \( e = 2^n \) is easy to explain. Assume that \( F = 2^e + 1, \) and \( e = ks, \) where \( k, s > 1, \) with \( s \) odd. The polynomial \( p(x) = x^s + 1 \) has the root \( x = -1, \) hence can be factor as

\[
p(x) = (x + 1)q(x),
\]

where \( q(x) \) is a polynomial with integer coefficients. Substituting \( x \) by \( 2^k, \) we conclude that

\[
F = 2^e + 1 = (2^k)^s + 1 = (2^k + 1)q(2^k),
\]

so the number \( F \) is not prime. In other words, if \( F = 2^e + 1 \) is prime then \( e \) has no odd factors greater than 1, so its only prime factor is 2, and we have \( e = 2^n, \) hence \( F \) is the Fermat number \( F_n. \)

In spite of the promising start of the sequence of Fermat numbers, we don’t know any Fermat number with \( n > 4 \) that is a prime, and it is known that some of these Fermat numbers are composed. For example, it is known that the smallest prime factor of the Fermat number corresponding to \( n = 1945 \) (a number with more than \( 10^{582} \) digits in its decimal expansion!) is a prime number \( p \) of 587 digits: \( p = 5 \cdot 2^{1947} + 1, \) and it is conjectured that no Fermat number with \( n > 4 \) is a prime. In spite of that, we will see in the exercises that these numbers can also be used to prove the existence of an infinite number of primes.

Questions concerning the number of primes in the interval \([1, n]\), or concerning the distribution of primes, are related to the probability of a randomly chosen natural number in the interval \([1, n]\) be a prime. It is also

\[\text{Leonhard Euler (1707-1783), Swiss mathematician. Euler was one of the most prodigious mathematicians of all times, having worked in various branches of Pure and Applied Mathematics, such as Analysis, Geometry, Number Theory, Fluid Mechanics, etc.}\]
related to the problem of finding an explicit formula for the \( n \)-esimal prime, as mentioned above. Legendre\textsuperscript{13} and Gauss were the first mathematicians to suggest an \textit{approximate} expression \( \pi(x) \) for the number of primes \(< x \).

At the end of the \textit{XIX} Century, Hadamard\textsuperscript{14} showed that

\[
\frac{\pi(x)}{x} \log x \to 1 \text{ when } x \to \infty.
\]

We will not discuss here any results of this nature that typically require analytic techniques.

Some of the problems are among the hardest problems in Mathematics, and illustrate the capabilities and limitations of the human mind. In spite of its abstract origin, these problems have interesting consequences in our day-to-day life. We will mention later some techniques from Cryptography that take advantage of the relative ease in manipulating “large” prime numbers when compared with the difficulties in finding prime factorizations. In practical applications, these “large” numbers can have more than 100 digits, while the determination of their prime factorizations by sequential verification of its possible prime factors may require a number of divisions of the order of \( 10^{50} \)! The security in the electronic transmission of the most secret communications or of simple financial transactions depends on our ignorance of a practical algorithm for prime factorization.

At this point it is hard to mention other “practical” applications where the properties of prime numbers are important. We mention only that questions concerning speech or image recognition require algorithms for the decomposition of signals into its fundamental frequencies, a technique known as Fourier Analysis. The theoretical maximum speed for these algorithms is directly related to the function \( \pi(x) \).

**Exercises.**

1. If \( n = \prod_{i=1}^{k} p_i^{e_i} \) and \( m = \prod_{i=1}^{k} p_i^{f_i} \) with \( e_i, f_i \geq 0 \) integers, find expressions for the \( \gcd(n, m) \) and the \( \text{lcm}(n, m) \).

2. Let \( p \) and \( q \) be distinct primes, and let \( n = p^2 q^3 \). Count all the natural numbers that are factors of \( n \), and show that their sum is \( (1 + p + p^2)(1 + q + q^2 + q^3) \).

\textsuperscript{13}Adrien Marie Legendre (1752-1833), French mathematician. Legendre became known for his fundamental contributions to Number Theory and to the study of elliptic function.

\textsuperscript{14}Jacques Hadamard (1865-1963), was one of the most influential French mathematicians of the second half of the \textit{XIX} Century and the first half of the \textit{XX} Century. His work spanned such different domains of Mathematics as, e.g., Number Theory or the Calculus of Variations.
3. Generalize the previous result to the case where \( n = \prod_{i=1}^{k} p_i^{e_i} \).

4. Prove a version of Corollary 2.7.2 for the ring \( 2\mathbb{Z} \) of all even integers.

5. Prove Corollary 2.7.6.

6. Show that, for any natural number \( n \), the interval \([n + 1, n! + 1]\) contains at least one prime.

7. The primes of the form \( 2^n - 1 \) are called Mersenne primes. Show that, if \( a^n - 1 \) is prime and \( n > 1 \), then \( a = 2 \) and \( n \) is prime.

8. Show that the sequence \( a_n = n^2 - n + 41 \) is not formed only by primes.

9. Show that, if \( p(x) \) is a non-constant polynomial with integer coefficients, then the set of integers \( n \) for which \( a_n = p(n) \) is not prime must be infinite.

10. Let \( F_n = 2^{2^n} + 1 \) be the \( n \)-esimal Fermat number \( (n \geq 0) \).
    
    (a) Show that \( F_{n+1} = 2 + \prod_{i=0}^{n} F_i \);
    
    (b) Show that if \( n \neq m \) then \( \gcd(F_n, F_m) = 1 \);
    
    (c) Explain why the previous result implies the existence of an infinite number of primes.

11. Show that, if \( n \) is not a perfect square, then \( \sqrt{n} \) is irrational (i.e., the equation \( x^2 = ny^2 \) has no solutions \( x, y \in \mathbb{Z} \)).

### 2.8 Congruences

We now turn to the study of other binary relations in \( \mathbb{Z} \), the so-called “congruence modulo \( m \)”, which are associated to the divisibility relation. We will use them to solve equations of the form \( ax + by = n \), where all the variables are integers. In the next chapter, we will also use them to exhibit an important class of finite rings.

**Definition 2.8.1.** Let \( m \in \mathbb{N}_0 \). If \( x, y \in \mathbb{Z} \), we say that \( x \) is congruent modulo \( m \) with \( y \) if and only if \( x - y \) is a multiple of \( m \). The integer \( m \) is called the modulus of congruence.

If \( x \) is congruent with \( y \) modulo \( m \), we will write \( x \equiv y \pmod{m} \). Therefore, we have:

\[
x \equiv y \pmod{m} \iff m|(x - y) \iff (x - y) \in \langle m \rangle.
\]
Recall that a binary relation is called an equivalence relation if it is reflexive, symmetric and transitive (see the Appendix).

**Proposition 2.8.2.** Congruence modulo $m$ is an equivalence relation.

**Proof.** We check that congruence modulo $m$ satisfies all three properties:

(i) $\equiv$ is reflexive: since 0 is a multiple of $m$, we have

$$x \equiv x \pmod{m}.$$

(ii) $\equiv$ is symmetric: obviously, $x - y = km$ if and only if $y - x = (-k)m$, hence

$$x \equiv y \pmod{m} \iff y \equiv x \pmod{m}.$$

(iii) $\equiv$ is transitive: if $x \equiv y \pmod{m}$ and $y \equiv z \pmod{m}$, then there exist integers $k$ and $n$ such that $x - y = km$ and $y - z = nm$. But then,

$$x - z = (x - y) + (y - z) = km + nm = (k + n)m,$$

hence $x \equiv z \pmod{m}$. $$

**Examples 2.8.3.**

1. Ignoring the variable $y$, the equation $3 = 21x + 30y$ ($x, y$ integers) is written

$$3 \equiv 21x \pmod{30}, \text{ or } 21x \equiv 3 \pmod{30}.$$

2. If $m = 0$, then since 0 is the only multiple of 0,

$$x \equiv y \pmod{0} \iff x = y.$$

Hence, the congruence relation modulo 0 is the usual equality relation. At the other extreme, if $m = 1$, then $x \equiv y \pmod{1}$, for all $x, y \in \mathbb{Z}$ (the integer $x - y$ is always a multiple of 1).

3. Any integer $n$ is even or odd, i.e., $n = 0 + 2k$ or $n = 1 + 2k$. Hence

$$n \equiv 0 \pmod{2}, \text{ or } n \equiv 1 \pmod{2}.$$

According to the Division Algorithm, if $m > 0$ and $x$ is any integer, then $x = mq + r$, where $q, r \in \mathbb{Z}$, and these integers are unique if $0 \leq r < m$. In other words, $x \equiv r \pmod{m}$, with $0 \leq r < m$, if and only if $r$ is the remainder of the division of $x$ by $m$. Therefore, we have the following generalization to any $m > 0$ of the example above for $m = 2$: 


2.8. CONGRUENCES

Proposition 2.8.4. If \( m > 0 \), any \( x \in \mathbb{Z} \) is congruent with exactly one integer \( r \) in the set \( \{0,1,2,\ldots,m-1\} \), where \( r \) is the remainder of the division of \( x \) by \( m \).

Example 2.8.5.

Any integer \( x \) satisfies exactly one of the following:

\[
\begin{align*}
x &\equiv 0 \pmod{4}, \quad x \equiv 1 \pmod{4}, \quad x \equiv 2 \pmod{4} \text{ or } x \equiv 3 \pmod{4}.
\end{align*}
\]

For a fixed \( m \neq 0 \), the set

\[
\mathcal{Z} = \{x \in \mathbb{Z} : x \equiv r \pmod{m}\}
\]

is called the congruence class or the residue class or simply the residue of the integer \( r \) modulo \( m \). We will see in the next chapter that the set of congruence classes \( \pmod{m} \) is the support of the ring \( \mathbb{Z}_m \), which generalizes to any \( m \) the examples of the rings \( \mathbb{Z}_2 \) and \( \mathbb{Z}_3 \) which we have already mentioned.

Equations that involve congruence relations can be dealt with just like the ordinary algebraic equations for numbers, with the exception of the “cancellation law” for the product. This is the content of the following proposition:

Proposition 2.8.6. If \( x \equiv x' \pmod{m} \) and \( y \equiv y' \pmod{m} \), then the following properties hold:

(i) \( x \pm y \equiv x' \pm y' \pmod{m} \);

(ii) \( xy \equiv x'y' \pmod{m} \).

In particular, we have that:

(iii) \( x \equiv x' \pmod{m} \Leftrightarrow x + a \equiv x' + a \pmod{m} \);

(iv) \( x \equiv x' \pmod{m} \Rightarrow ax \equiv ax' \pmod{m} \).

Proof. By assumption, both \( x - x' \) and \( y - y' \) are multiples of \( m \). Therefore, it is obvious that \( (x - x') \pm (y - y') = (x \pm y) - (x' \pm y') \) are multiples of \( m \), i.e., \( x \pm y \equiv x' \pm y' \pmod{m} \).

On the other hand, \( (x - x')y \) and \( x'(y' - y) \) are still multiples of \( m \). Hence, \( (x - x')y - x'(y' - y) = xy - x'y' \) is a multiple of \( m \), or equivalently, \( xy \equiv x'y' \pmod{m} \).

The proofs of the remaining statements are left for the exercises.
Example 2.8.7.

Since \(10 \equiv 3 \pmod{7}\) and \(11 \equiv -3 \pmod{7}\), according to Proposition 2.8.6:

\[
\begin{align*}
10 + 11 & \equiv 3 + (-3) \pmod{7} \iff 21 \equiv 0 \pmod{7}, \\
10 - 11 & \equiv 3 - (-3) \pmod{7} \iff -1 \equiv 6 \pmod{7}, \\
10 \cdot 11 & \equiv 3 \cdot (-3) \pmod{7} \iff 110 \equiv -9 \pmod{7}.
\end{align*}
\]

On the other hand, note that

\[
4 \cdot 5 \equiv 4 \cdot 8 \pmod{6},
\]

but \(5 \not\equiv 8 \pmod{6}\). Hence, in general, the cancellation law for the product fails: \(ax \equiv ax' \pmod{m}\) does not imply that \(x \equiv x' \pmod{m}\).

An important observation is that Proposition 2.8.6 uses precisely the properties of \(\langle m \rangle\) that make this set into an ideal. A corollary of this Proposition is the following, which is left as an exercise.

Corollary 2.8.8. If \(x \equiv y \pmod{m}\), then \(x^n \equiv y^n \pmod{m}\) for all natural numbers \(n\).

One of the most simple applications of Proposition 2.8.6 and of its Corollary is to obtain divisibility criteria in terms of the decimal expansion of \(x\).

Example 2.8.9.

Letting \(m = 3\), we observe that for any natural number \(k\):

\[
10 \equiv 1 \pmod{3} \implies 10^k \equiv 1 \pmod{3}.
\]

For example, if \(x = 1998\), we have that

\[
1998 = 1 \cdot 1000 + 9 \cdot 100 + 9 \cdot 10 + 8,
\]

\[
\equiv 1 + 9 + 9 + 8 \equiv 27 \equiv 0 \pmod{3}.
\]

We conclude that 1998 is divisible by 3, without using the Division Algorithm!

This example illustrates the following general divisibility criterium:

Proposition 2.8.10. A natural number \(n\) is divisible by 3 if and only if the addition of all the digits of its decimal representation is divisible by 3.

We discuss in the exercises similar divisibility criteria for 2, 5, 9 and 11.
Let us now turn to the study of linear equations of the form

$$ax \equiv b \pmod{m}.$$

Recall that the equation $ax = b$ has integer solutions if and only if $b$ is multiple of $a$, in which case the solution is unique (except if $a = 0$). We are interested in knowing for which values of $b$ (in terms of $a$ and $m$) does the equation $ax \equiv b \pmod{m}$ have solutions, and how can we find all its solutions.

The problem of existence of solutions of $ax \equiv b \pmod{m}$ is easily solved using the notion of greatest common divisor:

**Theorem 2.8.11.** The equation $ax \equiv b \pmod{m}$ has solutions if and only if $b$ is a multiple $d = \gcd(a, m)$.

**Proof.** Indeed, the equation $ax \equiv b \pmod{m}$ has solutions if and only if there exist integers $x$ and $y$ such that $b - ax = my$, i.e., $b = ax + my$. In other words, $ax \equiv b \pmod{m}$ has solutions precisely when $b$ is a linear combination of $a$ and $m$ with integers coefficients. As we saw in Section 2.6 (see Definition 2.6.9), the linear combinations of $a$ and $m$ with integers coefficients are exactly the multiples of $\gcd(a, m)$.

The previous result shows that Euclids Algorithm can be used to decide about the existence of solutions of the equation $ax \equiv b \pmod{m}$. Actually, since this algorithm allows to obtain $d = \gcd(a, m)$ as a linear combination of $a$ and $m$, it gives also one solution of $ax \equiv d \pmod{m}$. This leads easily to the solutions of the original equation, as we illustrate in the next example.

**Example 2.8.12.**

Consider the equation

$$15x \equiv b \pmod{40}.$$  

Using Euclides Algorithm, we have

$$40 = 15 \cdot 2 + 10, \quad \text{or} \quad 10 = 40 + 15 \cdot (-2),$$

$$15 = 10 \cdot 1 + 5, \quad \text{or} \quad 5 = 15 + 10 \cdot (-1) = 15 \cdot 3 + 40 \cdot (-1),$$

$$10 = 5 \cdot 2 + 0, \quad \text{hence} \quad 5 = \gcd(15, 40) = 15 \cdot 3 + 40 \cdot (-1).$$

We conclude from Theorem 2.8.11 that the equation has a solution if and only if $b$ is a multiple of 5. These very same computations show that we have

$$5 = 15 \cdot 3 + 40 \cdot (-1).$$

Hence, $x = 3$ is a solution of $15x \equiv 5 \pmod{40}$. More generally, by Proposition 2.8.6, $x = 3k$ is solution of $15x \equiv 5k \pmod{40}$. 

For example, if $k = 2$ we see that $x = 6$ is a solution of the equation $15x \equiv 10 \pmod{40}$, because:

$$15 \cdot 3 \equiv 5 \pmod{40} \implies 15 \cdot 2 \equiv 5 \cdot 2 \pmod{40},$$

Obviously, any integer $x$ that verifies $x \equiv 6 \pmod{40}$ is also a solution of $15x \equiv 10 \pmod{40}$, so this last equation has an infinite number of solutions.

However, the integers $x \equiv 6 \pmod{40}$ do not include all the solutions of $15x \equiv 10 \pmod{40}$. For example, $x = -2$ is also solution of $15x \equiv 10 \pmod{40}$, but $-2 \not\equiv 6 \pmod{40}$. This is again a consequence of the lack of a “cancellation law” for the product, because

$$3 \cdot 5 \cdot (-2) \equiv 5 \cdot 2 \pmod{40} \not\equiv 6 \equiv -2 \pmod{40}.$$

We can look under what circumstances this multiplicity of solutions is not possible. Notice that Theorem 2.8.11 also shows that the equation $ax \equiv b \pmod{m}$ has solutions for all $b$ precisely when $\gcd(a, m) = 1$, i.e., when $a$ and $m$ are relatively prime. Obviously, this happens if and only if the equation $ax \equiv 1 \pmod{m}$ has solutions.

**Definition 2.8.13.** One says that $a \in \mathbb{Z}$ is invertible $\pmod{m}$ if $ax \equiv 1 \pmod{m}$ has a solution, i.e., if $a$ and $m$ are relatively prime. If $c$ is a solution of the equation $ax \equiv 1 \pmod{m}$, we call $c$ is an inverse $\pmod{m}$ of $a$.

**Examples 2.8.14.**

1. The equation $4x \equiv b \pmod{9}$ has solutions for every $b$, since $\gcd(4, 9) = 1$. In particular, $4x \equiv 1 \pmod{9}$ has the solution $x = -2$ because $4(-2) + 9 = 1$.

   Hence,
   
   (a) $-2$ is an inverse of $4 \pmod{9}$, and
   
   (b) $x = -2b$ is a solution of $4x \equiv b \pmod{9}$, for every integer $b$.

2. Since $\gcd(21, 30) = 3$, $21$ is not invertible $\pmod{30}$.

Assume that $x = c$ is a particular solution of $ax \equiv b \pmod{m}$. As we have observed, any integer $x'$ congruent with $c \pmod{m}$ is also solution of the equation:

$$x' \equiv c \pmod{m} \implies ax' \equiv ac \equiv b \pmod{m}.$$

It is possible that there are other solutions $x''$, that are not congruent with $c$, i.e., we may have

$$ax'' \equiv b \pmod{m} \text{ with } x'' \not\equiv c \pmod{m}.$$
2.8. CONGRUENCES

However, when \( a \) and \( m \) are relatively prime this is impossible, due to the following “cancellation law”:

**Theorem 2.8.15.** If \( a \) and \( m \) are relatively prime,

\[
ax \equiv ay \pmod{m} \iff x \equiv y \pmod{m}.
\]

**Proof.** We know already that

\[
x \equiv y \pmod{m} \implies ax \equiv ay.
\]

On the other hand, let \( c \) be an inverse \((\text{mod } m)\) of \( a \), so that \( ac \equiv ca \equiv 1 \pmod{m} \). Then

\[
ax \equiv ay \pmod{m} \implies c(ax) \equiv c(ay) \pmod{m} \quad \text{(Proposition 2.8.6)},
\]

\[
\implies (ca)x \equiv (ca)y \pmod{m} \quad \text{(associativity)},
\]

\[
\implies x \equiv y \pmod{m} \quad \text{ca} \equiv 1 \pmod{m}.
\]

We conclude from Theorems 2.8.11 e 2.8.15 that

**Theorem 2.8.16.** If \( a \) and \( m \) are relatively prime and \( b \in \mathbb{Z} \), then:

(i) The equation \( ax \equiv b \pmod{m} \) has at least one solution \( c \in \mathbb{Z} \);

(ii) \( ax \equiv b \pmod{m} \implies x \equiv c \pmod{m} \).

Therefore, when \( a \) and \( m \) are relatively prime the solution of the linear equation \( ax \equiv b \pmod{m} \) is very simple. We illustrate it in the following example.

**Example 2.8.17.**

Since \( \gcd(4,7) = 1 \) and \( 4 \cdot 2 \equiv 1 \pmod{7} \), we conclude that the solutions of \( 4x \equiv 1 \pmod{7} \) (the inverses of \( 4 \pmod{7} \)) are precisely the integers which satisfy \( x \equiv 2 \pmod{7} \), i.e., the integers of the form \( x = 2 + 7k \). For example, if \( b = 3 \), we have

\[
4x \equiv 3 \pmod{7} \iff x \equiv 6 \pmod{7}.
\]

When \( a \) and \( m \) are not relatively prime, it is still easy to determine all the solutions of the equation

\[
ax \equiv b \pmod{m}.
\]

For that, recall that if \( a = a'd \) and \( m = m'd \) with \( d = \gcd(a,m) \neq 0 \), then \( a' \) and \( m' \) are relatively prime.
Example 2.8.18.

Let us find all solutions of the equation

\[ 15x \equiv 10 \pmod{40}. \]

Obviously, this equation is equivalent to \( 15x = 10 + 40y \) and this last equation can be divided by \( \gcd(15, 40) = 5 \), giving \( 3x = 2 + 8y \), or \( 3x \equiv 2 \pmod{8} \). Hence, we have:

\[ 15x \equiv 10 \pmod{40} \iff 3x \equiv 2 \pmod{8}, \]

where in the last equation obviously \( \gcd(3, 8) = 1 \). We have already seen that \( x = 6 \) is a solution of the original equation, and therefore also of \( 3x \equiv 2 \pmod{8} \). It follows from Theorem 2.8.16 that the solutions of the last equation are the integers which satisfy \( x \equiv 6 \pmod{8} \).

We conclude that the solutions of \( 15x \equiv 10 \pmod{40} \) are the integers of the form \( x = 6 + 8k \). In particular, this equation has 5 solutions that are not congruents (mod 40), namely, \( x = 6, 14, 22, 30, 38 \) (the solution that we found before, \( x = -2 \), is congruent to \( x = 38 \pmod{40} \)).

We discuss in the exercises how to determine the number of solutions of \( ax \equiv b \pmod{m} \) that are not congruent (mod \( m \)).

One can also solve systems of the form

(2.8.1) \[
\begin{align*}
  x &\equiv a \pmod{m}, \\
  x &\equiv b \pmod{n}.
\end{align*}
\]

The following result was found by the Chinese mathematician Sun-Tsu in the 1st Century AD! For that reason, it is known as the Chinese Remainder Theorem.

Theorem 2.8.19 (Chinese Remainder Theorem). The system (2.8.1) has solutions for all \( a \) and \( b \) if and only if \( m \) and \( n \) are relatively prime. In this case, if \( c \) is a solution, then (2.8.1) is equivalent to \( x \equiv c \pmod{mn} \).

Proof. It is obvious that the integers of the form \( x = a + ym \) are the solutions of the equation \( x \equiv a \pmod{m} \). Hence \( x = a + ym \) is a solution of (2.8.1) if and only if

\[ x \equiv b \pmod{n} \Rightarrow a + ym \equiv b \pmod{n} \Rightarrow my \equiv b - a \pmod{n}. \]

We saw above that this last equation has a solution for all \( a \) and \( b \) if and only if \( m \) and \( n \) are relatively prime. In this case, Theorem 2.8.16 shows that
the solutions are the integers of the form \( y = y' + zn \), where \( y' \) is any fixed solution, and \( z \) is arbitrary. We conclude that the solutions of the system \((2.8.1)\) are the integers of the form

\[
x = a + (y' + zn)m = c + z(nm),
\]

where \( c = a + y'm \), i.e., are the solutions of \( x \equiv c \pmod{mn} \).

The case where \( m \) and \( n \) are not relatively prime is dealt with in an exercise at the end of the section.

**Example 2.8.20.**

Consider the system

\[
\begin{aligned}
x &\equiv 1 \pmod{4}, \\
x &\equiv 2 \pmod{9}.
\end{aligned}
\]

The solutions of the first equation are the integers of the form \( x = 1 + 4y \).

They should also satisfy the second equation:

\[
1 + 4y \equiv 2 \pmod{9} \iff 4y \equiv 1 \pmod{9}
\]

\(\text{we saw that } -2 \text{ is inverse of } 4 \pmod{9} \). We conclude that \( c = 1 + 4(-2) = -7 \)

is a particular solution of the system. By the Chinese Remainder Theorem, the system is equivalent to the equation \( x \equiv -7 \pmod{36} \), so the solutions are the integers of the form \( x = -7 + 36z \), with \( z \in \mathbb{Z} \).

The equations studied in this section have as unknowns integers. Such equations are called Diophantine.\(^\text{15}\) Although, as we saw above, solving linear Diophantine equations is relatively simple, some of the hardest (and more famous) problems in Mathematics concern non-linear Diophantine equations. Perhaps the most famous one is “Fermat’s Last Theorem” concerning the integers solutions \( x, y \) and \( z \) of the equation

\[
x^n + y^n = z^n.
\]

Although this equation has an infinite number of solutions when \( n=2 \) (e.g., \( x = 3, y = 4 \) and \( z = 5 \)), solutions for \( n > 2 \) were never found. Fermat wrote in the margins of a copy of the *Arithmetica* of Diophantus, next to

\(^\text{15}\)After Diophantus of Alexandria, a Greek mathematician of the III Century, author of a famous series of books called *Arithmetica*, most of which are now lost. This treaty included, among many things, solutions (some of there extremely ingenious!) of algebraic equations. Diophantus was only interested in rational solutions, and called the irrational ones “impossible”.

the discussion of the Pythagoras’s theorem, that he knew how to prove the non-existence of solutions for \( n > 2 \), but that the margin was too small to describe the argument. Fermat’s proof remains unknown to these days and, in fact, it took more than 300 years, and the efforts of many generations of mathematicians, to reach a complete solution of Fermat’s Last Theorem. The proof, due to the american mathematician Andrew Wiles\(^\text{16}\), is no doubt one of the greatest discovers of Mathematics. The degree of sophistication of the proof, which culminates the works of many great mathematicians for more than 200 years, makes it one of the most elaborate intellectual constructions of all the humanity.

**Exercises.**

1. Find all integers solutions of \( 21x + 30y = 9 \).

2. For which integers \( b \) does the equation \( 533x \equiv b \pmod{4141} \) have solutions?

3. State and prove criteria of divisibility by 2, 5, 9 and 11, in terms of the decimal representation of a natural number \( n \).

4. Given some natural number \( k \), find the remainder of the division of \( 3^k \) be 7.

5. Determine all solutions of \( xy \equiv 0 \pmod{12} \).

6. For \( a, m \in \mathbb{Z} \), with \( m \neq 0 \), show that exactly one of following holds true:
   
   \( (a) \ ax \equiv 1 \pmod{m} \) has solutions (so gcd(\( a, m \)) = 1), or
   \( (b) \ ax \equiv 0 \pmod{m} \) has solutions \( x \neq 0 \pmod{m} \) (so gcd(\( a, m \)) \( \neq 1 \)).

7. Assume that gcd(\( a, m \)) = \( d \) and \( m = dn \). Show that:
   
   \( (a) \ ax \equiv 0 \pmod{m} \) has \( d \) solutions \( x \), with 0 < \( x \leq m \), namely \( x = n, 2n, \ldots, dn \);
   \( (b) \ ax \equiv b \pmod{m} \) has either 0 or \( d \) solutions \( x \), with 0 < \( x \leq m \).

8. Show that the equation \( x^2 + 1 \equiv 0 \pmod{11} \) has no solutions.

---

\(^{16}\)Andrew Wiles announced in the summer of 1993 a proof of Fermat’s Last Theorem, but it was soon recognized that a crucial step was missing. Finally, in September of 1994, Wiles together with Richard Taylor found an argument which allowed to circumvent the missing step. The correct proof was published in the paper “Modular elliptic curves and Fermat’s last theorem”, *Ann. of Math.* 141 (1995), no. 3, 443–551, and refers to the following paper in the *Annals*, which is precisely a joint paper of Wiles and Taylor, “Ring-theoretic properties of certain Hecke algebras”, *Ann. of Math.* 141 (1995), no. 3, 553–572.
9. Determine which numbers between 1 and 8 have an inverse (mod 9).

10. Find all solutions of the equation \( x^2 + 1 \equiv 0 \pmod{13} \).

11. Show that \( x^5 - x \equiv 0 \pmod{30} \) for every integer \( x \).

12. Show that, if \( a \equiv a' \pmod{m} \), then \( \gcd(a, m) = \gcd(a', m) \).

13. Find the values of \( c \in \mathbb{Z} \) for which the following system has solutions:
   \[
   \begin{cases}
   x \equiv c \pmod{14} \\
   2x \equiv 10 \pmod{42}
   \end{cases}
   \]

14. Solve the system of equations
   \[
   \begin{cases}
   2x + 3y \equiv 3 \pmod{5} \\
   3x + y \equiv 4 \pmod{5}
   \end{cases}
   \]

15. Which day of the week was March 15, 1800?
   (Hint: In the current Western calendar, called the Gregorian calendar, a leap year is one which is divisible by 4, but not 100, or divisible by 400. For example, the years 1700, 1800, and 1900 are not leap years, but 1600 and 2000 are).

16. Five shipwreck survivors arrive at an island where they find a chimpanzee. After they spend one day picking up coconuts, they decide to leave the division of the coconuts for the next day. During the night each survivor wakes up and goes pickup what he thought was his share of the coconuts. Each of them finds that he cannot divide the coconuts by 5, since there is always one leftover, which they leave for the chimpanzee. The next day, they decide to split the coconuts that were left from their night incursions and now they find that they can split them exactly among them. Knowing that that there were at most 10,000 coconuts in the island, how many coconuts did they pickup?

### 2.9 Prime Factorizations and Cryptography

We will now describe, as an example of the theory we have developed so far, an application to Cryptography. We start with some preliminary results.

We will denote by \( C^n_k \) the "number of arrangements of \( n \) elements in groups of \( k \)", so that:

\[
k!(n-k)!C^n_k = n!.
\]

Since \( n|n! \), this equation and Euclid’s Lemma, imply that whenever \( n = p \) is a prime and \( 0 < k < p \), we have \( p|C^n_k \). We state this as a lemma for future reference:
Lemma 2.9.1. If $p$ is a prime and $0 < k < p$, then $C_p^k \equiv 0 \pmod{p}$.

Recall that the well-known Newton's binomial formula states that:

\[(2.9.1) \quad (a+b)^n = \sum_{k=0}^{n} C^n_k a^k b^{n-k}.\]

(this formula is actually valid in any commutative ring, as can easily be seen by applying induction). When $n = p$ is prime, it leads to:

Proposition 2.9.2 (“The Freshman’s Formula”). If $p$ is prime, $(a+b)^p \equiv a^p + b^p \pmod{p}$.

Proof. By Lemma 2.9.1 the only terms in the binomial expansion that may fail to be divisible by $p$ correspond to $k = 0$ and $k = n$.

Theorem 2.9.3 (Fermat). If $p$ is prime, $a^p \equiv a \pmod{p}$.

Proof. Let us prove this result by induction in $a$. The result is obvious if $a = 0$. Assuming it is true for some integer $a \geq 0$, we have that $(a+1)^p \equiv a^p + 1 \pmod{p}$, by the freshman’s formula, and that $a^p \equiv a \pmod{p}$, by the induction hypothesis. We conclude that $(a+1)^p \equiv a + 1 \pmod{p}$, so the result is true for any integer $a \geq 0$. Since $(-1)^p \equiv (-1) \pmod{p}$ for any prime $p$ (why?), the result holds for all integers.

An interesting corollary of this result is a practical method to find inverses $(\pmod{p})$.

Corollary 2.9.4. If $p$ is prime and $a \not\equiv 0 \pmod{p}$, then $a^{p-1} \equiv 1 \pmod{p}$. In particular, $a^{p-2}$ is the inverse $(\pmod{p})$ of $a$.

Proof. Obviously, since $p$ is a prime, gcd$(a,p)$ must be 1 or $p$, so that either $a \equiv 0 \pmod{p}$ or gcd$(a,p) = 1$. Since the first case is excluded, we must have gcd$(a,p) = 1$ and $a$ is invertible $(\pmod{p})$. Using the Fermat’s theorem, we conclude that

\[ a^p = a^{p-1}a \equiv a \pmod{p} \implies a^{p-1} \equiv 1 \pmod{p}. \]

We can now describe a public key encryption mechanism, which is particularly ingenious. The reason for this name is that the encryption method
can be known to everyone, but the decryption method is kept secret. Vari-
ations of this method of encryption are used, e.g., to make secure financial
transactions in the Internet.\footnote{One of the most commonly used methods is the so-called RSA encryption method (discovered by Rivest, Shamir and Adleman). Such a public system, besides the encryption mechanism that we describe, includes also a simple signature verification scheme, which is crucial for private communications via public channels.}

The main ingredients are a natural number $N = pq$, which is the product
of two distinct primes $p$ and $q$, and another natural number $r$, relatively
prime to both $p - 1$ and $q - 1$. The numbers $N$ and $r$ are public, but
the prime factors of $N$ are kept secret. Hence, this system explores the
possibility of generating very large prime numbers, together with our lack of
knowledge of a practical, efficient, algorithm to determine the prime factors
of large natural numbers.

Now the encryption proceeds as follows: assume that we would like to
transmit numbers $a$ such that $0 \leq a < N$. Instead of transmitting $a$, one
sends the remainder of the division of $a^r$ by $N$, which we denote by
$b = \rho(a^r, N)$. The decryption amounts to finding $a$ given $b = \rho(a^r, N)$. This
computation can only be done efficiently if one knows the prime factors
of $N$, in which case it is a simple application of our previous results, as we now
explain.

On the one hand, setting $c = \rho(b, p)$ and $d = \rho(b, q)$, it is clear that
\[
c \equiv a^r \pmod{p} \quad \text{and} \quad d \equiv a^r \pmod{q}.
\]

On the other hand, since $r$ is relatively prime to both $p - 1$ and $q - 1$, there
exist integers $x, y, x', y'$ such that
\[
1 = rx + (p - 1)y = rx' + (q - 1)y'.
\]

It follows from the corollary above that
\[
\begin{align*}
a &= a^{rx + (p - 1)y} \equiv (a^r)^x \equiv c^x \pmod{p}, \\
a &= a^{rx' + (q - 1)y'} \equiv (a^r)^{x'} \equiv d^{x'} \pmod{q}.
\end{align*}
\]

Note that we can find $x$ and $x'$ directly by applying Euclides Algorithm.
Also, interpreting negative powers of $a$ as positive powers of an inverse of $a$, it is clear that one can always choose $x$ and $x'$ to be both positive.

We conclude that the original message $a$ satisfies the system
\[
\begin{align*}
a &\equiv (\rho(b, p))^x \pmod{p} \\
a &\equiv (\rho(b, q))^{x'} \pmod{q}.
\end{align*}
\]
According to the Chinese Remainder Theorem, this system determines uniquely the number $a \pmod{pq}$, i.e., $\pmod{N}$. Since we assumed $1 \leq a < N$, we have found $a$. Solving this system requires only the knowledge of an inverse of $p \pmod{q}$, and this can again be achieved by applying Euclid's Algorithm. We illustrate the method in the following example.

**Example 2.9.5.**

Let $N = 21 = 3 \cdot 7 = p \cdot q$. The number $r = 5$ is relatively prime to $p-1 = 2$ and $q-1 = 6$. Let us assume that we wish to transmit the message $a = 4$. According to the method just described, instead of transmitting $a$, one sends instead the remainder of the division of $a^r = 1024$ by $N = 21$, that is $b = \rho(1024, 21) = 16$ (since $1024 = 48 \cdot 21 + 16$).

Now assume that the message $b = 16$ reaches the receiver, who wishes to decrypt it and find $a$. Since

$$1 = 5 \cdot (-1) + 2 \cdot 3 = 5 \cdot (-1) + 6' \cdot 1,$$

we conclude that $x = -1$ and $x' = -1$. On the other hand, the numbers $c$ and $d$ are given by

$$c = \rho(b, p) = 1, \ (since \ 16 = 3 \cdot 5 + 1)$$

$$d = \rho(b, q) = 2, \ (since \ 16 = 7 \cdot 2 + 2).$$

Hence, the number $a$ satisfies the system

$$\begin{cases}
  a \equiv (1)^{-1} \pmod{3} \\
  a \equiv (2)^{-1} \pmod{7}
\end{cases}.$$

Noting that $2 \cdot 4 = 8 \equiv 1 \pmod{7}$, the inverse of $2 \pmod{7}$ is $4$, and we conclude that $a$ satisfies the system

$$\begin{cases}
  a \equiv 1 \pmod{3} \\
  a \equiv 4 \pmod{7}
\end{cases}.$$

The solutions of the first equation are $a = 1 + 3n$, with $n \in \mathbb{Z}$. Replacing in the second equation, one finds

$$1 + 3n \equiv 4 \pmod{7} \implies 3n \equiv 3 \pmod{7}.$$

In order to solve this last equation we note that the inverse of $3 \pmod{7}$ is $5$ (since $3 \cdot 5 = 15 \equiv 1 \pmod{7}$), so that:

$$n \equiv 3 \cdot 5 = 15 \pmod{7}.$$

The Chinese Remainder Theorem now gives that $a \equiv 1 + 3 \cdot 15 = 46 \pmod{21}$. Since $0 \leq a < 21$, the receiver concludes that the message transmitted was $a = 4$.  

Exercises.

1. A word is codified so that each letter of the (english) alphabet corresponds to a natural number: $a \mapsto 1$, $b \mapsto 2$, $c \mapsto 3$, \ldots. Then the message “10,21,23,10,16,1” is transmitted in a system of public key with $N = 35$ and $r = 5$. What was the original word?

2. Prove the following generalization of the Chinese Remainder Theorem: If $d = \gcd(m,n)$ the system

$$
\begin{cases}
    x \equiv a \pmod{m} \\
    x \equiv b \pmod{n}
\end{cases}
$$

has solutions if and only if $d|(a - b)$. In this case, if $c$ is a solution, then the system is equivalent to $x \equiv c \pmod{\text{lcm}(n,m)}$. 
Chapter 3

Other Examples of Rings

3.1 The Rings $\mathbb{Z}_m$

We have studied the ring $\mathbb{Z}$ in detail in Chapter 2. You should also be familiar with the fields $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$, and their properties. In this chapter we will study other important examples of rings.

We start by studying the rings associated with the congruence (mod $m$), the rings $\mathbb{Z}_m$. In Chapter 1 we have seen briefly the cases $\mathbb{Z}_2$ and $\mathbb{Z}_3$, without making any reference to the congruence (mod $m$). However, a more systematic study of these rings requires using these congruence relations. Let us start by observing that the congruence (mod $m$) can be replaced by an equality if we “identify” (i.e., treat as unique element) all the integers which are congruent between themselves. The method we use is actually valid for any equivalence relation and consists in replacing one element by the class made of all the elements which are equivalent to it.

Let us assume then that we have fixed a modulus of congruence $m \in \mathbb{Z}$. For any $x \in \mathbb{Z}$ we introduce:

**Definition 3.1.1.** The equivalence class (mod $m$) of $x$ is the set $\overline{x}$ formed by all integers congruent with $x$ (mod $m$). In other words:

$$\overline{x} = \{ y \in \mathbb{Z} : x \equiv y \pmod{m} \}.$$ 

**Example 3.1.2.**

For the congruence with modulus $m = 3$, we have:

$$\overline{0} = \{0, \pm 3, \pm 6, \ldots \},$$
$$\overline{1} = \{1, 1 \pm 3, 1 \pm 6, \ldots \},$$
$$\overline{2} = \{2, 2 \pm 3, 2 \pm 6, \ldots \}.$$
Obviously the notation \( \overline{x} \) is ambiguous since it does not contain any information concerning the modulus of congruence \( m \) it refers to. However, it is clear that
\[
\overline{x} = \{ x + ym : y \in \mathbb{Z} \},
\]
or that:
\[
\overline{x} = \{ x + z : z \in \langle m \rangle \}.
\]
For this reason, whenever necessary to clarify the modulus of congruence, we will write \( x + \langle m \rangle \) instead of \( \overline{x} \).

Example 3.1.3.

For \( m = 4 \) and \( m = 5 \) we have, respectively,
\[
\begin{align*}
m = 4 : & \quad \overline{3} = 3 + (4) = \{ \ldots, -5, -1, 3, 7, 11, \ldots \}, \\
m = 5 : & \quad \overline{3} = 3 + (5) = \{ \ldots, -7, -2, 3, 8, 13, \ldots \}.
\end{align*}
\]

In order to replace the congruence \( x \equiv y \pmod{m} \) by an equality we will use the following lemma. Since it is a direct consequence of Proposition 2.8.2, the lemma is actually valid for any equivalence relation.

Lemma 3.1.4. For any integers \( x, y \), the following statements are equivalent:

(i) \( x \equiv y \pmod{m} \);

(ii) \( \overline{x} = \overline{y} \)

(iii) \( \overline{x} \cap \overline{y} \neq \emptyset \).

Proof. We show that \((i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (i)\).

To prove that \((i) \Rightarrow (ii)\), observe that if \( x \equiv y \pmod{m} \), the transitivity property of the congruence relation, gives that \( \overline{x} \supseteq \overline{y} \). But, by the symmetry property, \( x \equiv y \pmod{m} \) if and only if \( y \equiv x \pmod{m} \), hence also \( \overline{y} \subseteq \overline{x} \). We conclude that \( \overline{x} = \overline{y} \).

To prove that \((ii) \Rightarrow (iii)\), we observe that the reflexivity property gives that \( y \in y \). Hence, if \( \overline{x} = \overline{y} \), then \( \overline{x} \cap \overline{y} \neq \emptyset \).

Finally, to prove that \((ii) \Rightarrow (iii)\), assume that \( z \in \overline{x} \cap \overline{y} \neq \emptyset \). From the definition of equivalence class, we have that \( x \equiv z \pmod{m} \) and \( y \equiv z \pmod{m} \). By the symmetry and transitivity properties, it follows that \( x \equiv y \pmod{m} \). \( \square \)
According to the reflexivity property, an integer $x$ belongs to the class $\overline{x}$, and hence the union of all the equivalence classes is the set $\mathbb{Z}$. Moreover, according to the previous lemma, distinct equivalence classes are necessarily disjoint. For this reason we say the set of all equivalence classes for a given modulus $m$, i.e., the set $\{\overline{x} : x \in \mathbb{Z}\}$, is a partition of $\mathbb{Z}$. We will use the notation introduced in the previous chapter and call $\overline{x}$ a congruence class.

Examples 3.1.5.

1. When $m = 2$, the resulting partition of $\mathbb{Z}$ is the usual classification of the integers into even and odd numbers.

2. When $m = 3$, the resulting partition amounts to the classification of integers in terms of the remainder of the division by 3:

$$\mathbb{Z} = \{0, \pm 3, \pm 6, \ldots\} \cup \{1, 1 \pm 3, 1 \pm 6, \ldots\} \cup \{2 \pm 3, 2 \pm 6, \ldots\}.$$ 

Obviously, an equivalence class $\overline{x}$ is uniquely determined once we know one of its members. For this reason, any integer $y$ in $\overline{x}$ is called a representative of the class $\overline{x}$, since we have $y = x$.

Example 3.1.6.

If $m = 3$, the integers $1, 4, 7, -2$ and $-5$ are all representatives of $\overline{1}$, and we have

$$1 = 4 = 7 = -2 = -5.$$ 

Given an equivalence relation “$\sim$” in a set $X$, the corresponding set of equivalence classes is called the quotient of $X$ by $\sim$ and denoted by $X/\sim$. The map $\pi : X \rightarrow X/\sim$ given by $\pi(x) = \overline{x}$, which associates to each element of $X$ its equivalence class, is called the quotient map. In the case where $X = \mathbb{Z}$ and $\sim$ is the equivalence relation modulo $m$, we denote the quotient set $\mathbb{Z}/\sim$ by $\mathbb{Z}_m$, and the quotient map by $\pi_m$, or simply by $\pi$. Therefore, we have $\pi_m(x) = x + \langle m \rangle = x$. More formally:

Definition 3.1.7. For a fixed modulus $m > 0$, we set:

$$\mathbb{Z}_m = \{\overline{x} : x \in \mathbb{Z}\}, \quad \pi_m : \mathbb{Z} \rightarrow \mathbb{Z}_m \quad x \mapsto \overline{x}.$$ 

In this notation, Proposition 2.8.4 amounts simply to count the number of elements of $\mathbb{Z}_m$: 
Proposition 3.1.8. If $m > 0$, $\mathbb{Z}_m = \{0, 1, \ldots, m-1\}$ has $m$ elements.

Notice that if $m = 0$, then $\mathbb{Z}_0 = \{x\}$, and hence $\mathbb{Z}_0$ is an infinite set. Actually, for the algebraic operations that we will define shortly, $\mathbb{Z}_0$ and $\mathbb{Z}$ are isomorphic rings.

Example 3.1.9.

The set $\mathbb{Z}_6$ has 6 elements, and we can write

$$\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\} = \{6, 7, 8, 9, 10, 11\} = \{36, -5, 2, 63, 610, -19\}, \text{ etc.}$$

Notice the ambiguity in notation: when we write $\mathbb{Z}_4 = \{0, 1, 2, 3\}$, the symbols in this list denote objects that are not elements of $\mathbb{Z}_6$. For example,

$$2 + \langle 4 \rangle \neq 2 + \langle 6 \rangle \text{ i.e., } \pi_4(2) \neq \pi_6(2).$$

According to Proposition 2.8.6, if $x \equiv x' \pmod{m}$ and $y \equiv y' \pmod{m}$, then $x + y \equiv x' + y' \pmod{m}$ and $xy \equiv x'y' \pmod{m}$. In terms of equivalence classes, we have

$$\overline{x} = \overline{x}' \text{ and } \overline{y} = \overline{y}' \implies \overline{x + y} = \overline{x' + y'} \text{ and } \overline{xy} = \overline{x'y'}.$$ In other words, the classes $x + y$ and $xy$ do not depend on the choice of representatives $x$ and $y$, but only on the congruence classes $\overline{x}$ and $\overline{y}$. We can use this fact to define operations of addition and product in $\mathbb{Z}_m$.

Definition 3.1.10. The addition and product in $\mathbb{Z}_m$ are defined by

$$\overline{x} + \overline{y} = \overline{x + y}, \quad \text{ and } \quad \overline{x} \cdot \overline{y} = \overline{xy}.$$ As one could expect, some of the properties of the algebraic operations in $\mathbb{Z}$ yield similar properties for the algebraic operations in $\mathbb{Z}_m$. For example:

$$\overline{x + (y + z)} = \overline{x + y + z}.$$

so we conclude that addition in $\mathbb{Z}_m$ is associative. We leave as an exercise to check that:
3.1. THE RINGS $\mathbb{Z}_M$

**Theorem 3.1.11.** $(\mathbb{Z}_m, +, \cdot)$ is an abelian ring with identity.

Note, however, that some properties of $\mathbb{Z}$ do not transfer to the ring $\mathbb{Z}_m$.

**Example 3.1.12.**

The addition and multiplication tables in $\mathbb{Z}_4$ are:

$$
\begin{array}{cccc}
+ & 0 & 1 & 2 & 3 \\
0 & 0 & 1 & 2 & 3 \\
1 & 1 & 2 & 3 & 0 \\
2 & 2 & 3 & 0 & 1 \\
3 & 3 & 0 & 1 & 2 \\
\end{array}
\quad
\begin{array}{cccc}
\cdot & 0 & 1 & 2 & 3 \\
0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 2 & 3 \\
2 & 0 & 2 & 0 & 2 \\
3 & 0 & 3 & 2 & 1 \\
\end{array}
$$

Now observe the following differences between $\mathbb{Z}_4$ and $\mathbb{Z}$:

- The equation $x = -x$ has two solutions in $\mathbb{Z}_4$ but only one in $\mathbb{Z}$;
- Since $2 \cdot 2 = 0$ it follows that 2 is a zero divisor, so $\mathbb{Z}_4$ is not an integral domain;
- The natural multiples of the identity 1 are
  $$
  1 \cdot 1 = 1, \ 2 \cdot 1 = 2, \ 3 \cdot 1 = 3, \ 4 \cdot 1 = 4 = 0, \ etc.
  $$
  and so this ring has characteristic 4;
- There are 3 subrings of $\mathbb{Z}_4$, which are all principal ideals (as in the ring of integers):
  $$
  \langle 1 \rangle = \langle 3 \rangle = \mathbb{Z}_4, \quad \langle 2 \rangle = \{0, 2\}, \quad \text{and} \quad \langle 0 \rangle = \langle 4 \rangle = \{0\}.
  $$
  Note that these ideals correspond precisely to the divisors of 4.

Returning to the general case of the ring $\mathbb{Z}_m$, with $m > 0$, let us identify the invertible elements of $\mathbb{Z}_m$, which is denoted by $\mathbb{Z}_m^*$ (according to the notation introduced in Chapter 1). Given $a \in \mathbb{Z}$, $a$ is invertible in $\mathbb{Z}_m$ if and only if the equation $a \cdot x = 1$ has a solution in $\mathbb{Z}_m$. From the results in Section 2.8, we have that:

$$
\begin{align*}
\mathbb{Z}_m^* & \iff a \cdot x = 1 \text{ has a solution in } \mathbb{Z}_m, \\
& \iff ax \equiv 1 \pmod{m} \text{ has solutions in } \mathbb{Z}, \\
& \iff \gcd(a, m) = 1.
\end{align*}
$$

Hence, the invertible elements of $\mathbb{Z}_m$ correspond to the natural numbers $k$, $1 \leq k \leq m$, that are relatively prime to $m$. The number of elements of $\mathbb{Z}_m^*$ is denoted by $\varphi(m)$, and the resulting function $\varphi : \mathbb{N} \to \mathbb{N}$ is called the **Euler function**.
Example 3.1.13.

The invertible elements in the ring \( \mathbb{Z}_9 \) form the set

\[ \mathbb{Z}_9^* = \{1, 2, 4, 5, 7, 8\}, \]

hence, \( \varphi(9) = 6 \).

We will see later how to compute the value of the Euler function \( \varphi(m) \) from the prime factorization of \( m \). For now we observe that for any prime \( p \) we have \( \varphi(p) = p - 1 \), and that all the non-zero elements of \( \mathbb{Z}_p \) are invertible:

**Theorem 3.1.14.** If \( p \) is a prime, then \( \mathbb{Z}_p \) is a finite field with \( p \) elements.

The characteristic of the rings \( \mathbb{Z}_m \) is also very easy to find. We already observed that \( \mathbb{Z}_4 \) has characteristic 4. In fact:

**Theorem 3.1.15.** The ring \( \mathbb{Z}_m \) has characteristic \( m \).

**Proof.** Using induction, we have that

\[(3.1.1) \quad \forall n \in \mathbb{N}, a \in \mathbb{Z} : na = n \cdot a = na.\]

In particular for \( a = 1 \), we have

\[ n1 = 0 \iff n = 0 \iff n \in \langle m \rangle, \]

so the result follows. \(\square\)

We have indicated in the example above all the subrings and ideals of \( \mathbb{Z}_4 \). In order to extend this to any value of \( m \), let us start by looking at the ideals generated by each of the elements of \( \mathbb{Z}_m \). The following proposition follows immediately from the commutativity of the product in \( \mathbb{Z}_m \) and \((3.1.1)\).

**Proposition 3.1.16.** If \( a \in \mathbb{Z} \), then \( \langle a \rangle = \{a \cdot n : n \in \mathbb{Z}_m\} = \{na : n \in \mathbb{Z}\} \).

Hence, given a generator, it is very easy to list the elements of an ideal of \( \mathbb{Z}_m \).

**Examples 3.1.17.**

1. In \( \mathbb{Z}_{40} \) we have:

\[ \langle 15 \rangle = \{15, 30, 45, 5, 20, 25, 50, 10, 25, 40 = 0\}. \]
2. In $\mathbb{Z}_{21}$ we have:

$$\langle 15 \rangle = \{15, 30 = 9, 24 = 3, 18, 33 = 12, 27 = 6, 21 = 0\}.$$

Looking more carefully, one finds that the elements of the ideal $\langle a \rangle$ correspond to the integers $b$ for which the equation $ax \equiv b \pmod{m}$ has solutions. This integers are, as we know, the multiples of $d = \gcd(a, m)$:

**Proposition 3.1.18.** If $d = \gcd(a, m)$, then we have $\langle a \rangle = \langle d \rangle$ in $\mathbb{Z}_m$.

*Proof.* On the one hand, since $d = ax + my$, we have $d = ax$, so that $d \in \langle a \rangle$ and:

$$\langle a \rangle \supset \langle d \rangle.$$

On the other hand, $a = dz$ and we have $a = dz$, so that $a \in \langle d \rangle$, and:

$$\langle d \rangle \supset \langle a \rangle.$$

\[\square\]

This result gives immediately the number of elements of $\langle a \rangle$: if $d = \gcd(a, m)$, then $m = dk$, and $\langle a \rangle = \langle d \rangle$ has $k$ elements.\footnote{This shows that the number of elements of the ideal $\langle a \rangle$ is a factor of the number of elements the ring $\mathbb{Z}_m$. We will see in the next chapter that this is a special case of Lagrange's Theorem.}

**Examples 3.1.19.**

1. In $\mathbb{Z}_{40}$ we have:

$$\langle 15 \rangle = \langle 5 \rangle = \{5, 10, 15, 20, 25, 30, 35, 0\},$$

with $40 \over 8 = 8$ elements.

2. In $\mathbb{Z}_{21}$ we have

$$\langle 15 \rangle = \langle 3 \rangle = \{3, 6, 9, 12, 15, 18, 0\},$$

with $21 \over 7 = 7$ elements.

In $\mathbb{Z}_4$ all subrings are principal ideals, just like for the ring of integers. In order to see that this is a general property of the rings $\mathbb{Z}_m$, we start by relating the subrings of $\mathbb{Z}_m$ and the subrings of $\mathbb{Z}$, via the quotient map $\pi : \mathbb{Z} \rightarrow \mathbb{Z}_m$. We will see in the next chapter that this is a special case of Lagrange's Theorem.
Proposition 3.1.20. If $I$ is a subring of $\mathbb{Z}_m$ and $J = \pi^{-1}(I)$, then $I$ is a subring of $\mathbb{Z}_m$ if and only if $J$ is a subring of $\mathbb{Z}$ that contains $\langle m \rangle$.

Proof. Note that:

$$J = \pi^{-1}(I) = \{ a \in \mathbb{Z} : a \in I \}.$$

We will show that if $I$ is a subring then $J$ is also a subring, leaving the rest as an exercise.

Obviously, if $a, b \in J$, then $a, b \in I$. Hence:

$$a - b = a - b \in I \implies a - b \in J,$$

$$a \cdot b = ab \in I \implies ab \in J.$$

\[ \square \]

Corollary 3.1.21. If $I$ is a subset of $\mathbb{Z}_m$, the following statements are equivalent:

(i) $I$ is a subring of $\mathbb{Z}_m$;

(ii) $I$ is an ideal of $\mathbb{Z}_m$;

(iii) there exists $d \in \mathbb{Z}$ such that $d|m$ and $I = \langle d \rangle$.

Proof. The implications (iii)$\Rightarrow$(ii)$\Rightarrow$(i) are obvious. Therefore, it remains to prove that (i)$\Rightarrow$(iii), which we leave for the exercises.\[ \square \]

It follows from this corollary that $\mathbb{Z}_m$ has exactly one subring (which is necessarily a principal ideal) for each divisor of $m$. One can use Proposition 3.1.18 to find the generators of each of these ideals.
Example 3.1.22.

The generators of \( \langle 1 \rangle = \mathbb{Z}_{40} \) correspond to the solutions of \( \gcd(x, 40) = 1 \):

\[
\langle 1 \rangle = \langle 3 \rangle = \langle 7 \rangle = \langle 9 \rangle = \langle 11 \rangle = \langle 13 \rangle = \langle 17 \rangle = \langle 19 \rangle = \\
\langle 21 \rangle = \langle 23 \rangle = \langle 27 \rangle = \langle 29 \rangle = \langle 31 \rangle = \langle 33 \rangle = \langle 37 \rangle = \langle 39 \rangle.
\]

The generators of \( \langle 2 \rangle \) correspond to the solutions of \( \gcd(x, 40) = 2 \):

\[
\langle 2 \rangle = \langle 6 \rangle = \langle 14 \rangle = \langle 18 \rangle = \langle 22 \rangle = \langle 26 \rangle = \langle 34 \rangle = \langle 38 \rangle.
\]

The generators of \( \langle 4 \rangle \) correspond to the solutions of \( \gcd(x, 40) = 4 \):

\[
\langle 4 \rangle = \langle 12 \rangle = \langle 28 \rangle = \langle 36 \rangle.
\]

The generators of \( \langle 8 \rangle \) correspond to the solutions of \( \gcd(x, 40) = 8 \):

\[
\langle 8 \rangle = \langle 16 \rangle = \langle 24 \rangle = \langle 32 \rangle.
\]

The generators of \( \langle 5 \rangle \) correspond to the solutions of \( \gcd(x, 40) = 5 \):

\[
\langle 5 \rangle = \langle 15 \rangle = \langle 25 \rangle = \langle 35 \rangle.
\]

The generators of \( \langle 10 \rangle \) correspond to the solutions of \( \gcd(x, 40) = 10 \):

\[
\langle 10 \rangle = \langle 30 \rangle.
\]

The ideals \( \langle 20 \rangle \) and \( \langle 0 \rangle \) obviously have a unique generator.

Assume now that \( n | m \), and that \( B \) is a subring of \( \mathbb{Z}_m \) with \( n \) elements. We have the following natural questions:

- Is the additive group \( B \) isomorphic to the group \( \mathbb{Z}_n \)?
- Is the ring \( B \) always isomorphic to the ring \( \mathbb{Z}_n \)?
CHAPTER 3. OTHER EXAMPLES OF RINGS

The first of these questions is very easy to answer:

**Proposition 3.1.23.** If \( n \mid m \) and \( B \) is a subring of \( \mathbb{Z}_m \) with \( n \) elements, then the additive groups \( B \) and \( \mathbb{Z}_n \) are isomorphic.

**Proof.** Let \( m = dn \), so that \( B = \langle d \rangle \). We define \( \phi : \mathbb{Z}_n \to \mathbb{Z}_m \) by setting \( \phi(x) = dx \), where here \( x \in \mathbb{Z}_n \), and \( dx \in \mathbb{Z}_m \).

Notice that one must show that \( \phi \) is a well defined function: one must show that the right side of \( \phi(x) = dx \) does not depend on the choice of the integer \( x \) representing the class \( x \) on the left side. For that it is enough to observe that

\[
bx = x' \text{ in } \mathbb{Z}_n \iff n|(x-x') \iff m = dn|(dx-dx') \implies dx = dx' \text{ in } \mathbb{Z}_m.
\]

Now it is immediate to check that \( \phi \) is a group homomorphism and also that

\[
\phi(\mathbb{Z}_n) = \{dx : x \in \mathbb{Z}\} = \langle d \rangle = B.
\]

It remains to show that \( \phi \) is an isomorphism, i.e. that \( \phi \) is injective. For that we need only to compute its kernel:

\[
\phi(x) = 0 \iff dx = 0 \text{ (in } \mathbb{Z}_m) \iff dn = m|dx \iff n|x \iff x = 0 \text{ (in } \mathbb{Z}_n).
\]

Since the kernel of \( \phi \) is trivial, \( \phi \) is an isomorphism between \( B \) and \( \mathbb{Z}_n \).

We will see later that this result is a general property of the so called cyclic groups. The second question above, concerning ring isomorphisms is not so simple. We illustrate its complexity with a few more examples.

**Examples 3.1.24.**

1. Consider in \( \mathbb{Z}_4 \) the subring \( B = \langle 2 \rangle = \{2, 0\} \). We have just seen that \( (B, +) \simeq (\mathbb{Z}_2, +) \). However, the rings \( B \) and \( \mathbb{Z}_2 \) are not isomorphic, since the product in \( B \) is always zero: \( x, y \in B \Rightarrow xy = 0 \).

2. Consider in \( \mathbb{Z}_6 \) the subrings \( B = \langle 2 \rangle = \{2, 4, 0\} \), and \( C = \langle 3 \rangle = \{3, 0\} \). Again, we have \( (B, +) \simeq (\mathbb{Z}_3, +) \) and \( (C, +) \simeq (\mathbb{Z}_2, +) \), in this case, these are also ring isomorphisms, although this is not obvious.

The following proposition gives a complete answer to this question:

\[\text{We could also wrote, maybe more precisely, that } \phi(\pi_n(x)) = \pi_m(dx).\]
Proposition 3.1.25. If \( m = nd \) and \( B = \langle d \rangle \) is the subring of \( \mathbb{Z}_m \) with \( n \) elements, then the following statements are equivalent:

(i) The rings \( B \) and \( \mathbb{Z}_n \) are isomorphic,

(ii) The ring \( B \) is unitary,

(iii) \( \gcd(n, d) = 1 \).

In this case, the identity of \( B \) is the unique \( x \in \mathbb{Z}_m \) such that

\[
x \equiv 0 \pmod{d}, \quad \text{and} \quad x \equiv 1 \pmod{n}.
\]

Example 3.1.26.

The ring \( \mathbb{Z}_{36} \) has 9 subrings, because 36 has 9 divisors. With the exception of the trivial subrings \( \langle 1 \rangle = \{1\} \) and \( \langle 36 \rangle = \mathbb{Z}_{36} \), only the subrings \( B = \langle 4 \rangle \), with 9 elements, and \( C = \langle 9 \rangle \), with 4 elements, have identity.

The solution of the system \( x \equiv 0 \pmod{4} \) and \( x \equiv 1 \pmod{9} \) is \( x \equiv 28 \pmod{36} \), and hence the identity of \( B \) is \( x = 2\overline{8} \). Similarly, the solution of \( x \equiv 0 \pmod{9} \) and \( x \equiv 1 \pmod{4} \) is \( x \equiv 9 \pmod{36} \), and hence the identity of \( C \) is \( x = \overline{9} \).

Exercises.

1. Prove Theorem 3.1.11

2. Show that the only subrings of \( \mathbb{Z}_4 \) are \( \langle 1 \rangle, \langle 2 \rangle \) and \( \langle 0 \rangle \).

3. Show that, if \( m > 1 \), then \( \mathbb{Z}_m \) is either a field or has zero divisors.

4. Show that \( na = na \) (in particular, \( n = n \\overline{1} \)).

5. Show that, if \( n > 1 \), then \( M_n(\mathbb{Z}_m) \) is a non-abelian ring, with characteristic \( m \), and \( m^2 \) elements.

6. Consider the ring consisting of all maps \( f : \mathbb{Z} \to \mathbb{Z}_m \) and show that for any \( m > 1 \) there exist infinite rings with characteristic \( m \).

7. Show that \( A \in M_n(\mathbb{Z}_m) \) is invertible if and only if \( \det(A) \in \mathbb{Z}_m^3 \)

---

\( ^3 \) The determinant of a matrix \( A = (a_{ij}) \) of size \( n \times n \) with entries in a commutative ring is defined as usual by:

\[
\det A = \sum_{\pi \in S_n} \text{sgn}(\pi)a_{1\pi(1)} \cdots a_{n\pi(n)}.
\]
CHAPTER 3. OTHER EXAMPLES OF RINGS

8. Give an example of a finite vector space over a finite field (compare $\mathbb{R}^n$ with $\mathbb{Z}_p^n$).

9. Determine all matrices in $GL(2, \mathbb{Z}_2)$ ($2 \times 2$ invertible matrices with entries in $\mathbb{Z}_2$).

10. What is the cardinal of $GL(n, \mathbb{Z}_p)$, if $p$ is prime? (Hint: Note that the rows of the matrix $M \in GL(n, \mathbb{Z}_p)$, that are vectors of $\mathbb{Z}_p^n$, must be linearly independent.)

11. Find the inverse of the matrix
   \[
   \begin{pmatrix}
   1 & 0 & 0 \\
   2 & 3 & 4 \\
   3 & 2 & 4
   \end{pmatrix}
   \in GL(3, \mathbb{Z}_5).
   \]

12. Solve the system
   \[
   \begin{align*}
   x + 2y &= a \\
   -3x + 3y &= b
   \end{align*}
   \]
   in $\mathbb{Z}_5$.

13. Show that $\langle a \rangle = \mathbb{Z}_m$ if and only if $a \in \mathbb{Z}_m^*$.


15. Find all the elements of the ideal $\langle 85 \rangle \subset \mathbb{Z}_{204}$. Which ones are generators of this ideal?

16. How many elements does the ideal $\langle 28, 52 \rangle \subset \mathbb{Z}_{204}$ have?

17. Determine all the ideals of $\mathbb{Z}_{30}$. Which of these ideals are unitary rings, and what are the respective identities?

18. If $p$ is a prime number, and $n \in \mathbb{N}$, show that $\varphi(p^n) = p^n - p^{n-1}$. Hint: Show that $x \in \mathbb{Z}_{p^n}^* \iff x \notin \langle p \rangle$.

19. Find $\varphi(3000)$. Hint: Show that $x \in \mathbb{Z}_{3000}^* \iff x \notin \langle 2 \rangle \cup \langle 3 \rangle \cup \langle 5 \rangle$.

20. Assume that $d = \gcd(a, m)$, $m = dn$, and $\phi : \mathbb{Z}_m \to \mathbb{Z}_m$ is given by $\phi(x) = ax$.
   (a) Show that $\phi$ is a homomorphism of groups.
   (b) Show that the kernel of $\phi$ is $\langle a \rangle$, and $\phi(\mathbb{Z}_m)$ has $n$ elements.
3.2. FRACTIONS AND RATIONAL NUMBERS

(c) Assuming that $m = 12$, for which $\mathfrak{a}$ is $\phi$ a group automorphism?
(d) Assuming that $m = 12$, for which $\mathfrak{a}$ is $\phi$ a ring homomorphism?

21. Assuming that $n|m$, show that the function $\phi : \mathbb{Z}_m \rightarrow \mathbb{Z}_n$ given by $\phi(\mathfrak{a}) = \mathfrak{x}$, i.e., $\phi(\pi_m(x)) = \pi_n(x)$, with $x \in \mathbb{Z}$, is well defined, and that it is a homomorphism of rings. What is its kernel?

22. Prove Proposition 3.1.25 proceeding as follows:
(a) Show that “(i) $\implies$ (ii)”.
(b) Show that “(ii) $\implies$ (iii)”, by showing first that if $\mathfrak{a}$ is the identity of $B$ then $(\mathfrak{a}) = (\mathfrak{d})$, and $\mathfrak{a}^2 = \mathfrak{a}$. Conclude that $a \equiv 0 \pmod{d}$, and $a \equiv 1 \pmod{n}$.
(c) Solve the system $a \equiv 0 \pmod{d}$, and $a \equiv 1 \pmod{n}$, and consider the function $\phi : \mathbb{Z}_n \rightarrow \mathbb{Z}_m$ given by $\phi(\mathfrak{x}) = \mathfrak{a} \mathfrak{x}$. Show that $\phi$ is well defined, is an injective ring homomorphism, and $\phi(\mathbb{Z}_n) = B$, therefore completing the proof.

23. Consider maps $\phi : \mathbb{Z}_4 \rightarrow \mathbb{Z}_{36}$.
(a) Find the group homomorphisms $\phi$. Which ones are injective?
(b) List the ring homomorphisms $\phi$. Which ones are injective?

24. Assume that $\mathfrak{a} \in \mathbb{Z}_m^*$, and consider $\Psi : \mathbb{Z}_m^* \rightarrow \mathbb{Z}_m$ given by $\Psi(\mathfrak{a}) = \mathfrak{a} \mathfrak{a}$.
(a) Show that $\Psi$ is injective, and that in fact $\Psi(\mathfrak{a}) \in \mathbb{Z}_m^*$, for all $\mathfrak{a} \in \mathbb{Z}_m^*$.
(b) Writing $\mathbb{Z}_m^* = \{x_1, x_2, x_3, \ldots, x_k\}$, where $k = \phi(m)$ and $\phi$ is the Euler function, show that $\prod_{i=1}^{k} \Psi(x_i) = \prod_{i=1}^{k} x_i$, and use this fact to prove the THEOREM OF EULER : $\mathfrak{a}^k = 1$.
(c) Prove also the following THEOREM OF FERMAT : If $m = p$ is prime, then $\mathfrak{a}^p = \mathfrak{a}$.

3.2 Fractions and Rational Numbers

Our main aim in this section is to define the field of rational numbers and to show that its properties (often introduced in an axiomatic form) are logical consequences of the axioms of the integers studied in Chapter 2. At the same time, we will see that the method to define the rational numbers from the integer numbers can be applied to any abelian ring where the cancellation law holds. This will allow us to define later other important fields.
The rational numbers (fractions, roots, etc.) often are informally defined as “expressions of the type \( \frac{m}{n} \), where \( m \) and \( n \) are integers, and \( n \neq 0 \). From a more formal point of view, we can observe that an ordered pair of integers \((m,n)\) determines a rational number, provided that \( n \neq 0 \). On the other hand, we know that distinct ordered pairs can correspond to the same rational number: we can have \((m,n)\neq(m',n')\) and \(\frac{m}{n} = \frac{m'}{n'}\). This happens precisely when \(mn' = m'n\).

These remarks suggest that instead of defining the rational numbers as ordered pairs of integers we can define them as equivalence classes of ordered pairs of integers. As we will see, the fact that this works does not depend on specific properties of the integers, but only on the fact that \(\mathbb{Z}\) is an abelian ring with more than one element, where the cancellation law holds. Therefore, we will formalize this construction in this more general abstract setting.

In what follows, \(A\neq\{0\}\) denotes an abelian ring where the cancellation law for the product is valid (i.e., without zero divisors).

**Definition 3.2.1.** Let \(B = \{(a,b) : a,b \in A, \text{ and } b \neq 0\}\). We denote by “\(\sim\)” the binary relation in \(B\) defined by

\[(a,b) \sim (a',b') \iff ab' = a'b.\]

The relation just defined is, of course, suggested by the equality of fractions that we mentioned before. Its usefulness depends on being an equivalence relation, which we now check.

**Lemma 3.2.2.** The relation \(\sim\) is an equivalence relation.

**Proof.** Obviously, the relation “\(\sim\)” is reflexive and symmetric. To check that is also transitivity, assume that \((a,b) \sim (a',b')\), and that \((a',b') \sim (a'',b'')\). Using the commutativity and associativity of the product, we find:

\[(a',b') \sim (a'',b'') \iff a'b'' = a''b' \implies a'b'b'' = a''b'b',\]

\[(a,b) \sim (a',b') \iff ab' = a'b \implies a'b'b'' = ab'b'.\]

We conclude that \(a''b'b' = ab''b'\) and, hence, that \(a''b = ab''\), e \((a,b) \sim (a'',b'')\), since \(b' \neq 0\) and \(A\) has no zero divisors.

An equivalence relation “\(\sim\)” in \(B\) determines a partition of \(B\) into equivalence classes. If \((a,b) \in B\), we call the corresponding equivalence class \((a,b)\) a fraction of \(A\) and denoted it by \(a/b\), or by \(\frac{a}{b}\). More formally:
Definition 3.2.3. Let $A \neq \{0\}$ be an abelian ring where the cancellation law holds.

(i) If $a, b \in A$ and $b \neq 0$, the fraction $\frac{a}{b}$ is defined by

$$\frac{a}{b} = \{(a', b') \in A \times A : b' \neq 0 \text{ and } ab' = a'b\};$$

(ii) The set of all fractions $\frac{a}{b}$ is denoted by $\text{Frac}(A)$;

(iii) When $A = \mathbb{Z}$, a fraction $\frac{a}{b}$ is called a rational number, and the set $\text{Frac}(\mathbb{Z})$ is denoted by $\mathbb{Q}$.

When $A = \mathbb{Z}$, the set of rational numbers $\text{Frac}(\mathbb{Z}) = \mathbb{Q}$ is a ring. For the general case, we define on the set of fractions of $A$ algebraic operations of addition and product by copying simply the usual operations with rational numbers. These are given by:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, \quad (3.2.1)$$
$$\frac{ac}{bd} = \frac{ad}{bd}, \quad (3.2.2)$$

In order to formalize these definitions, which represent binary operations on equivalence classes, we must check that each of this operations is independent of the choice of representative for each class. We state this in the following lemma, whose proof is left as an exercise:

Lemma 3.2.4. If $(a, b) \simeq (a', b')$, and $(c, d) \simeq (c', d')$, then

$$(ad + bc, bd) \simeq (a'd' + b'c', b'd'),$$

$$ac, bd \simeq (a'c', b'd').$$

According to this lemma, (3.2.1) and (3.2.2) do indeed define addition and product operations on the set of fractions. It is obvious that both operations are commutative, and that the product is associative. With a bit more work, one can verify that addition is associative and that the product is distributive relative to addition. The existence of identities for both operations also does not offer any difficulties. Indeed, just like for the rationals, we have:

Theorem 3.2.5. The set $\text{Frac}(A)$ with the algebraic operations (3.2.1) and (3.2.2) is a field, called the field of fractions of $A$. 

Proof. Let \( b \neq 0 \) be any non-zero element in \( A \). We set \( 0' = \frac{0}{b} \) and \( 1' = \frac{b}{b} \). We leave as an exercise to check:

\[
0' = \{(0, c) : c \neq 0\} \quad \text{and} \quad 1' = \{(c, c) : c \neq 0\}
\]

are, respectively, identities for the addition and the product in \( \text{Frac}(A) \). Note-se that the elements \( 0' \) and \( 1' \) are independent of the choice of \( b \neq 0 \) in \( A \). In particular, and assuming that \( y \neq 0 \), we have:

\[
\frac{x}{y} = 0' \iff x = 0, \quad \text{and} \quad \frac{x}{y} = 1' \iff x = y.
\]

Note also that the existence of an identity for the product of fractions does not depend of the existence of an identity for the product in the original ring \( A \), and observe finally that

\[
\frac{a}{b} + \frac{(-a)}{b} = 0',
\]

and that if \( \frac{a}{b} \neq 0' \), then \( a \neq 0 \), so \( \frac{b}{a} \) is a fraction, and

\[
\frac{a b}{b a} = 1'.
\]

\( \square \)

When \( A = \mathbb{Z} \) is the ring of integers and \( \text{Frac}(A) = \mathbb{Q} \), we are used to consider \( \mathbb{Z} \) as subring of \( \mathbb{Q} \). For an arbitrary ring this “identification” is still possible: the field of fractions of a ring \( A \) contains a subring isomorphic to the ring \( A \).

**Proposition 3.2.6.** Let \( a \neq 0 \) be a fixed element in the ring \( A \). Then:

(i) \( \iota : A \to \text{Frac}(A) \), \( x \mapsto \frac{ax}{a} \) is a ring isomorphism between \( A \) and \( \iota(A) \);

(ii) the isomorphism \( \iota \) is independent of the choice of element \( a \in A - \{0\} \).

**Proof.** The identities \( \iota(x + y) = \iota(x) + \iota(y) \) and \( \iota(xy) = \iota(x)\iota(y) \) are easily verified, so \( \iota \) is a homomorphism of rings and \( \iota(A) \) is a subring of \( \text{Frac}(A) \). Moreover, \( \iota \) is an isomorphism between \( A \) and \( \iota(A) \) since:

\[
\iota(x) = 0' \iff \frac{ax}{a} = 0' \\
\iff ax = 0 \\
\iff x = 0.
\]
On the other hand, if \(a, a' \neq 0\) and \(x \in A\), then \(\frac{ax}{a} = \frac{a'x}{a'}\), because \((ax)a' = (a'x)a\). Hence, \(\iota\) is independent of the choice of \(a \in A - \{0\}\).

According to this result, the elements of \(\text{Frac}(A)\) of the form \(\frac{ax}{a}\), with \(a, x \in A\) and \(a \neq 0\) fixed, are “copies” of the elements \(x \in A\). For this reason, whenever one works with elements of the field \(\text{Frac}(A)\), the fraction \(\frac{ax}{a}\) is denoted by “\(x\)”, and one says that \(x\) is an element of the ring \(A\). We also write \(\text{Frac}(A) \supset A\), so we consider \(\text{Frac}(A)\) as an extension of the ring \(A\). This abuse of language is fully justified by the proposition above, and allows for a much lighter notation. For the same reason, we will use the same symbol \(0\) to denote the zeros of \(A\) and \(\text{Frac}(A)\). Similarly, when \(A\) as identity \(1\) we use the same symbol to denote the identity of \(\text{Frac}(A)\).

A similar issue arises when one considers “fractions of fractions”. When dealing with rational numbers, we have the equality of fractions

\[
\frac{a/b}{c/d} = \frac{ad}{bc}.
\]

However, according to our formal definition these two fractions belong to two different rings: the fraction \(\frac{a/b}{c/d}\) is an element of \(\mathbb{Q}\), while the fraction \(\frac{ad}{bc}\) is an element of \(\text{Frac}(\mathbb{Q})\). Hence, this equality cannot hold literally. The sense in which it holds is explained in the following proposition:

**Proposition 3.2.7.** If \(K\) is a field, the map \(\iota : K \to \text{Frac}(K)\) defined in Proposition 3.2.6 is surjective, and hence is a ring isomorphism.

**Proof.** The proof reduces to the observation that \(\frac{x}{y} = \iota(xy^{-1})\), which according with the conventions above is written in the form “\(\frac{x}{y} = xy^{-1}\)”. \(\square\)

In the case of the composite fraction above \((x = \frac{a}{b}\) and \(y = \frac{c}{d}\)), strictly speaking, we have that:

\[
\frac{a/b}{c/d} = \iota \left( \frac{a}{b} \left( \frac{c}{d} \right)^{-1} \right) = \iota \left( \frac{ad}{bc} \right).
\]

Using the identification of \(x\) with \(\iota(x)\), we find that \(\frac{a/b}{c/d} = \frac{ad}{bc}\).

In the previous section we saw that the existence of the finite fields \(\mathbb{Z}_p\) follows from the axioms for the integers. We have just seen that the existence of the field \(\mathbb{Q}\) is another consequence of those axioms. We will see later that these fields are, in some sense, the smallest fields: we will show that any field of characteristic \(p\) contains necessarily a subfield isomorphic to \(\mathbb{Z}_p\) (if \(p \neq 0\)) or isomorphic to \(\mathbb{Q}\) (if \(p = 0\)).
CHAPTER 3. OTHER EXAMPLES OF RINGS

Exercises.

1. Prove Lemma 3.2.4.

2. Complete the proof of Theorem 3.2.5.

3. What is the field of fractions of the ring of Gaussian integers?

4. If Frac(A) is isomorphic to Frac(B), is it true that A is isomorphic to B? If A is isomorphic to B, is it true that Frac(A) is isomorphic to Frac(B)?

5. Show that, if a field K is an extension of a ring A, then K contains a subfield isomorphic to Frac(A) (this result, generalizes Proposition 3.2.6 that Frac(A) is the smallest field that contains A). (Hint: If K is a field, the intersection of all subfields of K is the smallest subfield of K.)

6. Show that, if A is countable, then Frac(A) is countable.

7. Assume that A is an integral domain. Show that then the characteristics of Frac(A) and A are equal.

8. Show that Q can only be ordered so that: \( \frac{m}{n} > 0 \Leftrightarrow mn > 0 \). (Hint: In an ordered field one always has \( x > 0 \Leftrightarrow x^{-1} > 0 \).)

9. Show that, if A is an ordered ring, then Frac(A) is also an ordered field.

3.3 Polynomials and Power Series

The polynomials with real coefficients are often defined as the maps \( p : \mathbb{R} \to \mathbb{R} \) of the form

\[
p(x) = \sum_{n=1}^{N} p_n x^n,
\]

where the real numbers \( p_n \) are called the coefficients of the polynomial \( p \). However, one cannot define in this way polynomials with coefficients in a arbitrary ring \( A \) if we want to keep the property that polynomials with distinct coefficients correspond to distinct polynomials. In fact, if \( A \) has more than one element, there is an infinite number of choices for the coefficients of polynomials. However, if \( A \) is finite, there are only a finite number of maps \( f : A \to A \), so these cannot be used to define all the polynomials with coefficients in \( A \).
Example 3.3.1.

The set underlying the ring $\mathbb{Z}_2$ is \{0, 1\}, so has two elements. Therefore the set of maps $f : \mathbb{Z}_2 \to \mathbb{Z}_2$ has 4 elements. On the other hand, if polynomials with distinct coefficients correspond to distinct polynomials, there exist an infinite number of polynomials with coefficients in $\mathbb{Z}_2$.

This problem is solved identifying a polynomial with the sequence of its coefficients. We will come back later to the issue of defining an associated map. In fact, we will consider polynomials (which have a finite number of non-zero coefficients) as a special case of a “power series”. The last ones will be defined without any reference to convergence issues, so they will be “formal” power series.

In what follows, we denote by $A$ an abelian ring with identity 1.

Definition 3.3.2. A (formal) power series in $A$ is a map $s : \mathbb{N}_0 \to A$. A power series is called a polynomial if and only if there exists a $N \in \mathbb{N}_0$ such that $s(n) = 0$ for all $n > N$.

Examples 3.3.3.

1. The following polynomials with coefficients in $A$ have special roles:
   - $0 = (0, 0, 0, \ldots)$, called the zero polynomial;
   - $1 = (1, 0, 0, \ldots)$, called the identity polynomial;
   - $x = (0, 1, 0, \ldots)$, which we will call the indeterminate $x$.

2. The polynomial $a = (a, 0, 0, \ldots)$, where $a \in A$, is called a constant polynomial. In particular, the zero and identity polynomials are constant polynomials.

3. The set of all power series in $\mathbb{Z}_2$ is infinite and non-countable (it should be obvious that it is a set in bijection with $\mathcal{P}(\mathbb{N})$), and the set of all polynomials in $\mathbb{Z}_2$ is infinite but countable.

The terms $s(0), s(1), s(2), \ldots$ of a formal power series $s : \mathbb{N}_0 \to A$ are called the coefficients of the series. In order to avoid confusion with the values of a function associated with the power series, we will always denote these coefficients by $s_0, s_1, s_2,$ etc. We will denote by the symbols $A[[x]]$ and $A[x]$, respectively, the sets of power series and polynomials, respectively.
with coefficients in $A$. Obviously, $A[x] \subset A[[x]]$.\footnote{Note that the use of the letter “$x$” in the symbols $A[x]$ and $A[[x]]$ is caused by the choice of this letter to denote the indeterminate $(0,1,0,\ldots)$. In some situations we may denote this indeterminate by some other letter, e.g., $y$, in which case we will use the notation $A[y]$ or $A[[y]]$.}

The addition and product of polynomials with real coefficients is certainly familiar to us. For example, if we consider polynomials of degree 2, we have:

$$(a_0 + a_1 x + a_2 x^2) + (b_0 + b_1 x + b_2 x^2) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2$$

$$= (a_0 + a_1 x + a_2 x^2)(b_0 + b_1 x + b_2 x^2), = (a_0b_0) + (a_0b_1 + a_1b_0)x$$

$$+(a_0b_2 + a_1b_1+a_2b_0)x^2 + (a_1b_2 + a_2b_1)x^3 + (a_2b_2)x^4.$$

Our next definition amounts to recognize that the operations over the coefficients of the polynomials, on the right hand side of the previous identities, make sense in any ring. Notice that the resulting addition operation is just the usual addition of sequences, while the product is distinct. When there is some risk of ambiguity, we will call the product defined below the convolution product, and denote it by $s \star t$ instead of $st$.

**Definition 3.3.4.** Let $s, t : \mathbb{N}_0 \to A$ be power series, the sum $s + t$ and the (convolution) product $s \star t$ are the power series:

\begin{align*}
(s + t)_n &= s_n + t_n, \text{ and,} \tag{3.3.1} \\
(s \star t)_n &= \sum_{k=0}^{n} s_k t_{n-k}. \tag{3.3.2}
\end{align*}

**Examples 3.3.5.**

1. If $a = (a,0,0,\ldots)$ and $b = (b,0,0,\ldots)$ are constant polynomials, their sum and product are given by $a + b = (a + b,0,0,\ldots)$ and $a \star b = ab = (ab,0,0,\ldots)$. Hence, the set of constant polynomials is a ring isomorphic to the ring $A$.

2. If $a = (a,0,0,\ldots)$ is constant and $s = (s_0,s_1,s_2,\ldots)$ is any power series, the product $a \star s$ is the power series $(as_0,as_1,as_2,\ldots) = as$, since the sum $\sum_{k=0}^{n}a_k s_{n-k}$ reduces the term with $k = 0$.

3. Consider the power series $s = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \ldots)$ with coefficients in $\mathbb{Z}_2$. In order

to compute its square, note that $(ss)_n = \sum_{k=0}^{n} s_k s_{n-k} = \sum_{k=0}^{n} 1 = n + 1$. We conclude that

$$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \ldots)(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \ldots) = (1,0,1,0,\ldots).$$
4. If \( s = (s_0, s_1, s_2, \ldots) \) is any power series, the product \( xs \) is the power series obtained from \( s \) by translation of all its coefficients to the right:

\[
(xs)_0 = x_0s_0 = 0,
\]

\[
(xs)_{n+1} = \sum_{k=0}^{n+1} x ks_{n+1-k} = s_n.
\]

When \( s = x \), we conclude that \( x^2 = (0, 0, 1, 0, \ldots) \), \( x^3 = (0, 0, 0, 1, 0, \ldots) \), etc. We extend this to the case \( n = 0 \) by stipulating by convention that \( x^0 = (1, 0, 0, \ldots) = 1 \).

The next result has a straightforward proof, and so it is left as an exercise.

**Theorem 3.3.6.** If \( A \) is an abelian ring with identity, then both \( A[[x]] \) and \( A[x] \) are abelian rings with identity for the sum and product defined by (3.3.1) and (3.3.2), and \( A[x] \) is a subring of \( A[[x]] \).

We have already observed that the constant polynomials form a ring isomorphic to \( A \). Therefore, this justifies using the same symbol \( a \) to denote an element of the ring \( A \), and the corresponding constant polynomial \((a, 0, 0, \ldots)\). We also say that \( A[x] \) and \( A[[x]] \) are extensions of \( A \). Notice that the expression \( ax^n \) is then representing the product of the constant polynomial \((a, 0, 0, \ldots)\) by the \( n \)-th power of the indeterminate \( x \), i.e., \( x^n \).

This polynomial, according to what we saw above, has only one non-zero coefficient namely \((ax^n)_n = a\) (note that here, \( a \) has two distinct meanings!) We conclude that if \( p = (p_0, p_1, p_2, \ldots, p_N, 0, \ldots) \), then

\[
p = p_0 + p_1x + \cdots + p_Nx^N = \sum_{n=0}^{N} p_nx^n.
\]

The expression on the right hand side is called the **canonical form** of the polynomial \( p \). As usual, the coefficient is omitted whenever it equals 1.

**Example 3.3.7.**

Assume that \( p, q \in \mathbb{Z}_4[x] \) are given by \( p = \frac{1}{1} + x + 2x^2 \) and \( q = \frac{1}{1} + 2x^2 \).

In order to add and multiply these polynomials, we proceed exactly as we do

---

\(5\) The ring of power series \( A[[x]] \) and the ring of sequences in \( A \) have the underlying set, and the same addition operation, differing only on the multiplication operation. Whenever is necessary to distinguish the power series \( s \) in \( A[[x]] \) from the corresponding sequence \( a \) in \( S \), one often says that \( s \) is the \( z \)-transform of \( a \).
for polynomials with real coefficients, since the usual procedure only uses the
algebraic properties common to any ring. One can easily recognize these apart
from the specific properties of the ring $\mathbb{Z}_4$ in the following computations:

\[
(1 + x + 2x^2) + (1 + 2x^2) = (1 + 1) + (1 + 0)x + (2 + 2)x^2 = 2 + x,
\]

\[
(1 + x + 2x^2)(1 + 2x^2) = (1 + x + 2x^2)1 + (1 + x + 2x^2)2x^2 = (1 + x + 2x^2) + (2x^2 + 2x^3) = 1 + x + 2x^3.
\]

Sometimes it is possible to give a meaning to “infinite sums” $\sum_{n=0}^{\infty} s_n$, where each $s_n$ is a power series. If $s_{nk}$ denotes the coefficient $k$ of the power series $s_n$, the simplest meaning one can give to the “sum” $t = \sum_{n=0}^{\infty} s_n$ is

\[
t = \sum_{n=0}^{\infty} s_n \iff t_k = \sum_{n=0}^{\infty} s_{nk}, \text{ for all } k \in \mathbb{N}_0.
\]

We will always use this definition whenever the sequence $s_{nk}$ is eventually
zero for all $k \geq 0$, interpreting the “infinite sum” on the right and the (finite)
sum of all the non-zero terms. In particular, we will write for any power
series $s$:

\[
s = \sum_{n=0}^{\infty} s_n x^n.
\]

**Definition 3.3.8.** If $p \neq 0$ is a polynomial, the **degree** of $p$ is the integer
$\deg p$ defined by

\[
\deg p = \max\{n \in \mathbb{N}_0 : p_n \neq 0\}.
\]

When $p = 0$ we convention that $\deg p = -\infty$.

**Example 3.3.9.**

Obviously, $\deg x^n = n$, for any $n \geq 0$.

The example above with the product of polynomials in $\mathbb{Z}_4[x]$ shows that
it is not always true that the degree of the product of two polynomials is
the sum of of the degrees of the factors. The next result clarifies this issue,
which is related with the presence of zero divisors. In order to avoid dealing
separately with the zero polynomial, we convention that

\[
\deg p + \deg q = -\infty, \text{ whenever } p = 0 \text{ or } q = 0.
\]
Proposition 3.3.10. Let $p, q \in A[x]$. Then:

(i) $\deg(p + q) \leq \max\{\deg p, \deg q\}$, and $\deg(pq) \leq \deg p + \deg q$;

(ii) If $A$ is an integral domain, then $\deg(pq) = \deg p + \deg q$, and $A[x]$ and $A[[x]]$ are also integral domains.

We leave the proof as an exercise.

Notice that, according to (ii), when $A$ is an integral domain we can form the field of fractions of $A[x]$ (i.e., $\text{Frac}(A[x])$) in the notation of the previous section). This field of fractions if the formal abstract analogue of the usual field of rational functions.

Definition 3.3.11. If $A$ is an integral domain, $A(x)$ (respectively, $A((x))$) denotes the field of fractions de $A[x]$ (respectively, $A[[x]]$).

Example 3.3.12.

When $A = \mathbb{Z}$, the field $\mathbb{Z}(x)$ consist of fractions whose numerators and denominators are polynomials with integer coefficients. In this field, we have

$$\frac{x^2 - 1}{x + 1} = x - 1.$$

Notice, however, that the functions $f(x) = \frac{x^2 - 1}{x + 1}$ and $g(x) = x - 1$ are not the same, since their domains of definition are distinct.

If $\phi : A \to B$ is a ring homomorphism, then the map $\Phi : A[x] \to B[x]$ defined by

$$\Phi \left( \sum_{n=0}^{N} p_n x^n \right) = \sum_{n=0}^{N} \phi(p_n) x^n$$

is also a ring homomorphism. Frequently, for $p \in A[x]$, we will denote the polynomial $q = \Phi(p)$ by $p^\phi(x)$. Sometimes (example, if $A \subset B$ is a subring and $\phi : A \to B$ is the inclusion) we may use the same letter to denote this two polynomials, if it is clear from the context to which ring the coefficients belong to. The next examples shows that this is a reasonable practice, which we are already used to!

Example 3.3.13.

Let $p = 1 + 2x + 3x^2 \in \mathbb{Z}[x]$ be a polynomial with integer coefficients. Obviously, $\mathbb{Z} \subset \mathbb{Q}$ and we can consider the canonical inclusion $\iota : \mathbb{Z} \to \mathbb{Q} = \text{Frac}(\mathbb{Z})$ and we have $p' = 1 + 2x + 3x^2 \in \mathbb{Q}[x]$. Notice that the symbols “1”, “2”
and “3” denote now rational numbers, rather than integers. This is a reasonable abuse of notation, as we saw in our treatment of the field of fractions: the fraction $\frac{ax}{a}$ is denoted by $x$. Therefore, it is natural to represent this new polynomial by the same letter as the original polynomial.

Finally, observe that if $\phi: A \to B$ is an injective homomorphism of rings and $p \in A[x]$, then $p$ and $p^\phi$ have the same degree.

**Exercises.**

1. Compute the following product in $\mathbb{Z}_m[x]$, when $m = 2, 3$ and 6:

   $$(1 + x + 2x^2)(2 + 3x + 2x^2).$$

2. Proof Theorem 3.3.6

3. Show that the ideal $\langle 2, x \rangle$ in $\mathbb{Z}[x]$ is not a principal ideal.

4. Give a proof of Proposition 3.3.10

5. What are the invertible elements of $A[x]$, when $A$ is an integral domain?

6. Assume that $A$ is a commutative ring with identity. Show that the characteristic of $A[x]$ and the characteristic of $A$ are equal.

7. The polynomials in several variables can be defined in several ways. For example, in the case of two variables, we can consider:

   (i) The ring $A[x]$ and the polynomials with coefficients in $A[x]$, which we denote by $A[x][y]$ (in this case we denote the new indeterminate by $y$);

   (ii) The maps $s: \mathbb{N}_0 \times \mathbb{N}_0 \to A$, with an appropriate definition of sum and convolution product, and indeterminates $x$ and $y$, giving a ring which is denoted by $A[x, y]$.

   Both ways of defining polynomials in several variables are useful and the following exercises show that they are equivalent.

   (a) Give a detailed description of the definition sketched in (ii).

   (b) Show that the two definitions are equivalent, i.e., they lead to isomorphic rings.

   (c) How should one generalize these definitions for polynomials in several variables $x_1, \ldots, x_n$?

   (d) How should one generalize these definitions for polynomials in the variables $x_\alpha$, with $\alpha \in I$, where $I$ can be an infinite set?
3.3. POLYNOMIALS AND POWER SERIES

8. Show that the following statements are equivalent:
   (a) $A$ is an integral domain;
   (b) $A[x]$ is an integral domain;
   (c) $A[[x]]$ is an integral domain;
   (d) For any $p, q \in A[x]$, $\deg(pq) = \deg p + \deg q$.

9. Show that there exist rings with characteristic $p$, a prime, which are not fields, and fields with characteristic $p$ which are infinite (both countable and uncountable).

10. If $p \in A[x]$ is a polynomial $p = \sum_{n=0}^{N} p_n x^n$, its (FORMAL) DERIVATIVE $p'$ is the polynomial
    \[ p' = \sum_{n=1}^{\infty} np_n x^{n-1}. \]
    Is it true that $p' = 0$ implies that $p$ is a constant polynomial?

11. Use the previous exercise and the homomorphism $\phi : \mathbb{Z} \to \mathbb{Z}_p$ defined by $\phi(n) = \bar{n}$ to deduce the following generalization of Euclides Lemma: if $p \in \mathbb{N}$ is a prime, $a, b \in \mathbb{Z}[x]$ and $p | ab$, then either $p | a$ or $p | b$.

12. Let $D$ be an integral domain and $K = \text{Frac}(D)$ its field of fractions. Show that $D(x)$ is isomorphic to $K(x)$.

13. Define a ring, denoted $A[1/x][x]$, with underlying set consists of all power series of the form $\sum_{n=k}^{\infty} s_n x^n$, where $k \in \mathbb{Z}$ is arbitrary\(^6\). Show that if the coefficients belong to a field $K$, then the ring $K[1/x][x]$ is a field isomorphic to $K((x)) = \text{Frac}(K[[x]])$.

14. Show that in $K[[x]]$ one has
    \[ \frac{1}{1-x} = \sum_{n=0}^{\infty} x^n. \]
    (Notice the abuse of notation, which is justified by the previous exercise!)

15. Determine the power series inverse to $(a-x)$ and to $(a-x)(b-x)$ in $K[[x]]$.

16. Show that in $\mathbb{Z}_2[[x]]$, one has
    \[ \frac{1}{(1-x)^2} = \sum_{n=0}^{\infty} x^{2n}. \]

\(^6\) These series are usually called \textbf{Laurent series}. The corresponding ring, denoted $A[1/x][x]$, is called the ring of Laurent series with coefficients in $A$. The special case where $A = \mathbb{C}$ plays a crucial role in Complex Analysis and in Algebraic Geometry.
17. One can solve problems like the one of the Fibonacci sequence in Section 2.3 by formal power series computations. Recalling that this sequence is defined recursively by:
\[ a_{n+2} = a_{n+1} + a_n, \quad (n \geq 0), \]
one notes that is equivalent to
\[ \sum_{n=0}^{\infty} a_{n+2} x^n = \sum_{n=0}^{\infty} a_{n+1} x^n + \sum_{n=0}^{\infty} a_n x^n, \]
and also:
\[ \sum_{n=0}^{\infty} a_{n+2} x^{n+2} = x \sum_{n=0}^{\infty} a_{n+1} x^{n+1} + x^2 \sum_{n=0}^{\infty} a_n x^n. \]
If we let \( s = \sum_{n=0}^{\infty} a_n x^n \), then
\[ \sum_{n=0}^{\infty} a_{n+2} x^{n+2} = s - a_0 - a_1 x, \quad \sum_{n=0}^{\infty} a_{n+1} x^{n+1} = s - a_0, \]
and we conclude that the recurrence relation above is equivalent to:
\[ s - a_0 - a_1 x = x(s - a_0) + x^2 s. \]
Solving this equation for \( s \), we obtain:
\[ s = \frac{-a_0 + (a_0 - a_1)x}{x^2 + x - 1}. \]
If \( \alpha \) and \( \beta \) are the roots of \( x^2 + x - 1 \), we can factorize this rational fraction as:
\[ \frac{-a_0 + (a_0 - a_1)x}{x^2 + x - 1} = \frac{A}{\alpha - x} + \frac{B}{\beta - x}. \]
The exercise above now allows to compute explicitly the coefficients of \( s \), i.e., the terms of the Fibonacci sequence.
Verify all these statements and compute the coefficients of the Fibonacci sequence.

3.4 Polynomial Functions
We have argued in the previous sections that one should not define polynomials with coefficients in \( A \) as functions of some sort with domain and values in \( A \). However, there is nothing forbidding us to define functions \( A \to A \) from polynomials in \( A[x] \).
3.4. POLYNOMIAL FUNCTIONS

Definition 3.4.1. If \( p = \sum_{n=0}^{N} p_n x^n \) is a polynomial in \( A[x] \), the function \( p^*: A \rightarrow A \) defined by \( p^*(a) = \sum_{n=0}^{N} p_n a^n \) is called the POLYNOMIAL FUNCTION associated with \( p \).

Example 3.4.2.

Let \( A = \mathbb{Z}_2 \) and \( p = 1 + x + x^2 \). The polynomial function associated to the polynomial \( p \) is \( p^*: \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \) given by \( p^*(a) = 1 + a + a^2 \), for any \( a \in \mathbb{Z}_2 \).

In this case, we have \( p^*(0) = p^*(1) = 1 \), and hence \( p^* \) is a constant function, although \( p \) is not a constant polynomial. In particular, when \( q = 1 \), we have \( p^* \neq q \) and \( p^* = q^* \).

Given a ring \( A \) we will denote by \( A^A \) the set of all functions \( f: A \rightarrow A \).

In Chapter ?? we already observed that \( A^A \) is a ring, with the usual addition and multiplication of functions. We have a map

\[
\Psi: A[x] \rightarrow A^A, \quad p \mapsto p^*,
\]

which associates to each polynomial \( p \in A[x] \) the corresponding function \( p^*: A \rightarrow A \).

Proposition 3.4.3. \( \Psi: A[x] \rightarrow A^A \) is a ring homomorphism.

Proof. Let \( p, q \in A[x] \). We can write \( p = \sum_{n=0}^{N} p_n x^n \) and \( q = \sum_{n=0}^{N} q_n x^n \), by eventually letting some of the coefficients be zero. If \( a \in A \), we need to show that the following identities hold:

\[
(p + q)^*(a) = p^*(a) + q^*(a), \quad \text{and} \quad (pq)^*(a) = p^*(a)q^*(a).
\]

A more general version of the first identity was already proved in Chapter ???. The second identity can also be proved easily by induction.

\[\square\]

Examples 3.4.4.

1. Given a polynomial \( p \in A[x] \), the associated function \( p^* \) is defined not just in the ring \( A \), but also in an extension of \( A \). For example, if \( p = 1 + 2x + 3x^2 \in \mathbb{Z}[x] \), the associated polynomial function \( p^*: \mathbb{Z} \rightarrow \mathbb{Z} \) is of course \( p^*(n) = 1 + 2n + 3n^2 \), for any \( n \in \mathbb{Z} \). Since \( \mathbb{Z} \subset \mathbb{Q} \) we can also consider \( q^*: \mathbb{Q} \rightarrow \mathbb{Q} \), given also by \( q^*(r) = 1 + 2r + 3r^2 \), for any \( r \in \mathbb{Q} \).

\[\text{From now, we will not distinguish between the integer } n \text{ and the class } \overline{n}. \text{ It should be clear from the context if we are referring to an integer (element of } \mathbb{Z} \text{) or an equivalence classe in some } \mathbb{Z}_m.\]
CHAPTER 3. OTHER EXAMPLES OF RINGS

2. More formally, if the ring $B$ is an extension of the ring $A$, so that there exists an injective homomorphism $\phi : A \to B$, then $\phi(A)$ is a subring of $B$ isomorphic to $A$. Hence, if $p = \sum_{n=0}^{N} p_n x^n$ is a polynomial in $A[x]$, then $p$ determines a polynomial function $q^* : B \to B$, namely:

$$q^*(b) = \sum_{n=0}^{N} \phi(p_n)b^n.$$

Of course, $p$ also determines a polynomial $q \in B[x]$ which was denoted $p^\phi$ in the previous section, so in fact we have:

$$q^* = (p^\phi)^*.$$

As we have observed before, it is a matter of common sense to decide to distinguish between $p$ and $q^*$, to write different symbols for the respective functions $p^*$ and $q^*$, or to use different notations for their coefficients $p_n$ and $\phi(p_n)$.

3. Continuing with Example 1, recall that $M_2(\mathbb{Z})$ denotes the ring of $2 \times 2$ matrices with integer entries. The matrices of the form $nI$, where $I$ denotes the identity matrix and $n \in \mathbb{Z}$, is isomorphic to $\mathbb{Z}$: indeed we have an injective homomorphism $\phi : \mathbb{Z} \to M_2(\mathbb{Z})$ given by $\phi(n) = nI$. If $p = 1+2x+3x^2 \in \mathbb{Z}[x]$ we have that $(p^\phi)^* : M_2(\mathbb{Z}) \to M_2(\mathbb{Z})$ is given by $(p^\phi)^*(C) = \phi(1) + \phi(2)C + \phi(3)C^2$, where $C$ denotes now any matrix in $M_2(\mathbb{Z})$. Obviously, we can simplify the notation and write the expression $\phi(1) + \phi(2)C + \phi(3)C^2$ as $I + 2C + 3C^2$.

This is formally the same expression as in Example 1, with the small subtlety that the symbol “1” is replaced by the symbol “$I$”, representing the identity matrix.

In the first example, the identification of $\mathbb{Z}$ with its image in $\mathbb{Q}$ is quite natural. In the last example, one does not need to substitute the coefficients of degree $> 0$ (because the product of a matrix by the integer $n$ is the same as the product by the matrix $nI$), but we must replace the coefficient of degree zero, because the same of matrix with an integer is not a priori defined.

We can use this circle of ideas to formalize the notion of root of a polynomial:

Definition 3.4.5. Let $p \in A[x]$ and let $B$ be an extension of $A$. We say that $b \in B$ is a root of $p$ if $p^*(b) = 0$.

When we write $p^*(a)$ we usually think of $p^*$ as fixed and $a$ as a variable (the “independent variable” of the function $p^*$). However, we can also consider $p^*$ as an “independent variable”. Still assuming that $A$ is any abelian ring with identity and that $B$ is an extension of $A$, we make the following definition (recall the remarks above about identifications!):
Definition 3.4.6. The map \( \text{Val} : A[x] \times B \to B \) is defined by

\[
\text{Val}(p, b) := p^*(b).
\]

\( \text{Val}(p, b) \) is called the evaluation of the polynomial \( p \) at \( b \).

The function \( \text{Val} \) has two independent variables: the polynomial \( p \) and the element \( b \). If we fix the polynomial \( p \), then \( \text{Val} \) is just the function \( p^* : B \to B \) associated to the polynomial \( p \). We can also fix the element \( b \), resulting in an evaluation map \( \psi : A[x] \to B, p \mapsto p^*(b) \).

Examples 3.4.7.

1. Let \( A = \mathbb{Z}, B = \mathbb{C} \) and \( b = i \). Then \( \psi : \mathbb{Z}[x] \to \mathbb{C} \) is given by \( \psi(p) = p^*(i) \). In order to find out the image \( \psi(\mathbb{Z}[x]) \) notice that the division of any polynomial \( p \in \mathbb{Z}[x] \) by \( 1 + x^2 \), results in a reminder of degree less than 2. In other words, \( p = (1 + x^2)q + (a_0 + a_1x) \), where \( q \in \mathbb{Z}[x] \) and \( a_0, a_1 \in \mathbb{Z} \). Hence, \( p^*(i) = a_0 + a_1i \) so \( \psi(\mathbb{Z}[x]) \) consists of all Gaussian integers.

2. Let \( A = \mathbb{Q}, B = \mathbb{R} \) and \( b = \sqrt{2} \). Then \( \psi(p) = p^*(\sqrt{2}) \). Since we can always write \( p = (2-x^2)q + (a_0 + a_1x) \), where \( q \in \mathbb{Q}[x] \) and \( a_0, a_1 \in \mathbb{Q} \), we have that \( p^*(\sqrt{2}) = a_0 + a_1\sqrt{2} \), so we conclude that \( \psi(\mathbb{Q}[x]) \) consists of all real numbers of the form \( a_0 + a_1\sqrt{2} \) with \( a_0, a_1 \in \mathbb{Q} \).

3. The ring \( A[x] \) is always an extension of the ring \( A \). Hence, if we let \( A \) be any ring, \( B = A[x] \) and \( b = x \), it is obvious that \( \psi(p) = p^*(x) = p \). In other words, \( \psi : A[x] \to A[x] \) is the identity function. For this reason, we can denote the polynomial \( p \) by the symbol \( p^*(x) \), which usually simplified to \( p(x) \).

Motivated by the last example, whenever \( B \) is an extension of \( A, b \in B \), and \( \psi : A[x] \to B \) is the evaluation map \( \psi(p) = p^*(b) \), we will denote the image \( \psi(A[x]) \) by the symbol \( A[b] \). This is the reason why we used the symbol \( \mathbb{Z}[i] \) for the Gaussian integers. In the case of Example 3.4.7.2, we have

\[
\mathbb{Q}[\sqrt{2}] = \{ p^*(\sqrt{2}) : p \in \mathbb{Q} \} = \{ a_0 + a_1\sqrt{2} : a_0, a_1 \in \mathbb{Q} \}.
\]

The previous examples suggest that \( A[b] \) is always a subring of \( B \). It is easy to check that this always holds:

**Proposition 3.4.8.** For any fixed \( b \in B \), the evaluation map \( \psi : A[x] \to B, p \mapsto p^*(b) \), is a ring homomorphism. Hence, \( A[b] \) is a subring of \( B \).

**Proof.** Proceed as in the proof of Proposition 3.4.3. \( \square \)
Also, in analogy with Definition 3.3.11 whenever \( A[b] \) is an integral domain, we will denote its field of fractions by \( A(b) \).

Notice that \( A[b] \) always contains the values of the constant polynomials. These form a subring of \( A[b] \) isomorphic to \( A \). Hence, \( A[b] \) is always an extension of \( A \): the ring of Gaussian integers contains a subring isomorphic to \( \mathbb{Z} \), \( \mathbb{Q}[\sqrt{2}] \) contains a subring isomorphic to \( \mathbb{Q} \), etc.). On the other hand, in all Examples 3.4.7, only for the last one is the evaluation map \( \psi \) injective. Therefore, only for the last of these examples, we have \( A[b] \) isomorphic to \( A[x] \). Of course, \( \psi \) is injective if and only if its kernel \( N(\psi) \) consist of the zero polynomial zero. Since:

\[
N(\psi) = \{ p \in A[x] : p^*(b) = 0 \},
\]

we conclude that \( \psi \) is injective if and only if \( b \) is not a root of any non-zero polynomial with coefficients in \( A \).

**Definition 3.4.9.** Let \( B \) be an extension of \( A \) and \( b \in B \). We say that \( b \) is **algebraic over** \( A \) if there exists a non-zero polynomial \( p \in A[x] \) such that \( p^*(b) = 0 \). Otherwise, we say that \( b \) is **transcendental over** \( A \).

**Example 3.4.10.**

As we saw above, \( i \in \mathbb{C} \) is algebraic over \( \mathbb{Z} \), and \( \sqrt{2} \in \mathbb{R} \) is algebraic over \( \mathbb{Q} \) (and also over \( \mathbb{Z} \)). The polynomial \( x \in A[x] \) is always transcendental over \( A \).

In general, an extension \( B \) of \( A \) may (i) contain both algebraic elements and transcendental elements over \( A \), or (ii) may contain only algebraic elements over \( A \). We distinguish these two possibilities:

**Definition 3.4.11.** We call \( B \) an **algebraic extension** of \( A \) if all its elements are algebraic over \( A \). Otherwise, \( B \) is called a **transcendental extension** of \( A \).

**Examples 3.4.12.**

1. The ring of Gaussian integers is an algebraic extension of \( \mathbb{Z} \). In fact, any Gaussian integer \( m + ni \) is a root of the non-zero polynomial with integer coefficients \( (x - m)^2 + n^2 = x^2 - 2mx + m^2 + n^2 \).
2. \( \mathbb{Q}[\sqrt{2}] \) is an algebraic extension of \( \mathbb{Q} \), since \( a + b\sqrt{2} \) is a root of the non-zero polynomial with rational coefficients \( (x - a)^2 - 2b^2 \).
3. \( \mathbb{Q} \) is an algebraic extension of \( \mathbb{Z} \), because \( \frac{m}{n} \) is a root of \( nx - m \in \mathbb{Z}[x] \).
4. \( \mathbb{C} \) is an algebraic extension of \( \mathbb{R} \), because \( a + bi \) is a root of \( (x-a)^2 + b^2 \in \mathbb{R}[x] \).
5. $A[x]$ is a transcendental extension of $A$.

6. We will show in the next chapter that $\mathbb{R}$ is a transcendental extension of $\mathbb{Q}$.

We close this section with the following result which expresses the fact that $A[x]$ is the smallest transcendental extension of $A$. We leave its proof for the exercises.

**Theorem 3.4.13.** Any transcendental extension of $A$ contains a subring isomorphic to $A[x]$.

**Exercises.**

1. Conclude the proof of Proposition 3.4.3.

2. Assume that $\sqrt{n}$ is irrational. Show that $\mathbb{Q}[\sqrt{n}]$ is an algebraic extension of $\mathbb{Q}$, a subfield of $\mathbb{R}$, and a vector space of dimension 2 over $\mathbb{Q}$.

3. Prove Theorem 3.4.13.

4. Assume that $A \subset B$ are integral domains and $b \in B$. Show that:
   
   (a) $A[b]$ is the smallest integral domain containing $A$ and $b$.
   
   (b) $A(b)$ is the smallest field containing $A$ and $b$.

5. Let $K$ be a field and fix $n$ points $(a_k, b_k)$ in $K \times K$ with $a_i \neq a_j$ for $i \neq j$. Show that :
   
   (a) There exists a polynomial $p_i \in K[x]$ such that
       
       $p_i(a_j) = \begin{cases} 
       1, & \text{if } j = i, \\
       0, & \text{if } j \neq i. 
       \end{cases}$
   
   (HINT: Modify the polynomial $q_i = \prod_{k \neq i} \frac{(x-a_k)}{(x-a_i)}$.)

   (b) There exists a polynomial $p(x) \in K[x]$ of degree $\leq n - 1$ such that $p(a_k) = b_k$, for all $k = 1, \ldots, n$.

   The formula defining $p$ is called Lagrange interpolation formula.

6. Let $K$ be a finite field. Show that any function $K \to K$ is polynomial.

---

*The fields $\mathbb{Q}[\sqrt{a}]$ where $a$ is not a perfect square are called QUADRATIC FIELDS, and play an important role in Number Theory.*
7. Let $K$ be a subfield of $L$, and assume that $L$ is a finite dimensional vector space over $K$. Show that $L$ is an algebraic extension of $K$.

(HINT: if the dimension of $L$ over $K$ is $n$ and $a \in L$ then $\{a^k : 0 \leq k \leq n\}$ cannot be a linearly independent set.)

8. Let $K$ be a subfield of $L$ and let $b \in L$ be algebraic over $K$. Use the previous exercise to show that $K(b)$ is an algebraic extension of $K$.

### 3.5 Division of Polynomials

In this and in the next sections we will study in some detail the ring of polynomials $A[x]$. Underlying this study, playing a fundamental role, there will be the division algorithm for polynomials. Therefore, we need to find sufficient conditions on the ring $A$ for this algorithm to hold. Many of the results that we will obtain are then analogous to the results we have obtained in Chapter 2 for the ring of integers.

According to Example 3.4.7.3, we will adopt the following convention: a polynomial $p \in A[x]$ is represented by the symbol $p(x)$ and the value of the polynomial $p$ at $a \in A$ will now be represented by $p(a)$ instead of $p^*(a)$.

**Definition 3.5.1.** A polynomial $p(x) \in A[x]$ is called **monic** if $p_n = 1$, where $\deg p(x) = n$ and 1 is the identity in the ring $A$.

**Example 3.5.2.**

The polynomial $5 + 3x + 2x^2 + x^4 \in \mathbb{Z}[x]$ is monic.

The division algorithm presents obvious issues in rings with zero divisors. On the other hand, if $D$ is a integral domain, then by Proposition 3.3.10(b) $D[x]$ is also an integral domain. In fact, we have:

**Theorem 3.5.3** (Division Algorithm). Let $D$ be an integral domain. If $p(x), d(x) \in D[x]$ and $d(x)$ is monic, there exist unique polynomials $q(x)$ and $r(x)$, with $\deg r(x) < \deg d(x)$, such that $p(x) = q(x)d(x) + r(x)$.

Similarly to the case of the integers, the polynomials $q(x)$ and $r(x)$ are called, respectively, the **quotient** and the **remainder** of the division of $p(x)$ by $d(x)$. The case where $r(x) = 0$ corresponds, of course, to the case where $d(x)$ is a divisor (or factor) of $p(x)$. We recall that in this case we write $d(x)|p(x)$. 


Proof of Theorem 3.5.3. We will show separately the existence and uniqueness.

Existence: Let \( R = \{ p(x) - a(x)d(x) : a(x) \in D[x] \} \). There are two cases, depending if 0 belongs to \( R \) or not:

(a) If \( 0 \in R \) there exists \( a_0(x) \in D[x] \) such that \( p(x) - a_0(x)d(x) = 0 \). In this case we set \( q(x) = a_0(x) \) and \( r(x) = 0 \). It follows that:
\[
p(x) = q(x)d(x) + r(x),
\]
with \( \deg r(x) < \deg d(x) \).

(b) If \( 0 \not\in R \) the set \( G = \{ \deg(p(x) - a(x)d(x)) : a(x) \in D[x] \} \subseteq \mathbb{N}_0 \), formed by the degrees of the polynomials in \( R \) must have a minimum \( m \). Let \( a_0(x) \) be a polynomial with \( \deg(p(x) - a_0(x)d(x)) = m \). If we set \( q(x) = a_0(x) \) and \( r(x) = p(x) - q(x)d(x) \), it follows that:
\[
p(x) = q(x)d(x) + r(x),
\]
and we just need to check that \( m = \deg r(x) < \deg d(x) \). Assume that \( n = \deg d(x) \leq m \). Then we set \( k = m - n \) and \( r'(x) = r(x) - r_m x^k d(x) \). On the one hand, \( r'(x) = p(x) - (q(x) + r_m x^k) d(x) \in R \). On the other hand, since \( d(x) \) is monic, we have \( \deg r'(x) < m = \deg r(x) \), contradicting the definition of \( m \). Therefore, we must have \( m < n \), i.e., \( \deg r(x) < \deg d(x) \).

Uniqueness: Let \( p(x) = q(x)d(x) + r(x) = q'(x)d(x) + r'(x) \), where \( \deg r(x) \) and \( \deg r'(x) \) are both smaller than \( \deg d(x) \). We have that
\[
(q(x) - q'(x)) d(x) = r'(x) - r(x).
\]
If \( q(x) \neq q'(x) \), then
\[
\deg (r'(x) - r(x)) \geq \deg d(x),
\]
which contradicts
\[
\deg (r'(x) - r(x)) \leq \max\{\deg r'(x), \deg r(x)\} < \deg d(x).
\]
Hence we must have \( q(x) = q'(x) \), which also implies \( r(x) = r'(x) \).

The argument above for the existence part of the proof can be implemented as an algorithm as follows: given a polynomial \( p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \in D[x] \) of degree \( n \) denote by \( p^{\top}(x) = a_n x^n \) the highest degree term of \( p(x) \). To divide the polynomial \( p(x) \) by a polynomial \( d(x) \) one iterates as follows:
• Starting with \( q(x) = 0 \) and \( r(x) = p(x) \), one replaces at each step

\[
q(x) \to q(x) + \frac{r_{\text{top}}(x)}{d_{\text{top}}(x)}, \quad r(x) \to r(x) - \frac{r_{\text{top}}(x)}{d_{\text{top}}(x)}d(x).
\]

The iteration ends when \( \deg r(x) < \deg d(x) \).

Example 3.5.4.

Let \( D = \mathbb{Z}_5 \). The division of \( p(x) = 2x^4 + 4x^3 + x^2 + 2x + 3 \) by \( d(x) = x^2 + x + 1 \) results in a quotient \( q(x) = 2x^2 + 2x + 2 \), with remainder \( r(x) = 3x + 1 \).

Assuming that \( a \in D \), the polynomial \( d(x) = (x - a) \) is monic of degree 1. Hence, any polynomial \( p(x) \in D[x] \) can be divided by \( (x - a) \) and, according to the division algorithm, the remainder is a constant polynomial. The next corollary states that the corresponding element of \( D \) is just the value of \( p(x) \) at \( a \). We leave the proof as an exercise:

Corollary 3.5.5 (Remainder Theorem). If \( p(x) \in D[x] \) and \( a \in D \), the remainder of the division of \( p(x) \) by \( (x - a) \) is the constant polynomial \( r(x) = p(a) \). In particular, \((x - a)|p(x)\) if and only if \( a \) is a root of \( p(x) \).

Examples 3.5.6.

1. Consider the polynomial \( p(x) = 2x^4 + 4x^3 + x^2 + 2x + 3 \) in \( \mathbb{Z}_5[x] \). Since \( p(1) = 12 \equiv 2 \pmod{5} \), it follows that the remainder of the division of \( 2x^4 + 4x^3 + x^2 + 2x + 3 \) by \( (x - 1) = (x + 4) \) is \( r(x) = 2 \).

2. If we assume instead that \( p(x) = 2x^4 + 4x^3 + x^2 + 2x + 3 \) is a polynomial with coefficients in \( \mathbb{Z}_3 \), we have that \( p(1) = 12 \equiv 0 \pmod{3} \), so in this case \( (x - 1) = (x + 2) \) is a factor of \( 2x^4 + 4x^3 + x^2 + 2x + 3 \).

Another consequence of the Division Algorithm (or, more precisely, of Corollary 3.5.5) is the classical result about the maximal number of roots of a non-zero polynomial:

Proposition 3.5.7. If \( p(x) \in D[x] \) and \( \deg p(x) = n \geq 0 \), then \( p(x) \) has at most \( n \) roots in \( D \).

Proof. We use induction in the degree of the polynomial \( p(x) \). If \( n = 0 \), the polynomial \( p(x) \) is constant and non-zero. It is then obvious that it has no roots.

Assume now that the result holds for an integer \( n \geq 0 \). Let \( \deg p(x) = n + 1 \) and assume that \( a \) is a root of \( p(x) \) (if \( p(x) \) has no roots there is nothing
3.5. DIVISION OF POLYNOMIALS

...to prove!). According to the Remainder Theorem, \( p(x) = q(x)(x-a) \), where of course \( \deg q(x) = n \). On the one hand, by the induction assumption, \( q(x) \) has at most \( n \) roots. On the other hand, if \( b \in D \) is distinct from \( a \), then \( p(b) = q(b)(b-a) \). Since \( D \) is an integral domain, we can only have \( p(b) = 0 \) if \( q(b) = 0 \). In other words, the remaining roots of \( p(x) \) must also be roots of \( q(x) \). Therefore, \( p(x) \) has at most \( n + 1 \) roots. \( \square \)

When \( D \) is an integral domain, the only invertible polynomials \( q(x) \in D[x] \) are the constant polynomials \( q(x) = a \), with \( a \in D \) invertible. These polynomials can always be used to obtain trivial factorizations of any polynomial \( p(x) \): if \( a(x)b(x) = 1 \) then \( p(x) = a(x)b(x)p(x) \). For this reason, if \( q(x) \mid p(x) \), one often says that \( q(x) \) is a proper factor of \( p(x) \) if and only if \( p(x) = a(x)q(x) \), where neither \( a(x) \) nor \( q(x) \) are invertible. Hence, a factorization is non-trivial if and only if it contains at least one proper factor.

Our next definition, which should be compared with the definition of prime numbers presented in Chapter ??, says that an irreducible polynomial is a non-invertible polynomial with only trivial factorizations. Notice that the restricting to a non-invertible polynomial is entirely analogous to the exclusion of 1 from the set of prime numbers.

Definition 3.5.8. A non-invertible polynomial \( p(x) \in D[x] \) is said to be irreducible in \( D[x] \) when it has no proper factors in \( D[x] \). In other words, if for \( q_1(x), q_2(x) \in D[x] \),

\[
p(x) = q_1(x)q_2(x) \implies q_1(x) \text{ or } q_2(x) \text{ is invertible in } D[x].
\]

We say that \( p(x) \) is reducible in \( D[x] \) if it is not irreducible.

Examples 3.5.9.

1. \( p(x) = x - a \) is always irreducible, no matter what the domain \( D \) is.

2. If \( D = \mathbb{Z} \), then \( p(x) = 2x - 3 \) is irreducible (check this!), but \( q(x) = 2x + 6 \) is reducible, because \( 2x + 6 = 2(x + 3) \), and both 2 and \( x + 3 \) are not invertible in \( \mathbb{Z}[x] \).

3. If \( \deg p(x) \geq 2 \) and \( p(x) \) has at least one root in \( D \), it follows from the Remainder Theorem that \( p(x) \) is necessarily reducible in \( D[x] \).

4. If \( p(x) \) is monic of degree 2 or 3, then \( p(x) \) is reducible in \( D[x] \) if and only if it had at least one root in \( D \) (why?).

5. It is possible for \( p(x) \) to have no roots in \( D \) and be reducible in \( D[x] \): for example, \( x^4 + 2x^2 + 1 \) is reducible in \( \mathbb{Z}[x] \), since \( x^4 + 2x^2 + 1 = (x^2 + 1)^2 \), but it has no roots in \( \mathbb{Z} \).
6. You probably have learned in elementary Algebra that the only irreducible polynomials in \( \mathbb{R}[x] \) are the polynomials of degree 1 and the quadratic polynomials \( p(x) = ax^2 + bx + c \), with negative discriminant \( \Delta = b^2 - 4ac \). We will see later that this is a consequence of the Fundamental Theorem of Algebra.

7. The property of a polynomial being irreducible depends heavily on the domain \( D \). For example, the polynomial \( x^2 + 1 \) is irreducible in \( \mathbb{R}[x] \), but reducible in \( \mathbb{C}[x] \supseteq \mathbb{R}[x] \). On the other hand, the polynomial \( p(x) = 2x + 6 \) is reducible in \( \mathbb{Z}[x] \), but irreducible in \( \mathbb{Q}[x] \supseteq \mathbb{Z}[x] \).

In some cases it is possible to describe all the irreducible polynomials in \( D[x] \), as in the case \( D = \mathbb{R} \). However, there are cases where it is practically impossible to describe them. The next results illustrate how complex this issue is in the case of \( \mathbb{Z}[x] \) and \( \mathbb{Q}[x] \).

Given \( p(x) \in \mathbb{Z}[x] \), \( p(x) = a_0 + a_1x + \cdots + a_nx^n \), we will say that \( c(p) = \gcd(a_0, a_1, \ldots, a_n) \) is content of \( p(x) \). We say that \( p(x) \) is primitive if its coefficients are relatively prime, i.e., if \( c(p) = 1 \). Obviously \( p(x) \) is primitive if and only if it has no factorizations of type \( p(x) = kq(x) \), with \( k \in \mathbb{Z} \), \( k \neq \pm 1 \) and \( q(x) \in \mathbb{Z}[x] \).

**Lemma 3.5.10.** If \( p(x) \in \mathbb{Z}[x] \) and \( p(x) = a(x)b(x) \), with \( a(x), b(x) \in \mathbb{Q}[x] \), then there exist polynomials \( a'(x), b'(x) \in \mathbb{Z}[x] \), and \( k \in \mathbb{Q} \), such that

\[
p(x) = a'(x)b'(x), \quad a(x) = ka'(x) \quad \text{and} \quad b(x) = k^{-1}b'(x).
\]

**Proof.** Obviously, there are integers \( n, m \) such that \( \tilde{a}(x) = na(x) \in \mathbb{Z}[x] \), \( \tilde{b}(x) = mb(x) \in \mathbb{Z}[x] \), and \( nmp(x) = \tilde{a}(x)\tilde{b}(x) \). Let \( q \) be any prime factor of \( nm \). By the generalization of Euclid's Lemma given in Exercise [3.4 Section 3.5](#),

\[
q|\tilde{a}(x)\tilde{b}(x) \implies q|\tilde{a}(x) \text{ or } q|\tilde{b}(x).
\]

Hence, we can divide both sides of the identity \( nmp(x) = \tilde{a}(x)\tilde{b}(x) \) by \( q \), still obtaining on the right hand side a product of two polynomials in \( \mathbb{Z}[x] \). If we repeat this procedure for all prime factors of \( nm \), we arrive at an identity

\[
p(x) = a'(x)b'(x),
\]

where \( a'(x), b'(x) \in \mathbb{Z}[x] \), \( \tilde{a}(x) = sa'(x) \) and \( \tilde{b}(x) = tb'(x) \), with \( s, t \in \mathbb{Z} \). We conclude that \( a(x) = \frac{s}{n}a'(x) \), \( b(x) = \frac{t}{m}b'(x) \), and \( \frac{s}{n} = \frac{t}{m} = 1 \).

**Lemma 3.5.11 (Gauss).** If \( p(x) \in \mathbb{Z}[x] \) is not constant, then \( p(x) \) is irreducible in \( \mathbb{Z}[x] \) if and only if it is primitive and irreducible in \( \mathbb{Q}[x] \).
Proof. Assume first that \( p(x) \) is reducible and primitive in \( \mathbb{Z}[x] \). We claim that \( p(x) \) is reducible in \( \mathbb{Q}[x] \). Indeed, in this case, \( p(x) = a(x)b(x) \), with \( a(x), b(x) \in \mathbb{Z}[x] \). Note that if any of the polynomials \( a(x) \) or \( b(x) \) is constant then it must be invertible, i.e., \( \pm 1 \), since \( p(x) \) is primitive. We conclude that \( a(x) \) and \( b(x) \) are not constant, hence not invertible in \( \mathbb{Q}[x] \). Therefore, the factorization \( p(x) = a(x)b(x) \) is non-trivial in \( \mathbb{Q}[x] \), so \( p(x) \) is reducible in \( \mathbb{Q}[x] \), as claimed.

If \( p(x) \) is not primitive, it is obvious that it is reducible in \( \mathbb{Z}[x] \). Therefore, it remains to show that if \( p(x) \) is reducible in \( \mathbb{Q}[x] \) it is also reducible in \( \mathbb{Z}[x] \). In this case, \( p(x) = a(x)b(x) \), where \( a(x), b(x) \in \mathbb{Q}[x] \) are non constant. According to Lemma 3.5.10 there exist polynomials \( a'(x), b'(x) \in \mathbb{Z}[x] \) such that \( p(x) = a'(x)b'(x) \), and \( a'(x), b'(x) \) are not constant. Hence, \( p(x) \) is reducible in \( \mathbb{Z}[x] \). \( \square \)

The next theorem and the previous lemma allows one to easily obtain examples of irreducible polynomials in \( \mathbb{Z}[x] \) and \( \mathbb{Q}[x] \).

**Theorem 3.5.12** (Eisenstein’s Criterion). Let \( a(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Z}[x] \) be a polynomial of degree \( n \). If there exists a prime \( p \in \mathbb{Z} \) such that \( a_k \equiv 0 \pmod{p} \) for \( 0 \leq k < n \), \( a_n \not\equiv 0 \pmod{p} \) and \( a_0 \not\equiv 0 \pmod{p^2} \) then \( a(x) \) is irreducible in \( \mathbb{Q}[x] \).

**Proof.** Assume that in \( \mathbb{Q}[x] \) we have a factorization

\[
a(x) = b(x)c(x) = (b_0 + b_1 x + \cdots)(c_0 + c_1 x + \cdots).
\]

According to Lemma 3.5.10, we may assume without loss of generality that \( b(x), c(x) \in \mathbb{Z}[x] \). If \( b_0 \equiv c_0 \equiv 0 \pmod{p} \), then \( a_0 = b_0 c_0 \equiv 0 \pmod{p^2} \), which contradicts the assumption that \( a_0 \not\equiv 0 \pmod{p^2} \). Hence we can assume, e.g., that \( c_0 \not\equiv 0 \pmod{p} \).

It is obvious that if \( p | b(x) \) then \( p | a(x) \), which is impossible because \( a_n \not\equiv 0 \pmod{p} \). We conclude that the set \( \{ k \geq 0 : b_k \not\equiv 0 \pmod{p} \} \) is non-empty. Let us denote by \( m \) its minimum.

Notice that

\[
a_m = \sum_{i=0}^{m} b_i c_{m-i} \equiv b_m c_0 \not\equiv 0 \pmod{p},
\]

so we must have \( m = n \), since \( a_n \) is the only coefficient of \( a(x) \) not divisible by \( p \).

Therefore, \( \deg b(x) \geq \deg a(x) \), and since \( a(x) = b(x)c(x) \), we have \( \deg b(x) = \deg a(x) \), and \( c(x) \) must be constant. This shows that \( a(x) \) is irreducible in \( \mathbb{Q}[x] \). \( \square \)
Examples 3.5.13.

1. If \( p \in \mathbb{Z} \) is a prime, \( q(x) = x^n - p \) is irreducible in \( \mathbb{Z}[x] \) and in \( \mathbb{Q}[x] \).

2. Eisenstein's Criterion does not apply to the polynomial \( q(x) = x^6 + x^3 + 1 \). However, noting that
   \[
   q(x + 1) = (x + 1)^6 + (x + 1)^3 + 1 = x^6 + 6x^5 + 15x^4 + 21x^3 + 18x^2 + 9x + 3,
   \]
   Eisenstein's Criterion show that \( q(x + 1) \) is irreducible in \( \mathbb{Q}[x] \). We conclude that \( q(x) \) is irreducible in \( \mathbb{Q}[x] \).
   Since \( q(x) \) is monic, it is also irreducible in \( \mathbb{Z}[x] \).

The examples and remarks above lead to the following conclusions:

- In any domain, the polynomials \( x - a \) are irreducible.
- There are domains for which there exist irreducible polynomials of degree as large as we wish.
- There are domains for which there are irreducible polynomials only up to same degree larger or equal to 1.

It remains to show that there are fields where the only irreducible polynomials take the form \( x - a \). We leave it as an exercise to check that if this is true then the any polynomial of degree \( > 1 \) must have a root. For this reason, we introduce the following:

**Definition 3.5.14.** A field \( K \) is called *algebraically closed* if every non-constant polynomial \( p(x) \in K[x] \) has at least one root in \( K \).

We will not prove the next result, which will only be used in examples and exercises. One can find a proof in any text of Complex Analysis. We leave as an exercise to use it to determine the irreducible polynomials em \( \mathbb{R} \).

**Theorem 3.5.15** (Fundamental Theorem of Algebra). The field \( \mathbb{C} \) of complex numbers is algebraically closed, i.e., any non-constant complex polynomial has at least one complex root.

**Exercises.**

1. If \( p(x) \in D[x] \), \( p(x) \neq 0 \), and \( a \in D \) is a root of \( p(x) \), the largest natural number \( m \) such that \( p(x) \) is a multiple of \( (x - a)^m \) is called the *multiplicity* of the root \( a \). Show that the sum of the multiplicities of the roots of \( p(x) \) is \( \leq \deg p(x) \).
2. Show that \( p(x) \in A[x] \) can have more than \( \deg p(x) \) roots, if \( A \) has zero divisors.

3. Show that \( x^2 + 1 \) is irreducible in \( \mathbb{Z}_3[x] \).

4. Determine all the irreducible polynomials \( p(x) \in \mathbb{Z}_3[x] \) with \( \deg p(x) \leq 2 \).

5. How many irreducible polynomials are there in \( \mathbb{Z}_5[x] \) with degree 5? (Hint: count the reducible and irreducible polynomials in \( \mathbb{Z}_5[x] \) of degrees \( \leq 5 \).)

6. Show that the following statements are equivalent:
   
   (a) The field \( K \) is algebraically closed.
   
   (b) Any polynomial in \( K[x] \) of degree \( \geq 1 \) is a product of polynomials of degree 1.

7. Assume that \( K \) is an algebraically closed field and show that if \( p(x) \in K[x] \) and \( \deg p(x) = n \geq 1 \), then the sum of the multiplicities of the roots of \( p(x) \) is exactly \( n \).

8. Assume that \( p(x) \in \mathbb{R}[x] \) and using the Fundamental Theorem of Algebra show that:
   
   (a) If \( c \in \mathbb{C} \) is a root of \( p(x) \), the complex conjugate of \( c \) is also a root of \( p(x) \).
   
   (b) If \( p(x) \) is irreducible in \( \mathbb{R}[x] \) and \( \deg p(x) > 1 \), then \( p(x) = ax^2 + bx + c \) and \( b^2 - 4ac < 0 \).

9. If \( K \) is an algebraically closed field and \( D \) is an integral domain and an algebraic extension of \( K \), show that \( K = D^{\mathbb{A}} \).

10. If \( p(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Z}[x] \), and \( c(p) = \gcd(a_0, a_1, \ldots, a_n) \), show that:
    
    (a) \( p(x) = c(p)q(x) \) where \( q(x) \) is primitive,
    
    (b) If \( p(x) \) and \( q(x) \) are primitive then \( p(x)q(x) \) is primitive.

---

\(^9\)In particular, according to the Fundamental Theorem of Algebra, there is no field \( L \) which is an algebraic extension of \( \mathbb{C} \). This gives a complete answer to Hamilton’s problem, discussed in Chapter ??.
3.6 The Ideals of $K[x]$

As we saw in Chapter ??, the structure of the ideals in $\mathbb{Z}$ is specially simple, a fact which is behind Euclid’s Algorithm for the computation of the greatest common divisor and least common multiple of two integers. For a general integral domain $D$ the structure of the ideals of $D[x]$ can be very complex, as the example of $\mathbb{Z}[x]$ shows. In general, there is no analogue of the Euclid’s Algorithm. However, if $D = K$ is a field, then the structure of the ideals in $K[x]$ is easy to describe, as we shall see now.

We start by remarking that the Division Algorithm, as stated in Theorem 3.5.3, in the case of $K[x]$ can be strengthen as follows:

**Theorem 3.6.1.** If $p(x), d(x) \in K[x]$ and $d(x) \neq 0$, there exist unique polynomials $q(x)$ and $r(x)$, such that $p(x) = q(x)d(x) + r(x)$ and $\deg r(x) < \deg d(x)$.

This result is similar to the result we have proved in Chapter ?? for the integers, and so can be used to establish various analogies between the rings $K[x]$ and $\mathbb{Z}$. In particular:

**Theorem 3.6.2.** Every ideal in $K[x]$ is principal.

*Proof. Assume that $I \subset K[x]$ is an ideal. If $I = \{0\}$, then $I = (0)$ is a principal ideal.

Therefore, assume that $I \neq \{0\}$, so there exists some non-zero polynomial $p(x) \in I$, and consider the set

$$N = \{n \in \mathbb{N}_0 : \exists p(x) \in I, n = \deg p(x)\} \subset \mathbb{N}_0.$$

The set $N$ is non-empty, so it must have a minimum. Let $m(x) \in I$ be such that $\deg m(x) = \min N$. Then

$$r(x) \in I \text{ and } r(x) \neq 0 \implies \deg m(x) \leq \deg r(x).$$

Since $m(x) \in I$, obviously $\langle m(x) \rangle \subset I$. On the other hand, if $p(x) \in I$ the Division Algorithm gives $p(x) = m(x)d(x) + r(x)$, where $\deg r(x) < \deg m(x)$. Since $I$ is an ideal, it follows that

$$r(x) = p(x) - m(x)d(x) \in I,$$

and we conclude from (3.6.1) that $r(x) = 0$ (otherwise we would have $\deg r(x) < \deg m(x)$, a contradiction). Hence, $p(x) \in \langle m(x) \rangle$ and $I = \langle m(x) \rangle$. 

$\square$
3.6. THE IDEALS OF $K[X]$

In Chapter ?? we have used the fact that all ideals in $\mathbb{Z}$ are principal to construct Euclides Algorithm, which allows one to find the greatest common divisor and the least common multiple of any two integers. We will now apply the same set of ideas to the ring $K[x]$.

We start with the following proposition, whose proof is left for the exercises:

**Proposition 3.6.3.** Let $I = \langle p(x) \rangle$ and $J = \langle q(x) \rangle$ be ideals in $K[x]$. Then:

(a) $I \subset J$ if and only if $q(x) | p(x)$;

(b) $I$ is maximal if and only if $p(x)$ is irreducible;

(c) If $I = J$ and $p(x)$ and $q(x)$ are monic or zero, then $p(x) = q(x)$.

If $p(x), q(x) \in K[x]$, then $I = \langle p(x), q(x) \rangle$ is the ideal in $K[x]$, given by:

$$I = \{ a(x)p(x) + b(x)q(x) : a(x), b(x) \in K[x] \}.$$  

According to Theorem 3.6.2, this ideal is principal. Therefore, there exists a polynomial $d(x) \in K[x]$ such that $\langle d(x) \rangle = \langle p(x), q(x) \rangle$. Note that $d(x)$ satisfies:

- $d(x) | p(x)$ and $d(x) | q(x)$;
- there exist polynomials $a(x)$ and $b(x)$ such that $d(x) = a(x)p(x) + b(x)q(x)$;
- if $c(x) | p(x)$ and $c(x) | q(x)$, then $c(x) | d(x)$.

In other words, $d(x)$ is a *common divisor* of $p(x)$ and $q(x)$, and it is a *multiple* of any other common divisor of these two polynomials.

Similarly, $\langle p(x) \rangle \cap \langle q(x) \rangle$ is a principal ideal, so there exists $m(x) \in K[x]$ such that $\langle m(x) \rangle = \langle p(x) \rangle \cap \langle q(x) \rangle$. In this case, we have:

- $p(x) | m(x)$ and $q(x) | m(x)$;
- if $p(x) | n(x)$ and $q(x) | n(x)$, then $m(x) | n(x)$.

Hence, $m(x)$ is a *common multiple* of $p(x)$ and $q(x)$, and is a *divisor* of any other polynomial which is a common multiple of these two polynomials.

Finally, note that according to Proposition 3.6.3(c), if $p(x)$ and $q(x)$ are monic polynomials or zero and $\langle p(x) \rangle = \langle q(x) \rangle$, then $p(x) = q(x)$. Hence we can introduce:
**Definition 3.6.4.** Let \( p(x), q(x) \in K[x] \).

(i) If \( \langle d(x) \rangle = \langle p(x), q(x) \rangle \), with \( d(x) \) monic or zero, then \( d(x) \) is called the **greatest common divisor** of \( p(x) \) and \( q(x) \), abbreviated to \( d(x) = \gcd(p(x), q(x)) \).

(ii) If \( \langle m(x) \rangle = \langle p(x) \rangle \cap \langle q(x) \rangle \), with \( m(x) \) monic or zero, then \( m(x) \) is called the **least common multiple** of \( p(x) \) and \( q(x) \), abbreviated to \( m(x) = \lcm(p(x), q(x)) \).

Similarly to the integers, notice that

\[
p(x) = q(x)a(x) + r(x) \quad \implies \quad \langle p(x), q(x) \rangle = \langle q(x), r(x) \rangle,
\]

so Euclid's Algorithm remains valid in \( K[x] \). We illustrate it with an example in \( Z_5[x] \).

**Example 3.6.5.**

In order to find the greatest common divisor of \( p(x) = x^4 + x^3 + 2x^2 + x + 1 \) and \( q(x) = x^3 + 3x^2 + x + 3 \) in \( Z_5[x] \), we proceed as follows:

1. Divide \( p(x) \) by \( q(x) \), so one obtains

\[
x^4 + x^3 + 2x^2 + x + 1 = (x^3 + 3x^2 + x + 3)(x + 3) + 2x^2 + 2.
\]

Hence:

\[
\langle x^4 + x^3 + 2x^2 + x + 1, x^3 + 3x^2 + x + 3 \rangle = \langle x^3 + 3x^2 + x + 3, 2x^2 + 2 \rangle.
\]

2. Then divide \( x^3 + 3x^2 + x + 3 \) by \( 2x^2 + 2 \), leading to

\[
x^3 + 3x^2 + x + 3 = (2x^2 + 2)(3x + 4).
\]

Hence:

\[
\langle x^3 + 3x^2 + x + 3, 2x^2 + 2 \rangle = \langle 2x^2 + 2 \rangle = \langle x^2 + 1 \rangle.
\]

3. One concludes that

\[
\gcd(x^4 + x^3 + 2x^2 + x + 1, x^3 + 3x^2 + x + 3) = x^2 + 1.
\]

The exact same argument that we used for the integers, also leads to the following result:

**Lemma 3.6.6.** If \( p(x), q(x) \in K[x] \) are monic polynomials, then

\[
\gcd(p(x), q(x)) \lcm(p(x), q(x)) = p(x)q(x).
\]
Example 3.6.7.

We saw above that the greatest common divisor of \( p(x) = x^4 + x^3 + 2x^2 + x + 1 \)
and \( q(x) = x^3 + 3x^2 + x + 3 \) em \( \mathbb{Z}_5[x] \) is \( d(x) = x^2 + 1 \). Hence, we conclude
that the least common multiple of these polynomials is the polynomial
\[
m(x) = \frac{(x^4 + x^3 + 2x^2 + x + 1)(x^3 + 3x^2 + x + 3)}{x^2 + 1} = x^5 + 4x^4 + 2x^2 + 4x + 3.
\]

Exercises.

1. Prove Theorem 3.6.1

2. Prove Proposition 3.6.3

3. Let \( p(x), q(x) \in K[x] \). Show that
\[
I = \langle p(x), q(x) \rangle = \{a(x)p(x) + b(x)q(x) : a(x), b(x) \in K[x]\}.
\]

4. Let \( p(x), q(x) \in K[x] \). Show that if \( \langle d(x) \rangle = \langle p(x), q(x) \rangle \), then:
   
   (a) There exist \( a(x), b(x) \in K[x] \) such that \( d(x) = a(x)p(x) + b(x)q(x) \).
   
   (b) \( d(x) \mid p(x) \) and \( d(x) \mid q(x) \).
   
   (c) If \( c(x) \mid p(x) \) and \( c(x) \mid q(x) \), then \( c(x) \mid d(x) \), hence \( \deg c(x) \leq \deg d(x) \).

5. Show that the following generalization of Euclid’s Lemma holds: if \( p(x), q_1(x), q_2(x) \in K[x] \), \( p(x) \) is irreducible and \( p(x) \mid q_1(x)q_2(x) \), then \( p(x) \mid q_1(x) \) or \( p(x) \mid q_2(x) \).

6. If \( p(x), q(x) \in K[x] \), shows that
\[
p(x) = q(x)a(x) + r(x) \implies \langle p(x), q(x) \rangle = \langle q(x), r(x) \rangle.
\]

7. Let \( d(x) \) be the greatest common divisor of \( x^4 + x^3 + 2x^2 + x + 1 \) and \( x^3 + 3x^2 + x + 3 \) in \( \mathbb{Z}_5[x] \). Determine \( a(x) \) and \( b(x) \) in \( \mathbb{Z}_5[x] \) such that
\[
d(x) = a(x)(x^4 + x^3 + 2x^2 + x + 1) + b(x)(x^3 + 3x^2 + x + 3).
\]

8. Let \( p(x), q(x) \in K[x] \), \( d(x) = \gcd(p(x), q(x)) \) and \( m(x) = \text{lcm}(p(x), q(x)) \).
Show that:
   
   (a) If \( p(x) \mid r(x) \) and \( q(x) \mid r(x) \), then \( p(x)q(x) \mid r(x)d(x) \).
   
   (b) There exists \( k \in K \) such that \( kd(x)m(x) = p(x)q(x) \).
9. Let \( q(x) \in K[x] \) be non-zero and non-invertible. Prove the following statements (recall the Fundamental Theorem of Arithmetic):

(a) There exist irreducible polynomials \( p(x) \) such that \( p(x) | q(x) \).
(b) There exist irreducible monic polynomials \( m_1(x), \ldots, m_k(x) \in K[x] \) and \( a_0 \in K \) such that \( q(x) = a_0 \prod_{i=1}^{k} m_i(x) \).
(c) The decomposition in (b) is unique up to the order of the factors.

10. Show that the ideal \( \langle x, y \rangle \) in \( K[x, y] \) is not principal.

11. Assume that the ring \( A \) is an extension of the field \( K \), and let \( a \in A \) be algebraic over \( K \). If \( J = \{ p(x) \in K[x] : p(a) = 0 \} \), show that:

(a) \( J = \langle m(x) \rangle \) is an ideal of \( K[x] \).
(b) If \( A \) does not have zero divisors, then \( m(x) \) is irreducible, and \( K[a] = K(a) \) is a field.
(c) If \( A \) does not have zero divisors, and \( B \) is the set of all elements in \( A \) which are algebraic over \( K \), then \( B \) is a field and it is the largest algebraic extension of \( K \) in \( A \).

12. Show that \( \mathbb{Q}[\sqrt{2}] \) and \( \mathbb{Q}[\sqrt{-2}] \) are algebraic extensions of \( \mathbb{Q} \) and subfields of \( \mathbb{R} \). What are their dimensions, when viewed as vector spaces over \( \mathbb{Q} \)?

13. Let \( A \subset \mathbb{R} \) be the set of all real numbers which are algebraic over \( \mathbb{Q} \). Show that:

(a) \( A \) is a countable field.
(b) \( A \) is an algebraic extension of \( \mathbb{Q} \).
(c) \( A \) has infinite dimension, when viewed as a vector space over \( \mathbb{Q} \).

3.7 Divisibility and Prime Factorization

We saw before in our studies of the ring of integers \( \mathbb{Z} \) and of the ring of polynomials \( K[x] \) that in these rings any non-zero element which is non-invertible has an essential unique factorization has a product irreducible or prime elements. It is natural to wonder if this fact holds in other rings. We will study in this and in the next section how the notions of divisibility and factorization can be extended to any integral domain \( D \).

\footnote{One say that \( m(x) \) is the \textit{minimal polynomial} of the element \( a \).}
Recall that if $a, b \in D$ we say that $a$ divides (or is a factor of) $b$ if there exists $d \in D$ such that $b = da$. In this case we write “$a \mid b$”. The following notions are partially inspired by the usual concepts we have seen in $\mathbb{Z}$ and $K[x]$, and they play a basic role.

**Definition 3.7.1.** Let $D$ be an integral domain, $a, b, p \in D$, and $p$ a non-invertible element. We say that:

(i) $a$ is associate to $b$, if $a \mid b$ and $b \mid a$;

(ii) $p$ is prime, if $p \neq 0$ and $p|ab \Rightarrow p|a$ or $p|b$;

(iii) $p$ is irreducible, if $p = ab \Rightarrow a$ is invertible or $b$ is invertible.

Notice that in this more general context, Euclides Lemma becomes a definition of a prime element, and the irreducible elements are those elements that only admit trivial factorizations. We will see below that in the rings $\mathbb{Z}$ and $K[x]$ the prime elements in the sense of the above definition are exactly the irreducible elements. It is only for historical reasons that we use the terms “prime” in $\mathbb{Z}$ and “irreducible” in $K[x]$. For a general integral domain, the notion of prime and irreducible may not coincide, but we will identify many classes of domains where these notions are equivalent, and where one can establish appropriate generalizations of the Fundamental Theorem of Arithmetics, concerning the existence and uniqueness of factorizations into irreducible elements.

The binary relation associate to is actually an equivalence relation: it is easy to check that $a$ is associate to $b$ if and only if $a = ub$ for some invertible element $u$. Hence, if $a, b \in D$ are associates, we write $a \sim b$. In Factorization Theory, it is common to call the invertible elements units, and we will follow this practice. Notice that the units are precisely those elements which are associate to the identity of $D$, and that given $p, q \in D$, if $p \sim q$, then $p$ is a prime (respectively irreducible) if and only if $q$ is a prime (respectively, irreducible). In particular, if $p$ is a prime then all the elements obtained from $p$ by multiplying by units are also prime elements (and this is *not* the usual convention in $\mathbb{Z}$).

We start by observing the following general fact:

**Lemma 3.7.2.** In any integral domain, a prime element is always irreducible.

\footnote{We also use the symbol “$a \nmid b$” to say that $a$ does not divide $b$.}
Proof. If \( p \in D \) is prime and \( p = ab \), then either \( p|a \) or \( p|b \). If, for example, \( p|a \), then there exists \( x \in D \) such that \( a = px \), and we conclude that

\[
p = ab \implies p = pxb, \text{ and since } p \neq 0, \\
\implies 1 = xb, \\
\implies b \text{ is invertible.}
\]

Similarly, if \( p|b \), we conclude that \( a \) is invertible.

\[
\square
\]

Examples 3.7.3.

1. In \( \mathbb{Z} \) the units are \( \{1, -1\} \). An element \( p \in \mathbb{Z} \) is irreducible in the sense of Definition 3.7.1 if and only if the natural number \( |p| \) is prime in the sense of Chapter ???. Also, if \( p|n \) then \( |p||n \), so we see that Euclides Lemma shows that if \( p \) is irreducible then

\[
p|ab \implies |p||ab \implies |p||a \text{ or } |p||b \implies p|a \text{ or } p|b.
\]

We conclude that in \( \mathbb{Z} \) irreducible and prime elements, in the sense of Definition 3.7.1 coincide.

2. The units in \( K[\mathbf{x}] \) are the zero degree polynomials, while an irreducible polynomial in the sense of Definition 3.7.1 is just an irreducible polynomial in the usual sense. The generalization of Euclides Lemma given in Exercise ?? shows that if \( p(\mathbf{x}) \in K[\mathbf{x}] \) is irreducible then it is prime. We conclude again that in \( K[\mathbf{x}] \) irreducible and prime elements, in the sense of Definition 3.7.1 coincide.

3. The units in the ring of Gaussian integers \( \mathbb{Z}[i] \) are \( \{1, -1, i, -i\} \). The element \( 2 \in \mathbb{Z}[i] \) is not irreducible in \( \mathbb{Z}[i] \) (although it is irreducible in \( \mathbb{Z} \)) since we have:

\[
2 = (1 + i)(1 - i), \text{ with } 1 \pm i \text{ non invertible.}
\]

We claim that \( 1 + i \) and \( 1 - i \) are irreducible. In order to show this, we observe that the function \( N : \mathbb{Z}[i] \to \mathbb{N}_0 \) defined by

\[
N(a + bi) = |a + bi|^2 = a^2 + b^2.
\]

satisfy the following two properties:

(a) if \( z_1, z_2 \in \mathbb{Z}[i] \), then \( N(z_1 z_2) = N(z_1)N(z_2) \);

(b) \( N(z) = 1 \) if and only if \( z \) is invertible.

To check, for example, that \( 1 + i \) is irreducible, assume that \( 1 + i = z_1 z_2 \) in \( \mathbb{Z}[i] \). By property (a), we find that

\[
2 = N(1 + i) = N(z_1 z_2) = N(z_1)N(z_2).
\]
Since $2$ is irreducible in $\mathbb{Z}$, we must have $N(z_1) = 1$ or $N(z_2) = 1$. By property (b), we conclude that $z_1$ or $z_2$ are invertible, and $1 + i$ is irreducible in $\mathbb{Z}[i]$. Observe that $-(1 + i) = -1 - i$, $i(1 + i) = -1 + i$ and $-i(1 + i)$ are also irreducible. We will show later that in $\mathbb{Z}[i]$ any irreducible element is also a prime, and how one can determine all the irreducible in $\mathbb{Z}[i]$.

4. There are integral domains where there are irreducible elements which are not prime. We have already observed in Chapter ?? that in the ring of even integers $2\mathbb{Z}$ we have two distinct factorizations

$$36 = 6 \times 6 = 2 \times 18,$$

so $2$ divides $6 \times 6$ but it does not divide $6$ (note however, that we assume integral domains to have a unit, which is not the case with $2\mathbb{Z}$). Consider instead the ring $\mathbb{Z}[\sqrt{-5}]$, which is an integral domain. In this ring, we have:

$$9 = 3 \cdot 3 = (2 + \sqrt{-5})(2 - \sqrt{-5}) \implies 3|(2 + \sqrt{-5})(2 - \sqrt{-5}).$$

We leave it to the exercises to check that $3$ is irreducible but is not a factor neither of $(2 + \sqrt{-5})$ nor of $(2 - \sqrt{-5})$. Hence $3 \in \mathbb{Z}[\sqrt{-5}]$ is irreducible but not a prime.

Using these elementary notions we can introduce domains where factorizations exist:

**Definition 3.7.4.** A domain $D$ is called a **unique factorization domain** (abbreviated UFD) if: satisfies:

(i) If $0 \neq d \in D$ is non-invertible there exists irreducible elements $p_1, \ldots, p_n$ such that

$$d = \prod_{i=1}^{n} p_i,$$

(ii) If $p_1, \ldots, p_n$, and $p'_1 \cdots p'_m$ are irreducible, and $\prod_{i=1}^{n} p_i = \prod_{i=1}^{m} p'_i$, then $n = m$, and there exists a permutation $\pi \in S_n$ such that $p'_i \sim p_{\pi(i)}$.

In other words, in a UFD every non-zero and non-invertible element has a factorization into a product of irreducible elements and this factorization is unique up to the order of the factors and multiplication of each factor by an appropriate unit (note that if $p'_i = u_ip_{\pi(i)}$, then we must have $\prod_{i=1}^{n} u_i = 1$).

Note that the factorization $[3.7.1]$ can also be expressed in *powers* of de irreducible elements, but in this case we may need to include a unit in the factorization:

$$d = u \cdot p_1^{e_1} \cdots p_n^{e_n}.$$
CHAPTER 3. OTHER EXAMPLES OF RINGS

Examples 3.7.5.

1. The ring $\mathbb{Z}$ is a UFD: it follows from the Fundamental Theorem of Arithmetic that every integer can be factored in the form

$$m = p_1 \cdots p_m,$$

where each $p_i$ is irreducible (i.e., $|p_i|$ is a prime). This factorization is unique up to the order of the factor and multiplication by $\pm 1$. For example, we have:

$$-15 = (3) \cdot (-3) = (-1)3^2,$$

where 3 and $-3$ are both irreducible.

2. By Exercise 9 in Section 3.6, given a polynomial $q(x) \in K[x]$, there exist irreducible polynomials $p_1(x), \ldots, p_n(x) \in K[x]$ such that

$$q(x) = \prod_{i=1}^{n} p_i(x).$$

This factorization is unique up to the order of the factor and multiplication by units. Hence, $K[x]$ is a UFD. For example, in $\mathbb{Q}[x]$ we have

$$2x^2 + 4x + 2 = (2x + 2)(x + 1) = 2(x + 1)^2,$$

where $2x + 2$ and $x + 1$ are both irreducible.

3. We will see below that the ring $\mathbb{Z}[i]$ of Gaussian integers is a UFD.

It is useful to observe that all the elementary properties concerning factors and divisibility introduced above can be expressed in terms of ideals. For that, we will say that an ideal $0 \subseteq P \subseteq D$ is a prime ideal if, for any ideals $I, J \subseteq D$,

$$IJ \subseteq P \implies I \subseteq P \text{ or } J \subseteq P.$$

Proposition 3.7.6. Let $a, b, p, u \in D$. Then:

(i) $a|b$ if and only if $\langle a \rangle \supseteq \langle b \rangle$;

(ii) $a \sim b$ if and only if $\langle a \rangle = \langle b \rangle$;

(iii) $u$ is a unit if and only if $\langle u \rangle = D$;

(iv) $p$ is a prime if and only if $\langle p \rangle$ is a prime ideal;

(v) $p$ is irreducible if and only if $\langle p \rangle$ is maximal among all principal ideals in $D$. 
Proof. We leave the proof of (i), (ii) and (iii) as an easy exercise.

In order to show that (iv) holds, let \( p \in D \) be a prime, and \( I, J \subset D \) ideals such that \( IJ \subset \langle p \rangle \). If \( I \not\subset \langle p \rangle \), then there exists \( a \in I \) such that \( a \not\in \langle p \rangle \), i.e., such that \( p \nmid a \) (by (i)). Hence, for all \( b \in J \), we have \( ab \in \langle p \rangle \iff p|ab \) and \( p \nmid a \). Since \( p \) is a prime, we must have \( p|b \), i.e., \( b \in \langle p \rangle \) (by (i)). We conclude that \( J \subset \langle p \rangle \). This proves that \( \langle p \rangle \) is a prime ideal.

For the converse, assume that \( \langle p \rangle \) is a prime ideal. Then

\[
\langle ab \rangle = \langle a \rangle \langle b \rangle \subset \langle p \rangle \implies \langle a \rangle \subset \langle p \rangle \text{ or } \langle b \rangle \subset \langle p \rangle.
\]

By (i), this means that either \( p|a \) or \( p|b \), so \( p \) is a prime.

In order to show that (v) holds, consider first an irreducible element \( p \in D \) and let \( \langle p \rangle \subset \langle a \rangle \), for some \( a \in D \). Then \( p = ax \) hence either \( a \) is a unit or \( x \) is a unit. If \( a \) is a unit, then by (iii) \( \langle a \rangle = D \). If \( x \) is a unit, then \( p \sim a \) and, by (ii), \( \langle p \rangle = \langle a \rangle \). Hence, \( \langle p \rangle \) is maximal among all principal ideals of \( D \).

Conversely, assume that \( \langle p \rangle \) is maximal among all principal ideals of \( D \). If \( p = ab \), since \( \langle p \rangle \subset \langle a \rangle \), we see that either \( \langle a \rangle = D \) and \( a \) is invertible (by (iii)), or \( \langle p \rangle = \langle a \rangle \) and \( a \sim p \) (by (ii)). In the later case, there exists a unit \( u \in D \) such that \( a = pu \), so that

\[
p = ab \implies p = pub, \quad \implies 1 = ub \implies b \text{ is invertible}.
\]

Therefore, either \( a \) or \( b \) is invertible, which shows that \( p \) is irreducible.

The previous proposition suggests one maybe able to construct a Factorization Theory based in the ideals of \( D \) rather than the elements of \( D \). This is indeed true and useful, and the name “ideal” derives historically from this idea (see Chapter 7). For now, we consider only factorization of elements of \( D \), but we use ideals to give a characterization of a UFD.

Given a domain \( D \) we will say that it satisfies the ascending chain condition on principal ideals if every chain of principal ideals:

\[
\langle d_1 \rangle \subset \langle d_2 \rangle \subset \cdots \subset \langle d_n \rangle \subset \cdots
\]

stabilizes, i.e., after some natural number \( n_0 \in \mathbb{N} \) we have

\[
\langle d_{n_0} \rangle = \langle d_{n_0+1} \rangle = \cdots.
\]

This property turns out to be crucial for factorizations to exist:
Theorem 3.7.7. Let $D$ be an integral domain. Then $D$ is a UFD if and only if the following two conditions hold:

(i) every irreducible element is prime;

(ii) $D$ satisfies the ascending chain condition on principal ideals.

Proof. Let $D$ be a UFD and $p \in D$ an irreducible element. If $p | ab$, then there exist $x \in D$ such that $px = ab$, where $x$, $a$ and $b$ admit factorizations of the type

$$x = p_1 \cdots p_r, \ a = p'_1 \cdots p'_s, \ b = p''_1 \cdots p''_t$$

with $p_i, p'_j, p''_k$ irreducible in $D$. Hence:

$$p \cdot p_1 \cdots p_r = p'_1 \cdots p'_s \cdot p''_1 \cdots p''_t,$$

and, by uniqueness of the factorization, we must have $p \sim p'_1$ or $p \sim p''_1$. In the first case, $p | a$ and in the second case $p | b$. Hence, $p$ is a prime.

On the other hand, consider an ascending chain condition of principal ideals

$$\langle d_1 \rangle \subset \langle d_2 \rangle \subset \cdots \subset \langle d_n \rangle \subset \cdots$$

We can assume, without loss of generality (why?), that $d_1 \neq 0$ and $d_i$ is non-invertible, $\forall i$. Since $d_i | d_1$, for all $i$, the factorizations of $d_1$ and $d_i$ take the form

$$d_1 = p_1 \cdots p_r, \ d_i = p'_1 \cdots p'_s.$$ 

The irreducible factors of $d_i$ are factors of $d_1$, so $s \leq r$. In particular, the chain cannot contain more than $r$ distinct ideals, so there exists a natural $n_0$ such that $\langle d_{n_0} \rangle = \langle d_k \rangle$, for all $k \geq n_0$. This completes half of the proof.

Conversely, let $D$ be an integral domain where conditions (i) and (ii) in the statement hold. Let $d \in D$ be a non-zero and non-invertible element. Let us assume that $d$ does not admit a factorization, and see that this leads to a contradiction. By induction, we construct as follows a sequence $\{d_n\}_{n \in \mathbb{N}}$ where $d_1 = d$, $d_{n+1} | d_n$, $d_n \not\sim d_{n+1}$ and none of the elements $d_n$ admits a factorization into a product of irreducible elements.

The case $n = 1$ is trivial, so assume that $n > 0$, and we have already constructed elements $d_1, \ldots, d_n$. Since $d_n$ cannot be irreducible, we have $d_n = a_n b_n$, where $a_n$ and $b_n$ are non-invertible. Obviously, if both $a_n$ and $b_n$ can be factored into a product of irreducible elements, then $d_n$ can also be factored. So assume, without loss of generality that $b_n$ cannot be factored.
3.7. DIVISIBILITY AND PRIME FACTORIZATION

Then we set $d_{n+1} = b_n$, and observe that obviously $d_{n+1} \mid d_n$, and $d_n \not\sim d_{n+1}$.

This shows that the sequence $\{d_n\}_{n \in \mathbb{N}}$ exists.

Now, the principal ideals generated by the $d_n$’s satisfy

$$\langle d_1 \rangle \subsetneq \langle d_2 \rangle \subsetneq \cdots \subsetneq \langle d_n \rangle \subsetneq \cdots$$

but this contradicts the ascending chain condition. Hence, we conclude that every non-zero and non-invertible element can be factored into irreducible elements.

In order to verify the uniqueness of factorization, assume that

$$p_1 \cdots p_n = p_1' \cdots p_m',$$

where, say, $n \leq m$. Since the $p_i$, $p_i'$ are irreducible, by (i) they are also prime. Since $p_n | p_1' \cdots p_m'$, we have that $p_n$ is associate to some $p_{\pi(n)}'$, which we denote by $p_{\pi(n)}'$. Removing these two elements from the products, and repeating the argument, we obtain by exhaustion that $n = m$ and $p_i \sim p_{\pi(i)}'$ for some permutation $\pi \in S_n$.

This theorem justifies the use of the term “prime factorization” to denote factorizations of the type (3.7.1) or (3.7.2).

As we have already observed, the crucial property of the rings $\mathbb{Z}$ and $K[x]$, in what concerns factorization, is that all its ideals are principal.

**Definition 3.7.8.** An integral domain $D$ is called a principal ideal domain (abbreviated PID) if all its ideals are principal, i.e., of the form $\langle d \rangle$.

As a consequence of Theorem 3.7.7 we have:

**Corollary 3.7.9.** Every PID is a UFD.

*Proof.* First we show that the irreducible elements in $D$ are prime. For that, assume that $p \in D$ is irreducible and that $p \mid ab$. The ideal $\langle a, p \rangle$ is principal, so there exists $d \in D$ such that $\langle d \rangle = \langle a, p \rangle$. Since $\langle p \rangle \subset \langle d \rangle \subset D$ and $\langle p \rangle$ is maximal (Proposition 3.7.6), we have

$$\langle p \rangle = \langle d \rangle \text{ or } \langle d \rangle = D.$$ 

In the first case, $p \sim d$ and since $d \mid a$, we conclude that $p \mid a$. In the second case there exist $r, s \in D$ such that $1 = ra + sp$, hence

$$b = 1 \cdot b = (ra + sp)b = rab + srb.$$
Since $p$ divides each of the terms in the right hand side, we conclude that $p|b$. In any case, $p|a$ or $p|b$, so $p$ is a prime.

Last we verify the ascending chain condition on principal ideals. Consider the chain:

$$\langle d_1 \rangle \subset \langle d_2 \rangle \subset \cdots \subset \langle d_n \rangle \subset \ldots$$

Note that $\bigcup_{i=1}^{\infty} \langle d_i \rangle$ is an ideal that, by assumption, must be principal: $\bigcup_{i=1}^{\infty} \langle d_i \rangle = \langle d_0 \rangle$. Obviously, there exists a natural number $n_0$ such that $d_0 \in \langle d_{n_0} \rangle$, and it follows that:

$$\langle d_{n_0} \rangle = \langle d_{n_0+1} \rangle = \ldots$$

We have verified both conditions in Theorem 3.7.7, so $D$ is a UFD.

**Examples 3.7.10.**

1. The ring of Gaussian integers is a UFD, since we know that $\mathbb{Z}[i]$ is a PID, by Exercise ?? in Section 2.6.

2. We will see in the next section that if $D$ is a UFD then $D[x]$ is also a UFD. In particular, $\mathbb{Z}[x]$ is a UFD, although it is not a PID.

In general, the problem of determining if a given integral domain is a UFD maybe hard to solve. For example, it is known that the quadratic domains $\mathbb{Z}[\sqrt{m}]$, for $m < 0$, are UFD if and only if $m = -1, -2, -3, -7$ and $-11$, but this is a very non-trivial result. For $m > 0$ is it not known which domains $\mathbb{Z}[\sqrt{m}]$ are UFD. Also, even if one knows that $D$ is a UFD, it maybe very hard to determine its prime elements. We will illustrate this problem in the case of the Gaussian integers.

In order to simplify the exposition, we will call *Euclidean primes* the prime natural numbers in $\mathbb{Z}$, and *Gaussian primes* the prime Gaussian integers in $\mathbb{Z}[i]$.

**Proposition 3.7.11.** Let $p \in \mathbb{Z}$ be an Euclidean prime.

(i) If $p = n^2 + m^2$ has solutions $n, m \in \mathbb{Z}$, then $z = n + mi$ is a Gaussian prime;

(ii) If $p = n^2 + m^2$ has no solutions in $\mathbb{Z}$, then $p$ is a Gaussian prime;

In particular, $z \in \mathbb{Z}[i]$ is a Gaussian prime if and only if $z \sim p$, where $p \in \mathbb{Z}$ is a Gaussian prime, or $z \sim n + mi$, where $n^2 + m^2 = p$ is an Euclidean prime.
3.7. DIVISIBILITY AND PRIME FACTORIZATION

Proof. Let \( N(n + mi) = n^2 + m^2 \) denote the norm of the Gaussian integer \( n + mi \). To show that (i) holds, assume that \( p = n^2 + m^2 \) is an Euclidean prime, and consider a Gaussian factorization:

\[
 n + mi = zw.
\]

Then we have:

\[
 p = N(n + im) = N(z)N(w).
\]

Hence, either \( N(z) = 1 \) and \( z \) is invertible, or \( N(w) = 1 \) and \( w \) is invertible. Hence, we conclude that \( n + mi \) is a Gaussian prime.

Next, to show that (ii) holds, observe that if \( p \) is not a Gaussian prime then there exist non-invertible Gaussian integers \( z \) and \( w \) such that \( p = zw \). Then \( p^2 = N(p) = N(z)N(w) \), and since \( N(z) \) and \( N(w) \) are integers \( \neq \pm 1 \), one must have \( p = N(z) = N(w) \). Hence the equation \( p = n^2 + m^2 \) has solutions.

Finally, let \( z = a + bi \) be a Gaussian prime, so \( N(z) > 1 \). If \( p \) is any prime factor (in \( \mathbb{Z} \)) of \( N(z) \), there exists \( w \in \mathbb{Z}[i] \) such that

\[
 N(z) = (a + bi)(a - bi) = pw.
\]

Since \( p \in \mathbb{Z} \), we have that \( p|a + bi = z \) if and only if \( p|a - bi \). Hence, one of the following alternative must hold:

(a) \( p \) is also a Gaussian prime: Then either \( p|a + bi \) or \( p|a - bi \), i.e., \( p \) is a factor of \( z \). Hence, \( z \sim p \), since \( z \) and \( p \) are both Gaussian primes;

(b) \( p \) is not a Gaussian prime: We conclude from (ii) that \( p = n^2 + m^2 \) has solutions, and we then have

\[
 N(z) = (a + bi)(a - bi) = pw = (n + mi)(n - mi)w.
\]

From (i), \( n + mi \) is a Gaussian prime, and we conclude that either \( a + bi \sim n + mi \) or \( a + bi \sim n - mi \).

\[\square\]

Examples 3.7.12.

1. Obviously, the equation \( 3 = n^2 + m^2 \) has no solutions in \( \mathbb{Z} \), so \( 3 \) is both an Euclidean prime and a Gaussian prime.

2. Since \( 5 = 1^2 + 2^2 \), it follows that \( 5 \) is not a Gaussian prime, but the Gaussian integers \( \pm 1 \pm 2i \) and \( \pm 2 \pm i \) are all Gaussian primes.
As we have just seen, the identification of Gaussian primes relies on the solutions of $p = n^2 + m^2$, where $p$ is an Euclidean prime. Fermat found a very elegant result concerning the values of $p$ for which this equation has solutions:

**Theorem 3.7.13** (Fermat). Let $p$ be an Euclidean prime. Then the following statements are equivalent:

(i) The equation $p = n^2 + m^2$ has solutions in $\mathbb{Z}$,

(ii) $p \not\equiv 3 \pmod{4}$,

(iii) The equation $x^2 = -1$ has solutions in $\mathbb{Z}_p$.

**Proof.** We leave it as exercise to check that “(i) $\implies$ (ii)”.

To show that “(ii) $\implies$ (iii)”, notice that we can assume $p \neq 2$, since $x = 1$ is a solution of $x^2 = -1 = 1$. Let $2 \neq p \equiv 1 \pmod{4}$. If $x \in \mathbb{Z}_p$ set $C(x) = \{x, -x, x^{-1}, -x^{-1}\}$ and define:

$$x \approx y \iff C(x) = C(y).$$

One checks easily that “$\approx$” is an equivalence relation in $\mathbb{Z}_p^*$, and that $x \neq -x$ for any $x \in \mathbb{Z}_p^*$ (because $p \neq 2$). The equivalence class of $x$ is just the set $C(x)$. Denoting by $\#(C(x))$ the number of elements of the class $C(x)$, notice that the possible values of $\#(C(x))$ are:

$$\#(C(x)) = 2,$$

if $x = x^{-1}$, i.e., if $x = \pm 1$,

$= 2,$ if $x = -x^{-1}$, i.e., if $x$ and $x^{-1}$ are the solutions of $x^2 = -1$,

$= 4,$ if $x$ is not a solution of $x^2 = \pm 1$ (or, equivalently, of $x^4 = 1$).

The equivalence classes of $\approx$ give a partition of $\mathbb{Z}_p^*$, which has $p - 1$ elements. We have just saw that there exists at least one class with 2 elements, namely $C(1) = \{1, -1\}$. Possibly, there exists another class with 2 elements, formed by the roots of $x^2 + 1$, if this polynomial has roots in $\mathbb{Z}_p^*$. If $n$ is the number of equivalence classes with 4 elements, we conclude that the $p - 1$ elements of $\mathbb{Z}_p^*$ are assembled as follows:

- if there are no solutions of $x^2 = -1$, then $p - 1 = 2 + 4n \iff p = 4n + 3$, since there exists only one equivalence class with 2 elements, and the remaining $n$ classes have 4 elements each, or

- there exist solutions of $x^2 = -1$ and then $p - 1 = 2 + 2 + 4n \iff p = 4(n + 1) + 1$, because there are 2 equivalence classes each with 2 elements, and the remaining $n$ classes have 4 elements.
3.7. DIVISIBILITY AND PRIME FACTORIZATION

Since \( p \neq 2 \) is a prime, we have that \( p \not\equiv 3 \pmod{4} \implies p \equiv 1 \pmod{4} \) and we conclude that the equation \( x^2 = -1 \) has solutions in \( \mathbb{Z}_p^* \).

Finally, to show that “(iii) \implies (i)” we note that \( x^2 = -1 \) has solutions in \( \mathbb{Z}_p \) if and only if there exists an integer \( k \) such that \( p | 1 + k^2 = (1 + ki)(1 - ki) \).
If \( p \) is a Gaussian prime then \( p | 1 + ki \) or \( p | 1 - ki \), a contradiction, since \( p / 1 \). Hence \( p \) is not a Gaussian prime, and according to Proposition [3.7.11] the equation \( p = n^2 + m^2 \) must have solutions.

**Examples 3.7.14.**

1. The Euclidean primes 7, 11 and 19 are also Gaussian primes.

2. 1973 is an Euclidean prime which is not a Gaussian prime because \( p \equiv 1 \pmod{4} \). It follows that the equation 1973 = \( n^2 + m^2 \) has solutions \( n, m \in \mathbb{Z} \), which is not an immediate fact (check that \( n = 23 \) and \( m = 38 \) is a solution).

An important property of a UFD is that any two elements have a greatest common divisor and a least common multiple. In order to discuss this, we need a more abstract definition of greatest common divisor and least common multiple, valid for any integral domain.

**Definition 3.7.15.** Let \( D \) be an integral domain, and \( a_1, \ldots, a_n \in D \).

(i) \( d \in D \) is called a **GREATEST COMMON DIVISOR** of \( a_1, \ldots, a_n \) if \( d | a_i \) (\( i = 1, \ldots, n \)) and for all \( b \in D \) such that \( b | a_i \) (\( i = 1, \ldots, n \)) we have that \( b | d \);

(ii) \( m \in D \) is called a **LEAST COMMON MULTIPLE** of \( a_1, \ldots, a_n \) if \( a_i | m \) (\( i = 1, \ldots, n \)) and for all \( b \in D \) such that \( a_i | b \) (\( i = 1, \ldots, n \)) we have that \( m | b \).

Note the lack of **uniqueness** in these notions: if \( d \) is a greatest common divisor, and \( c \sim d \), then \( c \) is also a greatest common divisor, and the same holds for a least common multiple. In \( \mathbb{Z} \) and in \( K[x] \), we could ensure uniqueness by requiring that \( d \) and \( m \) be non negative in \( \mathbb{Z} \), and monic in \( K[x] \). Apart from this detail, the definitions above are compatible with the definitions introduced in Chapters 2 and 3.

It is not at all obvious that given elements \( a_1, \ldots, a_n \in D \) a greatest common divisor and/or a least common multiple exist. However, if \( a_1, \ldots, a_n \) have at least one greatest common divisor (respectively, least common multiple) we will denote by \( \text{gcd}(a_1, \ldots, a_n) \) (respectively, \( \text{lcm}(a_1, \ldots, a_n) \)) any such element. We leave the proof of the following elementary properties as an exercise:
Lemma 3.7.16. Let $a, b, c \in D$. Then:

(i) $\gcd(a, \gcd(b, c)) \sim \gcd(\gcd(a, b), c) \sim \gcd(a, b, c)$;

(ii) $\gcd(ca, cb) \sim c \gcd(a, b)$.

Similarly, with $\gcd$ replaced by $\text{lcm}$.

When $D$ is a UFD the $\gcd(a_1, \ldots, a_n)$ and $\text{lcm}(a_1, \ldots, a_n)$ always exist:

Proposition 3.7.17. Let $D$ be a UFD, and $a, b \in D$. Then:

(i) $\gcd(a, b), \text{lcm}(a, b)$ exist and $ab \sim \gcd(a, b) \text{lcm}(a, b)$;

(ii) If $D$ is a PID, then $\gcd(a, b) = ra + sb$ for some $r, s \in D$.

Proof. (i) If $a = 0$ then $\gcd(a, b) = b$ and $\text{lcm}(a, b) = 0$. If $a$ is invertible, then $\gcd(a, b) = a$ and $\text{lcm}(a, b) = b$. Assume then that both $a$ and $b$ are non-zero and non-invertible. The prime factorizations of $a$ and $b$ can be written as

$$a = u \cdot p_1^{n_{a_1}} \cdots p_s^{n_{a_s}}, \quad b = u' \cdot p_1^{n_{b_1}} \cdots p_s^{n_{b_s}},$$

where the $p_i$ are non-associate, $n_{a_i} \geq 0$ and $n_{b_i} \geq 0$. Setting for $i = 1, \ldots, s$ $m_i = \min\{n_{a_i}, n_{b_i}\}$ and $M_i = \max\{n_{a_i}, n_{b_i}\}$, we see immediately that we can choose:

$$\gcd(a, b) = p_1^{m_1} \cdots p_s^{m_s}, \quad \text{lcm}(a, b) = p_1^{M_1} \cdots p_s^{M_s}.$$

(ii) Let $D$ be a PID. If $a, b \in D$, by (i) there exists $\gcd(a, b)$. We leave as an exercise to check that since $D$ is a PID one has: $\langle a, b \rangle = \langle \gcd(a, b) \rangle$. This means that there exist $r, s \in D$ such that $\gcd(a, b) = ra + sb$.

Exercises.

1. Verify items (i)-(iii) in Proposition 3.7.6.

2. In the ring $\mathbb{Z}[\sqrt{-5}]$ show that the elements 3 and $2 \pm \sqrt{-5}$ are irreducible.

3. Show that the ring $\mathbb{Z}[\sqrt{10}]$ is not a UFD.

4. If $p$ is a Euclidean prime and there are integers $n$ and $m$ such that $p = n^2 + m^2$, show that $p \not\equiv 3 \pmod{4}$.

5. Show that if $p$ is a Euclidean prime and $n, m, a, b \in \mathbb{Z}$ satisfy $p = n^2 + m^2 = a^2 + b^2$ then $\{n^2, m^2\} = \{a^2, b^2\}$.
6. An integral domain $D$ is called an Euclidean domain if there exists a map $\delta : D \to \mathbb{N}$ with the following property: $\forall a, b \in D - \{0\}$ there exist $q, r \in D$ such that

$$a = qb + r, \text{ with } r = 0 \text{ or } \delta(r) < \delta(b).$$

Such a map is called an Euclidean function. Show that:

(a) $\mathbb{Z}$ and $K[x]$ are Euclidean domains;

(b) The ring of Gaussian integers $\mathbb{Z}[i]$ is an Euclidean domain;

(c) A Euclidean domain $D$ always admits an Euclidean function $\tilde{\delta} : D \to \mathbb{N}$ satisfying $\tilde{\delta}(a) \leq \tilde{\delta}(ab)$ for all $a, b \in D - \{0\}$ and $\tilde{\delta}(u) = \tilde{\delta}(1)$ if and only if $u$ is a unit;

HINT: Define a new Euclidean function $\tilde{\delta}(a) = \min\{\delta(xa) : x \in D - \{0\}\}$.

(d) A Euclidean domain is a UFD (without using neither Theorem 3.7.7 or its Corollary);

(e) An Euclidean domain is a PID.

One can show that the algebraic integer $^{[12]}$ in the field $\mathbb{Q}(\sqrt{-19})$ form a PID which is not an Euclidean domain, so the following inclusions are strict:

Euclidean Domains $\subseteq$ PID $\subseteq$ UFD $\subseteq$ Integral Domains

7. Let $D$ be an integral domain.

(a) Show that if $D$ satisfies the ascending chain condition on principal ideals, then every non-zero and non-invertible element of $D$ has a factorization into irreducible elements (possibly, non unique).

(b) Give an example of an integral domain $D$ which does not satisfy the ascending chain condition on principal ideals.

8. Prove Lemma 3.7.16

9. Show that if $D$ is a PID then for any $a, b \in D$ one has $\langle a, b \rangle = \langle \gcd(a, b) \rangle$, $\langle a \rangle \cap \langle b \rangle = \langle \lcm(a, b) \rangle$ and $\gcd(a, b) \lcm(a, b) \sim ab$.

10. Assume that $k \in \mathbb{N}$ and show that $k = n^2 + m^2$ has solutions in $\mathbb{Z}$ if and only if any prime factor $p$ of $k$ with $p \equiv 3 \pmod{4}$ satisfies $p^{2N} | k$. What is the smallest natural number $k$ for which the equation $k = n^2 + m^2 = s^2 + t^2$ has solutions $n, m, s, t$ such that $\{n^2, m^2\} \neq \{s^2, t^2\}$?

$^{[12]}$If $K$ is a field extension of $\mathbb{Q}$, $d \in K$ is called an algebraic integer if $d$ is a root of some monic polynomial $p(x) \in \mathbb{Z}[x]$. 
11. Let \( P \) be the set of Euclidean primes and \( G \) the set of Gaussian primes. Show that both \( P - G \) and \( G - P \) are infinite sets. In other words, show that there are infinite Euclidean primes of the form \( p = 4n + 1 \) and of the form \( p = 4n + 3 \).

**Hint:** Consider the natural numbers of the form \( k = (2N!) - 1 \) and of the form \( k = (\prod_{i=1}^{N}(2i - 1))^2 + 4 \).

### 3.8 Factorization in \( D[x] \)

When \( K \) is a field, the ring of polynomials \( K[x] \) is a UFD, since it is a principal ideal domain. For an arbitrary domain \( D \) the structure of the ideals of \( D[x] \) can be very complex. In fact, \( D[x] \) is a PID if and only if \( D \) is a field, and this explains why some example of integral domains that we have mentioned before, like \( \mathbb{Z}[x] \) or \( K[x, y] = K[y][x] \), are not PIDs (see Exercise 10 in Section 3.6). In any case, it is obvious that if \( D \) is not a UFD then \( D[x] \) is also not a UFD. In fact, we will see in this section that the ring \( D[x] \) is a UFD if and only if \( D \) is a UFD, and this shows in particular that \( \mathbb{Z}[x] \) and \( K[x, y] \) are UFDs.

In this section we will always assume that \( D \) is a UFD, so greatest common divisors and least common multiples always exist in \( D \). Also, we will denote by \( K \) the field of fractions \( \text{Frac}(D) \). The definition of content of a polynomial, introduced in Section 3.5 for polynomials \( p(x) \in \mathbb{Z}[x] \), can be extended without changes to \( D[x] \).

**Definition 3.8.1.** If \( p(x) = a_0 + a_1 x + \cdots + a_n x^n \in D[x] \), we say that \( c(p) \in D \) is the CONTENT of \( p(x) \) if and only if

\[
(3.8.1) \quad c(p) = \gcd(a_0, \ldots, a_n).
\]

It should be clear that, just like the greatest common divisor, the content of a polynomial is only defined up to a multiple by a unit. Again we will call a polynomial \( p(x) \in D[x] \) PRIMITIVE if \( c(p) \sim 1 \).

**Lemma 3.8.2.** Let \( p(x) \in D[x] \). Then:

(i) There exists \( q(x) \in D[x] \) primitive such that \( p(x) = c(p)q(x) \).

(ii) If \( p(x) = dq(x) \), with \( q(x) \in D[x] \) primitive and \( d \in D \), then \( d \sim c(p) \).

**Proof.** Part (i) is obvious.
3.8. FACTORIZATION IN $D[X]$

In order to show that (ii) holds, let $p(x) = a_0 + a_1x + \cdots + a_nx^n$ and $q(x) = b_0 + b_1x + \cdots + b_nx^n$, with $q(x)$ a primitive polynomial, and assume that $p(x) = dq(x)$. Then $a_i = db_i$ and, by Lemma 3.7.16, we conclude that

$$c(p) = \gcd(a_0, \ldots, a_n) \sim d \gcd(b_0, \ldots, b_n) \sim d.$$  

The next two results are useful to express a polynomial $p(x) \in K[x]$ in terms of a primitive polynomial $q(x) \in D[x]$.

**Lemma 3.8.3.** Let $0 \neq p(x) \in K[x]$. Then:

(i) There exist $q(x) \in D[x]$ primitive and $k \in K$ such that $p(x) = kq(x)$;

(ii) If $p(x) = kq(x) = \tilde{k}q(x)$, with $k, \tilde{k} \in K$ and $q(x)$, $\tilde{q}(x) \in D[x]$ primitives, then $\tilde{q}(x) = uq(x)$ and $k = u^{-1}\tilde{k}$, where $u \in D$ is a unit.

**Proof.** (i) If

$$p(x) = a_0 + a_1x + \cdots + a_nx^n = \frac{a_0}{b_0} + \frac{a_1}{b_1}x + \cdots + \frac{a_n}{b_n}x^n \in K[x],$$

we let $b = \prod_{i=1}^{n} b_i$. Clearly, $r(x) = bp(x) \in D[x]$. If $c = c(r)$, by Lemma 3.8.2 there exists $q(x) \in D[x]$ primitive such that $r(x) = cq(x)$ and we have

$$p(x) = kq(x), \text{ where } k = \frac{c}{b} \in K.$$  

(ii) Exercise. □

**Corollary 3.8.4.** If $p(x), q(x) \in D[x]$ are primitives and $p(x) \sim q(x)$ in $K[x]$, then $p(x) \sim q(x)$ in $D[x]$.

**Proof.** If $p(x) \sim q(x)$ in $K[x]$, then $p(x) = \alpha q(x)$, with $\alpha \in K$. The corollary then follows from Lemma 3.8.3 item (ii). □

The next two results are just generalizations to UFD of results that we have obtained before in the cases $D = \mathbb{Z}$ and $K = \mathbb{Q}$.

**Lemma 3.8.5.** Let $p(x), q(x), r(x) \in D[x]$ be such that $p(x) = q(x)r(x)$. Then:

(i) If $d \in D$ is prime, then $d | c(p) \iff d | c(q)$ or $d | c(r)$, and

(ii) $p(x)$ is primitive if and only if $q(x)$ and $r(x)$ are both primitive.
Proof. Part (ii) follow from (i).

To prove (i), note first that if \( d \) is a prime and \( d \mid c(q) \) or \( d \mid c(r) \), obviously \( d \mid c(p) \). So it remains to prove the converse. Write \( p(x) = a_0 + a_1 x + \cdots \), \( q(x) = b_0 + b_1 x + \cdots \), and \( r(x) = c_0 + c_1 x + \cdots \). Let \( d \in D \) be a prime such that \( d \mid c(p) \) and assume, by contradiction, that \( d \nmid c(q) \) and \( d \nmid c(r) \). Let:

\[
s := \min \{ k \geq 0 : d \nmid b_k \}, \quad \text{and} \quad t := \min \{ k \geq 0 : d \nmid c_k \}.\]

If \( m = s + t \), then we observe that:

\[
a_m = \sum_{k=0}^{m} b_k c_{m-k} = \sum_{k=0}^{s-1} b_k c_{m-k} + b_s c_t + \sum_{k=s+1}^{m} b_k c_{m-k},
\]

\[
= \sum_{k=0}^{s-1} b_k c_{m-k} + b_s c_t + \sum_{k=0}^{t-1} b_{m-k} c_k,
\]

where the last sums are assume to represent zero, when \( s = 0 \) or \( t = 0 \). It follows that:

\[
b_s c_t = a_m - \sum_{k=0}^{s-1} b_k c_{m-k} - \sum_{k=0}^{t-1} b_{m-k} c_k.
\]

The right hand side of the last identity is a multiple of \( d \), while the left hand side cannot be a multiple of \( d \). Hence, we must have either \( d \mid c(q) \) or \( d \mid c(r) \). \( \square \)

**Example 3.8.6.**

The polynomials \( p(x) = 3x^2 + 2x + 5 \) and \( q(x) = 5x^2 + 2x + 3 \) in \( \mathbb{Z}[x] \) are both primitive since \( \gcd(3, 2, 5) = 1 \). Its product is the primitive polynomial \( p(x)q(x) = 15x^4 + 16x^3 + 38x^2 + 16x + 15 \).

We now generalize Lemma 3.5.10 from \( D = \mathbb{Z} \) to any UFD: the factors \( a(x) \in K[x] \) of a polynomial \( p(x) \in D[x] \) are associate in \( K[x] \) of the factors of \( p(x) \) in the original ring \( D[x] \).

**Lemma 3.8.7.** If \( p(x) \in D[x] \), and \( p(x) = a(x)b(x) \) with \( a(x), b(x) \in K[x] \), then there exist \( \tilde{a}(x), \tilde{b}(x) \in D[x] \), and \( k \in K \), such that

\[
p(x) = \tilde{a}(x)\tilde{b}(x), \quad a(x) = k\tilde{a}(x), \quad \text{and} \quad b(x) = k^{-1}\tilde{b}(x).
\]

**Proof.** Lemma 3.8.3 (i) shows that \( a(x) = sa'(x) \) and \( b(x) = tb'(x) \), where \( s, t \in K \), and \( a'(x) \) and \( b'(x) \) are primitive polynomials in \( D[x] \). On the
other hand, we have from Lemma 3.8.2 that \( p(x) = c(p)p'(x) \), where \( p'(x) \) is also primitive in \( D[x] \). We conclude that \( p(x) = c(p)p'(x) = sta'(x)b'(x) \).

By Lemma 3.8.3 (ii), \( a'(x)b'(x) \) is primitive and it follows from Lemma 3.8.3 (i) that \( c(p)u = st \) and \( p'(x) = u^{-1}a'(x)b'(x) \), for some unit \( u \in D \).

Let us set, for example, \( \tilde{a}(x) = c(p)ua'(x) \) and \( \tilde{b}(x) = b'(x) \). Then

\[
\tilde{a}(x)\tilde{b}(x) = c(p)ua'(x)b'(x) = sta'(x)b'(x) = p(x).
\]

It follows also that the constant in the statement can be taken to be \( k = \frac{c(p)u}{a} \) and \( k^{-1} = \frac{1}{x} \).

In this more general context Gauss’s Lemma takes the same form as in the case \( D = \mathbb{Z} \).

**Corollary 3.8.8** (Gauss’s Lemma). *If \( p(x) \in D[x] \) is not constant, then \( p(x) \) is irreducible in \( D[x] \) if and only if it is primitive in \( D[x] \) and irreducible in \( K[x] \).*

**Proof.** If \( p(x) \) is not primitive then it has a non-trivial factorization in \( D[x] \), of the form \( p(x) = c(p)p'(x) \). On the other hand, if \( p(x) = a(x)b(x) \) is a non-trivial factorization in \( K[x] \), then \( p(x) \) has a non-trivial factorization in \( D[x] \), by Lemma 3.8.7.

If \( p(x) \) is reducible in \( D[x] \), then it has a non-trivial factorization \( p(x) = a(x)b(x) \) in \( D[x] \). If any of these non-trivial factors is constant, then \( p(x) \) is not primitive. If all factors are non-constant, we obtain a non-trivial factorization in \( K[x] \).

Finally, we can prove

**Theorem 3.8.9.** \( D[x] \) is a UFD if and only if \( D \) is a UFD.

**Proof.** Assume that \( D[x] \) is a UFD. Since any \( 0 \neq d \in D \) can be identified with a zero degree polynomial, it is obvious that \( D \) must be a UFD.

Conversely, let \( D \) be a UFD. We will show that any \( q(x) \in D[x] \) of degree \( > 0 \) has a unique factorization.

**Existence:** We have that \( q(x) = c(q)p(x) \), where \( p(x) \) primitive. Since \( K[x] \) is a UFD, \( q(x) \) has a factorization \( q(x) = \prod_{k=1}^{n} q_k(x) \), where the polynomials \( q_k(x) \in K[x] \) are irreducible in \( K[x] \) and \( \deg(q_k(x)) \geq 1 \). As we saw above, there exist primitive polynomials \( p_k(x) \in D[x] \) and constant \( s_k \in K \) such that \( q_k(x) = s_k p_k(x) \). By Gauss’s Lemma, the polynomials \( p_k(x) \) are irreducible in \( D[x] \), and we have

\[
q(x) = c(q)p(x) = s \prod_{k=1}^{n} p_k(x), \text{ where } s = \prod_{k=1}^{n} s_k.
\]
On the other hand, according to Lemma 3.8.5 (ii), \( \prod_{k=1}^{n} p_k(x) \) is primitive. By Lemma 3.8.3 (ii), there exists a unit \( u \in D \) such that \( s = c(p)u \) and, in particular, \( s \in D \). If we factor \( s = \prod_{k=1}^{m} c_k \) into irreducible elements \( c_k \in D \) we have that

\[
q(x) = \left( \prod_{k=1}^{m} c_k \right) \left( \prod_{k=1}^{n} p_k(x) \right)
\]

is a factorization of \( q(x) \) into irreducible elements in \( D[x] \).

**Uniqueness:** Assume that \( q(x) = \prod_{k=1}^{m'} c_k' \prod_{k=1}^{n'} p'_k(x) \) is another factorization of \( q(x) \) into irreducible polynomials in \( D[x] \), where \( c_k' \in D \) and \( \deg(p'_k(x)) \geq 1 \). By Gauss’s Lemma the polynomials \( p'_k(x) \) are primitive and irreducible in \( K[x] \). Hence:

- \( \prod_{k=1}^{n'} p'_k(x) \) is primitive, so \( \prod_{k=1}^{m'} c_k' \sim c(q) \sim \prod_{k=1}^{m} c_k \) em \( D \). Since \( D \) is a UFD, we have \( m = m' \) and, after an appropriate permutation of these factorizations, we have \( c_k \sim c_k' \) in \( D \).

- \( \prod_{k=1}^{n} p_k(x) \sim \prod_{k=1}^{n'} p'_k(x) \) in \( K[x] \). Since \( K[x] \) is a UFD, we have that \( n' = n \) and also after an appropriate permutation of these factorizations \( p_k(x) \sim p'_k(x) \) in \( K[x] \). By Corollary 3.8.4 we also have that \( p_k(x) \sim p'_k(x) \) in \( D[x] \).

**Exercises.**

1. If \( p(x) \in \mathbb{Z}[x] \) is a monic polynomial with integers coefficients show that any rational root \( p(x) \) is actually an integer.

2. Let \( D \) be an integral domain which has an non-invertible element \( d \neq 0 \). Show that \( D[x] \) is not a PID.

3. Prove the following generalization of Eisenstein’s Criterion: let \( D \) be a UFD, \( K = \text{Frac}(D) \) and \( p(x) = a_0 + a_1 x + \cdots + a_n x^n \in D[x] \) with \( n \geq 1 \). If \( p \in D \) is a prime such that \( p | a_k \) for \( 0 \leq k < n, p \nmid a_n \) and \( p^2 \nmid a_0 \), then \( p(x) \) is irreducible in \( K[x] \).

4. Show that the polynomial \( p(x, y) = x^3 + x^2 y + x y^2 + y \) is irreducible in \( K[x, y] \).

5. Show that the polynomial \( p(x) = i x^3 + 4x^2 + 16ix + (3 - i) \) is irreducible in \( \mathbb{Z}[i][x] \).
6. Show that when $D$ is a UFD and $p(x) \in D[x]$ is monic, then any monic factor of $p(x)$ in $K[x]$ belongs to $D[x]$.

7. Let $D$ be a UFD and $p(x), q(x) \in D[x]$.
   (a) Can one use Euclid’s Algorithm to compute $\text{gcd}(p(x), q(x))$ in $D[x]$?
   (b) Does the equation $\text{gcd}(p(x), q(x)) = a(x)p(x) + b(x)q(x)$ always have solutions $a(x), b(x) \in D[x]$?

8. Let $D$ be an UFD. Which of the following rings are UFD?
   (a) The ring $D[[x]]$ of power series with coefficients in $D$.
   (b) The ring $D[1/x, x]$ of fractions $p(x)/q(x)$, with $p(x), q(x) \in D[x]$ and $q(x) \neq 0$.
   (c) The ring $D[1/x, x] \ [[x]]$ of Laurent series.
   (d) The ring $D[x_1, x_2, \ldots]$ of polynomials in an infinite number of indeterminates.
Chapter 4

Quotients and Isomorphisms

4.1 Groups and Equivalence Relations

In defining the ring operations in \( \mathbb{Z}_m \) one takes the following steps:

(i) First, one defines congruence modulo \( m \) (\( x \equiv y \pmod{m} \iff y - x \in \langle m \rangle \)), and shows that it is an equivalence relation.

(ii) Then, one considers the set \( \mathbb{Z}_m \) consisting of the equivalence classes \( x = \{ x + z : z \in \langle m \rangle \} \), which one often writes as \( x = x + \langle m \rangle \).

(iii) Finally, one defines operations on these equivalence classes, from the usual operations on the integers, through the identities \( x + y = x + y \) and \( xy = xy \).

We shall now see that this procedure can be extended to much more general contexts, and hence can be used to provide many new examples of algebraic structures.

We start by replacing the additive group \((\mathbb{Z}, +)\) by an arbitrary group \((G, \cdot)\). We use the multiplicative notation since the fact that addition was commutative plays here no role. We also replace the subgroup \( \langle m \rangle \subset (\mathbb{Z}, +) \) by an arbitrary subgroup \( H \subset G \), and we generalize the congruent relations modulo \( m \) as follows:

**Definition 4.1.1.** If \((G, \cdot)\) is a group and \( H \subset G \) is a subgroup, the congruence relation modulo \( H \) is given by:

\[
g_1 \equiv_g g_2 \pmod{H} \iff g_2^{-1} \cdot g_1 \in H.
\]
CHAPTER 4. QUOTIENTS AND ISOMORPHISMS

Note that, in fact, congruence modulo \( m \) is the special case of this definition where \( G = (\mathbb{Z}, +) \) and \( H = \langle m \rangle \). We leave it as an exercise to check that congruence modulo \( H \) is indeed an equivalence relation.

**Proposition 4.1.2.** If \((G, \cdot)\) is a group and \( H \subset G \) is a subgroup, then \( \equiv \) (mod \( H \)) is an equivalence relation in \( G \).

Proceeding as we did for the case of the integers, we observe that if \( g \in G \), the equivalence class of \( g \), denoted by \( \overline{g} \), can be written as:

\[
\overline{g} = \{ g' \in G : g' \equiv g \},
\]

\[
= \{ g' \in G : g^{-1}g' = h \in H \},
\]

\[
= \{ g' \in G : g' = gh, h \in H \}.
\]

If \( A \) and \( B \) are any subsets of the group \( G \), we will denote by \( AB \) the set of all products of elements of \( A \) and elements of \( B \):

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

If \( A = \{a\} \) (respectively, \( B = \{b\} \)) is a singleton, we will write \( aB \) (respectively, \( Ab \)) instead of \( AB \). We leave it as an exercise to check that \( A(BC) = (AB)C \) and that, in general, \( AB \neq BA \).

With these conventions, we will denote by \( gH \) the equivalence class of \( g \in G \) for the congruence (mod \( H \)), and call it the **left coset** of \( H \). The set of all equivalence classes \( \{ gH : g \in G \} \) is called the **quotient set** of \( G \) by \( H \), and denoted \( G/H \). Therefore, we have that:

\[ G/H = \{ gH : g \in G \}. \]

The number of elements of \( G/H \), which is just the number of left cosets, is called the **index of the subgroup** \( H \) in \( G \), and is denoted by \([G : H]\).

**Examples 4.1.3.**

1. Consider the symmetric group \( G = S_3 = \{ I, \alpha, \beta, \gamma, \delta, \varepsilon \} \) and take as a subgroup the alternating group \( H = A_3 = \{ I, \delta, \varepsilon \} \). Note that:
   - The equivalence class of \( I \) is the coset \( \overline{I} = IH = H = \{ I, \delta, \varepsilon \} \). We conclude that \( I \equiv \delta \equiv \varepsilon \), and \( \overline{I} = \overline{\delta} = \overline{\varepsilon} \). Also, \( H = \delta H = \varepsilon H \).
   - If we let \( g = \alpha \), it is immediate that \( \overline{\alpha} = \alpha H = \{ \alpha I, \alpha \delta, \alpha \varepsilon \} \) and a simple computation shows that \( \overline{\alpha} = \{ \alpha, \beta, \gamma \} \). It follows that \( \alpha \equiv \beta \equiv \gamma \), or \( \overline{\alpha} = \overline{\beta} = \overline{\gamma} = \{ \alpha, \beta, \gamma \} = \alpha H = \beta H = \gamma H \).

\footnote{If \( G \) is an additive group with operation \(+\), it is convenient to write \( A + B \) instead of \( AB \) and \( g + H \) instead of \( gH \). In this case, we have \( A + B = B + A \) and \( g + H = H + g \).}
We conclude that in this example there are 2 distinct cosets each with 3 elements. Hence, the quotient set is \( S_3/A_3 = \{I, \alpha\} = \{A_3, \alpha A_3\} \), and the index of \( A_3 \) in \( S_3 \) is \([S_3 : A_3] = 2\).

2. Consider again the group \( S_3 \), but let us fix now the subgroup \( H = \{I, \alpha\} \).

Then:

- The equivalence class of \( I \) is the coset \( I = H \), hence \( I \equiv \alpha \) and \( I = \alpha \);
- The equivalence class of \( \beta \) is the coset \( \beta = \beta H = \{\beta I, \beta \alpha\} = \{\beta, \varepsilon\} \), hence \( \beta \equiv \varepsilon \) and \( \beta = \varepsilon \);
- The equivalence class of \( \gamma \) is the coset \( \gamma = \gamma H = \{\gamma I, \gamma \alpha\} = \{\gamma, \delta\} \), hence \( \gamma \equiv \delta \) and \( \gamma = \delta \).

We conclude that there are now 3 distinct cosets, each with 2 elements. The quotient set is \( S_3/H = \{I, \beta, \gamma\} = \{H, \beta H, \gamma H\} \) while the index is \([S_3 : H] = 3\).

3. Obviously the index of the subgroup \( \langle m \rangle \) in \( \mathbb{Z} \) is the number of elements of \( \mathbb{Z}_m \), i.e., \([\mathbb{Z} : \langle m \rangle] = m\).

Instead of the binary relation in Definition 4.1.1, we can consider instead the binary relation defined by

\[ g_1 \equiv g_2 \pmod{H} \iff g_1g_2^{-1} \in H. \]

We still obtain an equivalence relation (distinct from the previous one, if the group multiplication is not commutative), and the equivalence class of \( g \in G \) is now given by

\[ \overline{g} = \{g' \in G : g' \equiv g\} = \{g' \in G : g'g^{-1} = h \in H\} = \{g' \in G : g' = hg, h \in H\}. \]

We denote this equivalence class by \( Hg \), and call it a right coset of \( H \). The set of right cosets is denoted by \( H \backslash G \). One may notice that the left and right cosets of a subgroup maybe equal, i.e., \( Hg = gH \), for every \( g \in G \), as in Example 4.1.3.1, or distinct, as in Example 4.1.3.2. We leave this easy check for the exercises.

If \( \equiv \) is an equivalence relation in a set \( X \), we know that the corresponding equivalence classes give a partition of \( X \). In other words, the equivalence classes are distinct, disjoint, subsets of \( X \), and their unions is the set \( X \).

---

\(^2\)By default, unless otherwise stated, we will always use left cosets. When it is clear which cosets (left or right) one is using, we may also use the notation \( G/H \).
Of course, if $X$ is a finite set, each equivalence class is also a finite set, and there are only a finite number of distinct equivalence classes. Denoting by $X_1, X_2, \ldots, X_n$ the equivalence classes of the equivalence relation $\equiv$ in the set $X$, we have:

\[(4.1.1) \quad |X| = |X_1| + \cdots + |X_n| = \sum_{i=1}^{n} |X_i|.
\]

This identity is sometimes referred to as the class equation.

When $X = G$ and one considers the equivalence relation (mod $H$), for some subgroup $H \subset G$, the number of elements in each equivalence class is always the same:

**Proposition 4.1.4.** If $H$ is a finite subgroup of $G$, then

$$|gH| = |Hg| = |H|,$$

for all $g \in G$.

**Proof.** Fix an element $g \in G$. The map $\phi : H \to gH$ defined by $\phi(h) = g \cdot h$ is obviously surjective. On the other hand, by the cancelation law, it is clear that $\phi$ is also 1:1:

$$\phi(h) = \phi(h') \Rightarrow g \cdot h = g \cdot h' \Rightarrow h = h'.$$

Therefore, $\phi$ is a bijection between $H$ and $gH$, so that $|H| = |gH|$. Similarly, one shows that $|Hg| = |H|$. \qed

The previous proposition, combined with the class equation \[(4.1.1),\] yields:

**Theorem 4.1.5 (Lagrange).** If $G$ is a finite group and $H \subset G$ is a subgroup, then:


In particular, both $|H|$ and $[G : H]$ are factors of $|G|$.

**Proof.** Since both $G$ and $H$ are finite, there exists only a finite number of left cosets so that $[G : H] = n$. Denote by $g_1H, \ldots, g_nH$ these left cosets. The class equation \[(4.1.1)\] gives

$$|G| = \sum_{i=1}^{n} |g_iH| = \sum_{i=1}^{n} |H| = n|H| = [G : H]|H|. \qed$$
4.1. GROUPS AND EQUIVALENCE RELATIONS

The number of elements of a group \( G \) is usually called the order of the group \( G \). Hence, by Lagrange’s Theorem, the order of a finite group \( G \) is a multiple of the order of any of its subgroups. Analogously, if \( g \in G \) is any element, then the order of the element \( g \) is the order of the subgroup generated by the element \( g \), i.e., the order of \( \langle g \rangle = \{g^n : n \in \mathbb{Z}\} \). It follows from Lagrange’s Theorem that the order of any element of \( G \) is also a factor of the order of \( G \).

Examples 4.1.6.

1. In Example 4.1.3.1 we find that \(|G| = 6\), \(|H| = 3\), and \([G : H] = 2\). In the case of Example 4.1.3.2, we find that \(|G| = 6\), but \(|H| = 2\) and \([G : H] = 3\).

2. If \( \pi \in S_3 \), it is easy to see that the order of \( \pi \) can be 3 (\( \delta \) and \( \epsilon \)), 2 (\( \alpha \), \( \beta \), and \( \gamma \)), or 1 (I). Of course, in any of these cases, the order of \( \pi \) is a factor of the order of \( S_3 \). Note that 6 is also a factor of the order of \( S_3 \), but there is no element in \( S_3 \) of order 6.

3. In the study of the rings \( \mathbb{Z}_m \) we saw that the subgroups of \( \mathbb{Z}_m \) take the form \( \langle d \rangle \), where \( d \) is a divisor of \( m \). Obviously, in this case, the number of elements of the subgroup \( \langle d \rangle \) is \( \frac{m}{d} \), which is a factor of \( m \). Note also that \( d = [\mathbb{Z}_m : \langle d \rangle] \).

4. If \((A, +, \cdot)\) is any ring with identity 1, the order of the additive subgroup generated by 1 is precisely the characteristic of the ring \( A \). Hence, for a finite ring \( A \) its characteristic is always a factor of the number of elements of \( A \).

In many cases it is important to study the factorization of groups into smaller pieces, i.e., to establish conditions for a group \( G \) to factor as a direct product of groups \( K \) and \( H \). These two groups then are more “elementary blocks”, which can lead to a better knowledge of the structure of the group, an idea which we will pursue in the next chapter. We give now some results in this direction. Their application is often made easier by Lagrange’s Theorem.

Lemma 4.1.7. Let \( G \) be a group with identity \( e \). If \( K \) and \( H \) are normal subgroups in \( G \) such that \( K \cap H = \{e\} \), then \( kh = hk \) for any \( k \in K \) and \( h \in H \).

Proof. Let \( k \in K \) and \( h \in H \). Consider the element \( k^{-1}h^{-1}kh \). We have that \( h^{-1}kh \in K \), since \( K \) is a normal subgroup of \( G \). Hence, \( k^{-1}h^{-1}kh \in K \). Similarly, we have that \( k^{-1}h^{-1}k \in H \), since \( H \) is a normal subgroup, and

\[ \text{When we refer to the “group \( \mathbb{Z}_m \),” with no further qualitative, we always mean the additive group (\( \mathbb{Z}_m, + \)).} \]
hence \( k^{-1}h^{-1}kh \in H \). But \( k^{-1}h^{-1}kh \in K \cap H = \{ e \} \), so we conclude that \( k^{-1}h^{-1}kh = e \), i.e., that \( kh = hk \).

\[ \square \]

**Theorem 4.1.8.** Let \( H \) and \( K \) be subgroups of \( G \), and assume that:

(i) \( H \) and \( K \) are normal in \( G \);

(ii) \( G = HK \), and

(iii) \( H \cap K = \{ e \} \).

Then \( G \cong H \times K \).

**Proof.** Recall that the group \( H \times K \) as underlying set the Cartesian product \( H \times K = \{ (h,k) : h \in H, k \in K \} \), and binary operation \((h_1,k_1)(h_2,k_2) = (h_1h_2,k_1k_2)\).

Define \( \phi : H \times K \to G \) by \( \phi(h,k) = hk \). Using Lemma 4.1.7 one checks easily that \( \phi \) is a group homomorphism. The assumption \( G = HK \), implies immediately that \( \phi \) onto.

In order to determine the kernel of \( \phi \), observe that if \( \phi(h,k) = e \), then \( hk = e \). This means that \( h = k^{-1} \), hence we conclude that \( h,k \in H \cap K = \{ e \} \). It follows that \( h = k = e \), so the kernel of \( \phi \) consists of \( (e,e) \).

Hence, \( \phi \) is a group isomorphism and \( G \cong H \times K \).

As we have mentioned above, in the case of finite groups Lagrange’s Theorem is sometimes useful in verifying the assumptions of the previous theorem. For example, if \( G \) is a finite group, then \( |H \cap K| \) is a factor of \( |H| \) and \( |K| \). Hence, if \( |H| \) and \( |K| \) happen to be relative primes then we must have \( |H \cap K| = 1 \). If this is the case, the homomorphism \( \phi \) used in the proof of the theorem is 1:1 and we can conclude that \( |HK| = |H||K| \), so we can check if \( HK = G \).

Here is an example in the case of abelian groups, where any subgroup is obviously a normal subgroup:

**Example 4.1.9.**

In the group \( \mathbb{Z}_6 \) consider the subgroups \( H = \{ 0,3 \} \) and \( K = \{ 0,2,4 \} \). We know that \( H \) is isomorphic to \( \mathbb{Z}_2 \) and \( K \) is isomorphic to \( \mathbb{Z}_3 \). We have then \( H \cap K = \{ 0 \} \), and \( |G| = |H||K| \), so we conclude that \( \mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3 \).

More generally, suppose that \( \gcd(n,d) = 1 \) and recall Proposition ??: if \( m = nd \), the group \( \mathbb{Z}_m \) has subgroups \( B \cong \mathbb{Z}_n \) and \( C \cong \mathbb{Z}_d \). Since \( |B \cap C| \) is a factor of both \( n \) and \( d \), it follows immediately that \( |B \cap C| = 1 \). Hence, \( |B + C| = |B||C| = nd = m = |\mathbb{Z}_m| \) and so \( B + C = \mathbb{Z}_m \). We conclude that \( \mathbb{Z}_m \cong B \oplus C \cong \mathbb{Z}_n \oplus \mathbb{Z}_d \).
Theorem 4.1.10. Let $H_1, \ldots, H_n$ be subgroups of a group $G$ such that:

(i) Each $H_i$ is normal in $G$;
(ii) $G = H_1 \cdots H_n$, and
(iii) $H_i \cap \prod_{k=1, k \neq i}^n H_k = \{e\}$, for $i = 1, \ldots, n$.

Then: 

$$G \cong H_1 \times \cdots \times H_n.$$ 

Exercises.

1. Proof Proposition 4.1.2.
2. Show that $\{gH : g \in G\} \neq \{Hg : g \in G\}$, when $G = S_3$ and $H = \{I, \alpha\}$.
3. Show that if $e$ is the identity of $G$, then $g \equiv e \pmod{H}$ if and only if $g \in H$.
4. Find the quotient set $G/H$ when $G = \mathbb{Z}_6$ and $H = \langle 2 \rangle$.
5. Find the quotient set $G/H$ when $G = S_4$ and $H = \langle (1234) \rangle$.
6. Determine the order of every element in the groups $S_3$, $(\mathbb{Z}_6, +)$ and $(\mathbb{Z}_9^*, \cdot)$.
7. Let $G = S_n$ and $H = A_n$. Show that $\pi \equiv \sigma \pmod{H}$ if and only if $\pi$ and $\sigma$ are permutations with the same parity. Conclude that $[S_n : A_n] = 2$.
8. Show that the function $\phi : G/H \to H\setminus G$ given by $\phi(gH) = Hg^{-1}$ is a well defined bijection. Conclude that $[G : H]$ coincides also with the number of right cosets.
9. Show that if $K \subset H$, where $K$ and $H$ are subgroups of a finite group $G$, then $[G : K] = [G : H][H : K]$.
10. If $A$, $B$ and $C$ are any subsets of a group $G$, show that:
   (a) $(AB)C = A(BC)$.
   (b) $AA = A$ when $A$ is a subgroup of $G$.
   (c) If $A$ and $B$ are subgroups of $G$, then $AB = BA$ if and only if $AB$ is a subgroup of $G$. 

11. Show that, if $A$ and $B$ are finite subgroups of $G$, then $|AB||A \cap B| = |A||B|$.
(HINT: Use the fact that $A \cap B$ is a subgroup of both $A$ and $B$.)

12. Show that any permutation of $S_3$ is of the form $\pi = \alpha^n \varepsilon_m$, without explicitly computing these powers.

13. If $|G| = p$, where $p$ is a prime, what are the subgroups of $G$, and what is the order of the elements of $G$?

14. Give an example of an infinite group, where all elements have order 2, with the exception of the identity.

15. Give an example of an infinite group, where all elements have finite order, but such that for each $n \in \mathbb{N}$ it has elements of order $n$.


17. Show the following converse of Theorem 4.1.8: if $G$ is a group isomorphic to the direct product $H \times K$, then $G$ has normal subgroups $H'$ and $K'$ such that $H'K' = G$ and $H' \cap K' = \{e\}$, where $e$ is the identity in $G$.

18. Let $A$ be a finite ring with identity. Show that the characteristic of $A$ is a factor of $|A|$.

19. Classify the rings with unit with 2, 3, 4 and 5 elements.
(HINT: use the previous exercise.)

20. Let $G$ be an abelian group with 9 elements, which has no element of order 9. Show that $G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3$.

21. Show that if $G$ is a finite group where all elements, except the identity, have order 2, then $G$ is abelian. What can you say if all elements distinct from the identity have order 3?

22. If $G$ is a group with order $2n$, show that there exists at least one element in $G$ which has order 2.
(HINT: If $x \in G$, define $C(x) := \{x, x^{-1}\}$. Show that the binary relation $x \sim y \iff C(x) = C(y)$ is an equivalence relation and that $C(x)$ is the equivalence class of $x$.)
4.2 Quotient Group and Quotient Ring

The ring operations in \( \mathbb{Z}_m \) were defined from the ring operations in \( \mathbb{Z} \). When can one generalize the procedure adopted for \( \mathbb{Z}_m \), to define operations in the quotient \( G/H \), from given operations in \( G \)? We will see that this is possible only under some restrictions in the subgroup \( H \).

In fact, as we will show next, it is possible to define a group operation in the quotient \( G/H \), provided \( H \) is a normal subgroup of \( G \).

Recall that in the case of \( \mathbb{Z}_m \), the addition operation was defined because of the following property of addition in \( \mathbb{Z} \):

\[
(4.2.1) \quad \begin{cases} \quad x \equiv x' \pmod{m} \\ \quad y \equiv y' \pmod{m} \end{cases}, \text{then } x + y \equiv x' + y' \pmod{m}.
\]

Indeed, this implies that for any \( x \) and \( y \) in \( \mathbb{Z}_m \), we can define

\[
x + y := \overline{x + y}
\]

without any ambiguity caused by a choice of representatives \( x \) and \( y \) for each of the equivalence classes. The following example shows that property (4.2.1) does not generalize for congruence \( \pmod{H} \).

**Example 4.2.1.**

If \( G = S_3 \) and \( H = \{I, \alpha\} \), we can have \( g_1 \equiv g_1' \pmod{H} \) and \( g_2 \equiv g_2' \pmod{H} \) without having \( g_1g_2 \equiv g_1'g_2' \pmod{H} \): for example, take \( g_1 = g_1' = \alpha \) and \( g_2 = \gamma, g_2' = \delta \); then \( \gamma \equiv \delta \pmod{H} \), but \( \alpha \gamma = \varepsilon \) and \( \alpha \delta = \gamma \) are not equivalent \( \pmod{H} \).

In fact, property (4.2.1) generalizes only in the case of normal subgroups:

**Proposition 4.2.2.** Let \( H \) be a subgroup of \( G \). The following statements are equivalent:

(i) \( H \) is a normal subgroup of \( G \);

(ii) \( gHg^{-1} = H \) for all \( g \in G \);

(iii) \( (g_1H)(g_2H) = (g_1g_2)H \) for all \( g_1, g_2 \in G \);

(iv) if \( g_1 \equiv g_1' \pmod{H} \) and \( g_2 \equiv g_2' \pmod{H} \), then \( g_1g_2 \equiv g_1'g_2' \pmod{H} \) for all \( g_1, g_2, g_1', g_2' \in G \).

\footnote{Note that this restriction has no consequences for \( \mathbb{Z}_m \): since \((\mathbb{Z}, +)\) is an abelian group, any of its subgroups is a normal subgroup.}
Proof. Let us show that

\[(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i).\]

\[(i) \Rightarrow (ii):\] From the definition of a normal subgroup we have:

\[gHg^{-1} \subset H.\]

Since \(g\) is arbitrary, we can replace \(g\) by \(g^{-1}\) obtaining \(g^{-1}Hg \subset H\). Note also that

\[g^{-1}Hg \subset H = \Rightarrow g(g^{-1}Hg)g^{-1} \subset gHg^{-1} = \Rightarrow H \subset gHg^{-1}.\]

Since we know that \(gHg^{-1} \subset H\), we conclude that \(gHg^{-1} = H\).

\[(ii) \Rightarrow (iii):\] Since \(gHg^{-1} = H\), one sees immediately that

\[(gHg^{-1})g = Hg, \text{ i.e., } gH = Hg.\]

Hence, we find:

\[(g_1H)(g_2H) = ((g_1H)g_2)H = (g_1(Hg_2))H = (g_1g_2)H\]

\[= (g_1g_2)H = (g_1g_2)(HH) = (g_1g_2)H.\]

\[(iii) \Rightarrow (iv):\]

\[g_1 \equiv g'_1 \pmod{H} \text{ and } g_2 \equiv g'_2 \pmod{H} \Rightarrow g'_1 \in g_1H \text{ and } g'_2 \in g_2H.\]

Hence \(g'_1g'_2 \in (g_1H)(g_2H)\). Since \((g_1H)(g_2H) = (g_1g_2)H\), we find

\[g'_1g'_2 \in (g_1g_2)H \iff g_1g_2 \equiv g'_1g'_2 \pmod{H}.\]

\[(iv) \Rightarrow (i):\] If \(g \in G\) and \(h \in H\), we must show that \(ghg^{-1} \in H\). For this, observe that \(gh \equiv g \pmod{H}\) and \(g^{-1} \equiv g^{-1} \pmod{H}\). Hence we must have \(ghg^{-1} \equiv gg^{-1} = e \pmod{H}\). It follows that \(ghg^{-1} \in H\) as claimed.

By the previous proposition, if \(H\) is a normal subgroup of \(G\), then \((g_1H)(g_2H) = (g_1g_2)H\) defines a binary operation in \(G/H\). It is easy to check that:

**Theorem 4.2.3.** If \(H\) is a normal subgroup of \(G\), then \(G/H\) is a group for the binary operation defined by

\[(g_1H)(g_2H) = (g_1g_2)H.\]

Moreover, the quotient map \(\pi : G \to G/H\) defined by \(\pi(g) = g = gH\) is a surjective group homomorphism with kernel \(N(\pi) = H\).
4.2. QUOTIENT GROUP AND QUOTIENT RING

Proof. We saw above that we have a well defined binary operation $G/H \times G/H \rightarrow G/H$ given by $(g_1H, g_2H) \rightarrow (g_1H)(g_2H) = (g_1g_2)H$.

This binary operation is associative since:

$$((g_1H)(g_2H)) (g_3H) = ((g_1g_2)H) (g_3H)$$
$$= (g_1g_2g_3)H$$
$$= (g_1H)((g_2g_3)H)$$
$$= (g_1H)((g_2H)(g_3H)).$$

If $e$ is the identity in $G$, obviously $eH = H$ and $(gH)H = H(gH) = gH$. Hence, $H$ is the identity in $G/H$.

It is also clear that $(gH)(g^{-1}H) = (g^{-1}H)(gH) = eH = H$. This means that every element in $G/H$ has an inverse. This shows that $G/H$ is a group.

Let $\pi : G \rightarrow G/H$ be defined by $\pi(g) = g = gH$. It is immediate to check that

$$\pi(g_1)\pi(g_2) = (g_1H)(g_2H) = (g_1g_2)H = \pi(g_1g_2),$$

so $\pi$ is a group homomorphism.

Finally, since $H$ is the identity in the group $G/H$, we have that the kernel of $\pi$ is $N(\pi) = \{g \in G : gH = H\} = H$.

Examples 4.2.4.

1. Let $G = S_3$. The subgroup $H = A_3$ is normal, so that $G/H = \{I, \alpha\}$ is a group. We easily find the multiplication table:

$$I I = \alpha \alpha = L,$$
$$I \alpha = \alpha I = \alpha.$$

2. When $G = \mathbb{Z}_6$ and $H = \langle 2 \rangle = \{0, 2, 4\}$, we find $G/H = \{0, 1\}$, where:

$$0 = 0 + \langle 2 \rangle = \{0, 2, 4\}$$
$$1 = 1 + \langle 2 \rangle = \{1, 3, 5\}.$$

In this case the multiplication table is given by

$$0 + 0 = 1 + 1 = 0,$$
$$0 + 1 = 1 + 0 = 1.$$

Obviously, the groups $S_3/A_3$ and $\mathbb{Z}_6/\{0, 2, 4\}$ are isomorphic (there exists only one group with two elements!).
We leave as an exercise the proof of the following result:

**Theorem 4.2.5.** If \( H \) is a normal subgroup of \( G \), then the subgroups (respectively, normal subgroups) of \( G/H \) take the form \( K/H \), where \( H \subseteq K \subseteq G \), and \( K \) is a subgroup (respectively, normal subgroup) of \( G \).

Let us turn now to the study of quotient rings, which is slightly more involved. On the one hand, if \( B \subset A \) is a subring of \( A \), then \((B,+)\) is a subgroup of \((A,+), so we can form the quotient group \( A/B \): this follows immediately from the fact that \((A,+)\) is an abelian group, so any of its subgroups is normal. The “addition” in \( A/B \) is given by:

\[
a_1 + a_2 = a_1 + a_2, \text{ or } (a_1 + B) + (a_2 + B) = (a_1 + a_2)B.
\]

This does not imply that \( A/B \) is a ring, since we need also a “multiplication” in \( A/B \), such that the axioms of a ring are satisfied.

Let us recall that in the case of \( \mathbb{Z}_m \) the definition of multiplication relied on the following property:

\[
(4.2.2) \quad \text{If } \begin{cases} x \equiv x' \pmod{m} \\ y \equiv y' \pmod{m} \end{cases}, \text{ then } xy \equiv x'y' \pmod{m}.
\]

This property shows that given elements \( x \) and \( y \) in \( \mathbb{Z}_m \), the binary operation

\[
x \cdot y = xy
\]

is well defined independent of the choice of representatives \( x \) and \( y \) for each equivalence class. This suggests that for an arbitrary ring \( A \) with a subring \( B \subset A \) we should set:

\[
a_1 a_2 = a_1 a_2 \text{ or } (a_1 + B)(a_2 + B) = a_1 a_2 + B.
\]

However, this definition is consistent only if whenever \( a_1 \equiv a'_1 \pmod{B} \) and \( a_2 \equiv a'_2 \pmod{B} \) we also have \( a_1 a_2 \equiv a'_1 a'_2 \pmod{B} \). This is clarified by the next result:

**Proposition 4.2.6.** Let \( B \) be a subring of \( A \) and set \( a \equiv a' \pmod{B} \) if and only if \( a' - a \in B \). Then the following statements are equivalent:

(i) \( B \) is an ideal in \( A \);

(ii) If \( a_1 \equiv a'_1 \pmod{B} \) and \( a_2 \equiv a'_2 \pmod{B} \) then \( a_1 a_2 \equiv a'_1 a'_2 \pmod{B} \), for any \( a_1, a'_1, a_2, a'_2 \in A \).
4.2. QUOTIENT GROUP AND QUOTIENT RING

Proof. We will show that both implications (i) \( \Rightarrow \) (ii) and (ii) \( \Rightarrow \) (i), hold.

(i) \( \Rightarrow \) (ii): If \( a_1 \equiv a'_1 \pmod{B} \) and \( a_2 \equiv a'_2 \pmod{B} \), then \( a'_1 = a_1 + b_1 \) and \( a'_2 = a_2 + b_2 \), where \( b_1, b_2 \in B \). Hence, \( a'_1 a'_2 = (a_1 + b_1)(a_2 + b_2) = a_1 a_2 + a_1 b_2 + a_2 b_1 + b_1 b_2 \). It is clear that \( b_1 b_2 \in B \), because \( B \) is a subring, and \( a_1 b_2, a_2 b_1 \in B \), because \( B \) is an ideal. We conclude that \( a'_1 a'_2 = a_1 a_2 + b \), where \( b = a_1 b_2 + a_2 b_1 + b_1 b_2 \in B \), and hence \( a_1 a_2 \equiv a'_1 a'_2 \pmod{B} \).

(ii) \( \Rightarrow \) (i): We must show that, if \( a \in A \) and \( b \in B \), then \( ab, ba \in B \). For that, we observe that \( b \in B \) if and only if \( b \equiv 0 \pmod{B} \), where 0 denotes the zero in the ring \( A \). According to (ii), we then have \( ab \equiv a0 \pmod{B} \) and \( ba \equiv 0a \pmod{B} \), or equivalently \( ab \equiv 0 \pmod{B} \) and \( ba \equiv 0 \pmod{B} \). We conclude that \( ab, ba \in B \), so \( B \) is an ideal.

We can now state an analogue of Theorem 4.2.3 for rings. We leave its proof as an exercise.

Theorem 4.2.7. If \( I \leq A \) is an ideal in a ring \( A \), then \( A/I \) is a ring with the operations:

\[
\begin{align*}
\bar{a} + \bar{b} &:= a + b \\
\bar{a} \cdot \bar{b} &:= a \cdot b.
\end{align*}
\]

If \( A \) is abelian (respectively, with identity 1), then \( A/I \) is abelian (respectively, with identity \( \bar{1} \)). Moreover, the quotient map \( \pi : A \to A/I \) defined by \( \pi(a) = \bar{a} = a + I \) is a ring homomorphism with kernel \( N(\pi) = I \).

The next examples show that often properties of the ring \( A \) do not pass to the quotient.

Examples 4.2.8.

1. The rings \( \mathbb{Z}_m \) are examples of the situation described in the theorem, where \( A = \mathbb{Z} \) and \( I = \langle m \rangle \), where \( m \) is a fixed natural number. In this case, \( A \) is always an integral domain, while the quotient \( A/I \) is not, if \( m \) is not a prime number: if \( d \) is a factor of \( m \) then \( \overline{d} \) is a zero divisor in \( \mathbb{Z}_m \).

2. Let \( A = \mathbb{Q}[x] \) and \( I = \langle m(x) \rangle \) where \( m(x) = x^2 + 1 \). Given any \( p(x) \in \mathbb{Q}[x] \), the division algorithm shows that there exists \( q(x) \in \mathbb{Q}[x] \) such that

\[
p(x) = q(x)(x^2 + 1) + (a + bx), \quad (a, b \in \mathbb{Q}).
\]

Since obviously \( q(x)(x^2 + 1) \in I \), we conclude that \( p(x) \equiv a + bx \), i.e., \( \overline{p(x)} = \overline{a + bx} \), so that

\[
\frac{\mathbb{Q}[x]}{\langle x^2 + 1 \rangle} = \{ a + bx : a, b \in \mathbb{Q} \}.
\]
One can find easily the operations in this ring. For the sum we have:
\[
\alpha + b\alpha + a' + b'\alpha = a + b\alpha + a' + b'\alpha \\
= (a + a') + (b + b')\alpha.
\]
To find the product we observe first that \(x^2 + 1 \equiv 0\), i.e., that \(\alpha^2 = -1\). Therefore:
\[
(a + b\alpha)(a' + b'\alpha) = (a + b\alpha)(a' + b'\alpha), \\
= a(a' + b'b) + (a' + b')\alpha \\
= (aa' - b'b) + (a' + b')\alpha.
\]
In order to make the notation lighter, let us right \(a\) instead of \(\alpha\), and \(i\) instead of \(\alpha\) (note that \(a = b\) if and only if \(a = b\)). With these conventions, the two operations can be written:
\[
(a + bi) + (a' + b'i) = (a + a') + (b + b')i, \\
(a + bi)(a' + b'i) = (aa' - bb') + (ab' + a'b)i.
\]
It should now be clear that \(Q[x]/(x^2 + 1)\) is actually isomorphic to \(Q[i]\), a “coincidence” which will be explained later. Observe also that \(Q[x]/(x^2 + 1)\) is a field extension of \(Q\), while the original ring \(Q[x]\) was not a field. In this field the polynomial \(x^2 + 1\) has roots and is reducible.

3. The previous example, extends to rings of polynomials over finite fields. Let \(A = \mathbb{Z}_2[x]\) and \(I = (x^2 + x + 1)\). Just as before, if \(p(x) \in \mathbb{Z}_2[x]\) there exists \(q(x) \in \mathbb{Z}_2[x]\) such that
\[
p(x) = q(x)(x^2 + x + 1) + (a + bx).
\]
Therefore, once more \(p(x) \equiv a + bx\), i.e., \(p(x) = a + bx\). In this case, this leads to a finite ring with \(4\) elements:
\[
\mathbb{Z}_2[x]/(x^2 + x + 1) = \{a + bx : a, b \in \mathbb{Z}_2\} = \{0, 1, x, 1 + x\}.
\]
Again, we write \(\alpha\) instead of \(a\), and \(\alpha\) instead of \(x\), where \(1 + \alpha + \alpha^2 = 0\), or still \(\alpha^2 = -1 = 1 + \alpha\). (Note that since \(a = -a\) in \(\mathbb{Z}_2\), we also have \(a = -a\) in the quotient ring. We write \(\beta = \alpha^2 = 1 + \alpha\) for convenience):

\[
\begin{array}{c|cccc}
+ & 0 & 1 & \alpha & \beta \\
0 & 0 & 1 & \alpha & \beta \\
1 & 1 & 0 & \beta & \alpha \\
\alpha & \alpha & \beta & 0 & 1 \\
\beta & \beta & \alpha & 1 & 0
\end{array}
\]
\[
\begin{array}{c|cccc}
- & 0 & 1 & \alpha & \beta \\
0 & 0 & 1 & \alpha & \beta \\
1 & 0 & 1 & \alpha & \beta \\
\alpha & \alpha & \beta & 1 & 0 \\
\beta & \beta & \alpha & 0 & 1
\end{array}
\]

These are the same tables as the field with \(4\) elements mentioned in an exercise in Chapter ???. It is a field extension of \(\mathbb{Z}_2\), and in this field the polynomial \(x^2 + x + 1\) has roots and is reducible.
4.2. QUOTIENT GROUP AND QUOTIENT RING

In the exercises one is asked to check that, in general, the quotient ring $K[x]/\langle m(x) \rangle$ is always an extension of the field $K$ and also a vector space over $K$ of dimension $n$, where $n$ is the degree of the polynomial $m(x)$.

In the last two examples above the quotient ring is actually a field. On the other hand, we know that the rings $\mathbb{Z}_m$ are fields when $m$ is a prime, and this occurs precisely when $\langle m \rangle$ is a maximal ideal in $\mathbb{Z}$. These facts are indeed related:

**Theorem 4.2.9.** If $A$ is an abelian ring with identity and $I \subseteq A$ is an ideal, the quotient $A/I$ is a field if and only if $I$ is a maximal ideal in $A$.

**Proof.** Let us assume first that $I$ is a maximal ideal in $A$. We need to show that $A/I$ has an identity $1 \neq 0$, and that if $a \neq 0$, there exists $x \in A$ such that $ax = 1$.

Let us observe first that $1 \notin I$, i.e., $1 \neq 0$, for otherwise we would have $I = A$. Next, given $a \neq 0$, i.e., $a \notin I$, consider the set $J = \{ax + b : x \in A$ and $b \in I\}$. Obviously, $a \in J$ and so $J \neq I$. It is also obvious that $I \subset J$. Since $A$ is abelian, one checks immediately that $J$ is an ideal in $A$. Since $I$ is assumed to be a maximal ideal we must have $J = A$. Therefore, we have $1 \in J$ which means that there exists $x \in A$ and $b \in I$ such that $1 = ax + b$, or $ax = 1$.

Conversely, assume now that $A/I$ is a field and let $\pi : A \to A/I$ be the quotient map. If $J \supseteq I$ is an ideal in $A$ containing $I$, then $\{0\} \neq \pi(J) \subset A/I$ is an ideal. It follows that there exists $a \in J$ such that $a \equiv 1$, i.e., $1 = a + b$, for some $b \in I$. Since $I \subset J$, we conclude that $1 \in J$, so that $J = A$. This shows that $I$ is a maximal ideal.

Whenever $D$ is an integral domain, the polynomial ring $D[x]$ is also an integral domain, and $m(x)$ is irreducible if and only if $\langle m(x) \rangle$ is a maximal ideal maximal among all principal ideals in $D[x]$ (see Proposition 29). When $D = K$ is a field, all ideals in $K[x]$ are principal, so the theorem gives:

**Corollary 4.2.10.** If $K$ is a field, the quotient ring $K[x]/\langle m(x) \rangle$ is a field if and only if $m(x)$ is an irreducible polynomial in $K[x]$.

This corollary explains why in the examples above the quotient rings are fields. As in those examples, this corollary can be used to obtain field extensions of known fields and, in particular, to construct new fields. On the other hand, Theorem 4.2.9 can be used to define the real numbers from the rationals numbers and to verify that the axiomatics of the real numbers is also a consequence of the axioms of the integers, presented in Chapter 20. We shall do this in the next section, where we will also present a formal
CHAPTER 4. QUOTIENTS AND ISOMORPHISMS

definition of the complex numbers, identified as the quotient of \( \mathbb{R}[x] \) by the ideal \( \langle x^2 + 1 \rangle \).

Exercises.

1. Show that if \( H \) is a subgroup of \( G \) and \([G : H] = 2\), then \( H \) is a normal subgroup of \( G \).

2. Show that \( N \) is a normal subgroup of \( G \) if and only if there exists a group \( H \) and a homomorphism \( \phi : G \to H \) whose kernel is \( N \).

3. Let \( N \) be a normal subgroup of \( G \), and \( \pi : G \to G/N \) the quotient map.
   (a) Show that if \( N \subset H \subset G \) where \( H \) is a subgroup of \( G \), then \( N \) is a normal subgroup of \( H \), and \( H/N \) is a subgroup of \( G/N \).
   (b) Show that the subgroups of \( G/N \) take the form \( H/N \), where \( H \) is a subgroup of \( G \) which contains \( N \).

4. Let \( N \) be a normal subgroup of \( G \) and \( x \in G \).
   (a) Assume that the order of \( x \) in \( G \) is finite and equal to \( m \). Show that the order of \( x \) in \( G/N \) is finite and divides \( m \).
   (b) Find examples where \( x \) has infinite order in \( G \) and \( x \) in \( G/N \) has (i) finite order and (ii) infinite order.

5. Let \( A \) be a ring and \( I \) an ideal in \( A \), verify that the product in the quotient ring \( A/I \), which is defined as \( (a_1 + I)(a_2 + I) = a_1a_2 + I \), does not in general correspond to the product of sets, which was defined as \( CD = \{cd : c \in C \text{ and } d \in D\} \).

6. Prove Theorem 4.2.7.

7. Show that the function \( \phi : \mathbb{Q} \oplus \mathbb{Q} \to \mathbb{Q}[x]/\langle x^2 + 1 \rangle \), given by \( \phi(a, b) = ax + b \), is a bijection.

8. Find the tables for addition and multiplication of the quotient ring \( \mathbb{Z}_2[x]/\langle x^2 + 1 \rangle \). Check that this ring is not a field. Why doesn’t this contradict Theorem 4.2.9?

9. Consider the ring \( \mathbb{Q}[x]/\langle m(x) \rangle \), where \( m(x) = x^6 + x^4 + x^2 + 1 \). Determine the inverse of \( \bar{2} \). Check if this ring has any zero divisors and if yes, give an example of one such element.

10. Show that \( \mathbb{Q}[x]/\langle x^2 - 3x + 2 \rangle \) is isomorphic to \( \mathbb{Q} \oplus \mathbb{Q} \). Hint: Show that the map \( \phi : \mathbb{Q}[x]/\langle x^2 - 3x + 2 \rangle \to \mathbb{Q} \oplus \mathbb{Q} \) given by \( \phi(p(x)) = (p(1), p(2)) \) is well defined and yields an isomorphism of rings.
11. Let $L = \mathbb{Z}_2[x]/(x^2 + x + 1)$. Find the factorization of the polynomial $x^2 + x + 1$ in $L[x]$.

12. Let $m(x)$ be an irreducible polynomial of degree $n$ in $K[x]$, and let $L = K[x]/(m(x))$. Show that:
   (a) $L$ is a field and a vector of dimension $n$ over $K$;
   (b) The field $L$ is an extension of $K$;
   (c) $m(x)$ has at least one root in $L$;
   (d) There exists a field extension of $K$, where $m(x)$ is a product of factors of degree 1.

13. Show that the polynomial $x^3 + x^2 + 1$ is irreducible in $\mathbb{Z}_2[x]$. Use this fact to give an example of a field $L$ with 8 elements. Find the factorization of the polynomial $x^3 + x^2 + 1$ in $L$.

14. Let $I \subset A$ be an ideal, and $\pi : A \to A/I$ the quotient map $\pi(a) = a$. Show that $J \subset A/I$ is an ideal in $A/I$ if and only if $\pi(J)$, where $J$ is an ideal in $A$ and $I \subset J$.

15. Find all the ideals of $\mathbb{Z}_2[x]/(x^2 + 1)$.

16. Show that a non-abelian group $G$ with 6 elements is isomorphic to $S_3$, by showing that:
   (a) $G$ has an element $x$ of order 3 and $H = \langle x \rangle$ is normal in $G$.
   (b) $G$ has an element $y$ of order 2 and $y \notin H$.
   (c) Show that $yx = xy^2$ (use $yx \in xH$). Conclude that $G \cong S_3$.

### 4.3 Real and Complex Numbers

It is more or less obvious that the rational numbers can by represented by points in a line: to determine which point corresponds to a given rational number one just needs to fix two points in the line that correspond to the numbers 0 and 1. This correspondence was of course known to the Ancient Greeks, who also discovered an interesting phenomena: although to each rational number corresponds a point in the line, there are points in the line which do not correspond to any rational number. The Greeks interpreted this as mistake by the Gods, since it suggested that the rational numbers, a byproduct of the natural numbers, were in some sense insufficient. In fact, for some time they even tried to hide this phenomena from general
knowledge, afraid that it could cause the wrath of the Gods. Actually, as
we shall see in this section, the Greeks were wrong since the real numbers,
which correspond to all points in the line, can be defined from the rational
numbers, and therefore also from the natural numbers.

In modern language, the “insufficiency” of the rational numbers can be
expressed in terms of Cauchy sequences. We start by recalling its definition,
adapted to the special case of the rational numbers.

**Definition 4.3.1.** Let \( x = (x_1, x_2, \ldots) \) be a sequence in \( \mathbb{Q} \). The sequence
is called:

(a) **bounded**, if there exists \( M \in \mathbb{Q} \) such that
\[
|x_n| \leq M, \forall n \in \mathbb{N}.
\]

(b) **convergent** in \( \mathbb{Q} \) with limit \( l \in \mathbb{Q} \), if
\[
\forall \varepsilon \in \mathbb{Q}^+, \exists N \in \mathbb{N} : n \geq N \implies |x_n - l| < \varepsilon.
\]

(c) **Cauchy**, if
\[
\forall \varepsilon \in \mathbb{Q}^+, \exists N \in \mathbb{N} : n, m \geq N \implies |x_n - x_m| < \varepsilon.
\]

As usual, if \( x = (x_1, x_2, \ldots) \) is a convergent sequence with limit \( l \), then
one writes \( x_n \to l \) or \( \lim_{n \to \infty} x_n = l \). One has the usual results about
convergence of sums, products and differences of convergent sequences. Also,
it is easy to show that in \( \mathbb{Q} \):

(i) Every convergent sequence is Cauchy, and

(ii) Every Cauchy sequence is bounded.

On the other hand, there are Cauchy sequences in \( \mathbb{Q} \) which are not conver-
gent, as illustrated in the following example:

**Example 4.3.2.**

Consider the map \( f : \mathbb{Q} \to \mathbb{Q} \) defined by \( f(x) = \frac{x^2 + 2}{2x} \). If \( x > 0 \), we observe
that \( f(x) > 1 \), because
\[
(x - 1)^2 + 1 > 0 \quad \implies \quad x^2 - 2x + 2 > 0,
\]
\[
\quad \implies \quad x^2 + 2 > 2x,
\]
\[
\quad \implies \quad \frac{x^2 + 2}{2x} > 1.
\]
If \( x, y > 0 \), we also observe that
\[
f(x) - f(y) = \frac{(xy - 2)(x - y)}{2xy} = \frac{xy - 2x - y}{xy} - \frac{1}{2}.
\]

If, additionally, \( x, y \geq 1 \), it is easy to check that \(-1 \leq \frac{xy - 2}{xy} < 1\), since \( g(z) = 1 - \frac{2}{z} \) is increasing for \( z > 0 \). Therefore,
\[
|f(x) - f(y)| \leq \frac{1}{2}|x - y|.
\]

Now let \( \{x_n\}_{n \in \mathbb{N}} \) be the sequence in \( \mathbb{Q} \) defined by
\[
 x_1 = 1, \text{ and } x_{n+1} = f(x_n) \text{ se } n \in \mathbb{N}.
\]

For \( n > 1 \) we have:
\[
|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \leq \frac{1}{2}|x_n - x_{n-1}|,
\]
so that
\[
|x_{n+1} - x_n| \leq \frac{1}{2^{n-1}}|x_2 - x_1|.
\]
We leave as an exercise to verify that if \( m > n \) then
\[
|x_m - x_n| \leq \frac{1}{2^{n-2}}|x_2 - x_1|,
\]
so the sequence \( \{x_n\}_{n \in \mathbb{N}} \) is a Cauchy sequence.

Although this sequence is a Cauchy sequence, it is not convergent in \( \mathbb{Q} \). In fact, we have
\[
x_{n+1} = f(x_n) \Rightarrow 2x_{n+1}x_n = x_n^2 + 2,
\]
so that if \( x_n \to x \), then \( 2x^2 = x^2 + 2 \) or \( x^2 = 2 \), an equation that we know has no solutions in \( \mathbb{Q} \).

Although the sequence in the previous example is not convergent in \( \mathbb{Q} \), it is obviously convergent in \( \mathbb{R} \) to the irrational number \( \sqrt{2} \). You probably know that every Cauchy sequence in \( \mathbb{Q} \) converges to a real number, which may or may not be a rational number. This leads to the following motto:

- Every Cauchy sequence in \( \mathbb{Q} \) determines a real number. \footnote{You should compare this with the idea used to define the rational numbers from the integers: each pair \((m, n)\) of integers with \( n \neq 0 \) determines a rational number.}

Of course, distinct Cauchy sequences can determine the same real number, i.e., can have the same limit. This happens precisely when the difference of the two sequences converges to zero. In other words:
• The Cauchy sequences \( \{x_n\}_{n \in \mathbb{N}} \) and \( \{y_n\}_{n \in \mathbb{N}} \) determine the same real number if and only if \( (x_n - y_n) \to 0 \).

We define two Cauchy sequences, \( \{x_n\}_{n \in \mathbb{N}} \) and \( \{y_n\}_{n \in \mathbb{N}} \), to be equivalent if \( (x_n - y_n) \to 0 \). This leads to the main idea we will use to define the real numbers out of the rational numbers: a real number is an equivalence classe of Cauchy sequences in \( \mathbb{Q} \).

In order to pursue this idea, we need the following proposition, which is in fact a special case of the general theory developed in the previous section. Its proof is left as an exercise.

**Theorem 4.3.3.** Let \( A \) be the set of all sequences of rational numbers. Then:

(i) The set \( A \) with the operations of addition and product of sequences is a ring.

(ii) The subset \( B \subset A \) consisting of all Cauchy sequences in \( \mathbb{Q} \) is a subring of \( A \).

(iii) The set \( I \) formed by all sequences in \( \mathbb{Q} \) which converge to 0 is a subring of \( A \) and an ideal in \( B \).

If \( x, y \in B \) are Cauchy sequences in \( \mathbb{Q} \), then \( x \) and \( y \) determine the same real number if and only if \( x - y \) converges to 0, i.e., if and only if \( x - y \in I \). This suggests:

**Definition 4.3.4** (Cantor). The ring \( B/I \) is denoted by \( \mathbb{R} \). Its elements (which are just equivalence classes of Cauchy sequences in \( \mathbb{Q} \)) are called real numbers.

It should be clear that \( \mathbb{R} \) is an extension of the ring \( \mathbb{Q} \): given any rational \( q \in \mathbb{Q} \), we can take the constant sequence \( q \) given by \( q_n = q, n \in \mathbb{N} \) (obviously a Cauchy sequence), and the map \( \iota : \mathbb{Q} \to \mathbb{R} \) given by \( \iota(q) = q \) is an injective homomorphism. Observe also that the zero in \( \mathbb{R} \) is the equivalence class of the zero sequence (the ideal \( I \)), and the identity is the equivalence class of the constant sequence with all terms equal to 1. Of course, any sequence of rationals converging to 0 also represents \( I = 0 \), and any sequence of rationals converging to 1 is a representative of 1.

---

6This way of defining the real numbers is due to Georg Cantor (1845-1918), a German mathematician who also made fundamental discoveries in Set Theory, including the theory of “transfinite numbers”.

In order to show that \( \mathbb{R} \) is a field (which amounts to show that \( I \) is a maximal ideal in \( B \)), one needs to show that if \( x \in \mathbb{R} - \{0\} \) then there exists \( y \in \mathbb{R} \) such that \( xy = 1 \). In terms of Cauchy sequences in \( \mathbb{Q} \), this means we need to show that:

**Proposition 4.3.5.** If \( x \) is a Cauchy sequence in \( \mathbb{Q} \) which does not converge to 0, then there exists a Cauchy sequence \( y \) in \( \mathbb{Q} \) such that \( x_n y_n \to 1 \).

**Proof.** Let \( x \) be a Cauchy sequence in \( \mathbb{Q} \) which does not converge to 0. We leave as an exercise to show that there exists a rational \( \delta > 0 \) and a natural number \( N \in \mathbb{N} \) such that \( |x_n| > \delta \) for \( n \geq N \).

Next, we define a sequence \( y \in \mathbb{Q} \) by

\[
y_n = \begin{cases} 
0, & \text{se } n \leq N \\
\frac{1}{x_n}, & \text{se } n > N.
\end{cases}
\]

Notice that if \( n > N \) then \( |y_n| = \frac{1}{|x_n|} \leq \frac{1}{\delta} \), hence for \( n, m > N \) we have

\[
|y_m - y_n| = \frac{|x_m - x_n|}{|x_n x_m|} \leq \frac{1}{\delta^2} |x_n - x_m| \to 0,
\]

This shows that \( y \) is a Cauchy sequence in \( \mathbb{Q} \).

Since \( x_n y_n = 1 \) for \( n > N \), it is obvious that \( x_n y_n \to 1 \). \( \square \)

Next we show that \( \mathbb{R} \) is an ordered field. For that we need to define a set \( \mathbb{R}^+ \) such that:

1. \( x, y \in \mathbb{R}^+ \Rightarrow x + y \in \mathbb{R}^+ \) and \( xy \in \mathbb{R}^+ \);

2. If \( x \in \mathbb{R}^+ \), exactly one of the following 3 conditions hold:

\[
x \in \mathbb{R}^+ \text{ or } x = 0, \text{ or } -x \in \mathbb{R}^+.
\]

After a little thought, one sees there is really just one way to proceed:

**Definition 4.3.6.** If \( x \in \mathbb{R} \) (i.e., \( x \) is a Cauchy sequence in \( \mathbb{Q} \)), we say that \( x \) is positive if there exists a rational number \( \varepsilon > 0 \) and \( N \in \mathbb{N} \), such that \( n > N \Rightarrow x_n \geq \varepsilon \). The set of positive real numbers is denoted by \( \mathbb{R}^+ \).

It is now easy to show that

**Theorem 4.3.7.** \( \mathbb{R} \) is an ordered field.
It follows that, as it was explained in Chapter ?? in our discussion of ordered fields, that one can define \( |x| = \max\{x, -x\} \) for any \( x \in \mathbb{R} \).

If \( q \in \mathbb{Q} \) is a rational number we will denote by \( q \) the constant sequence where \( q_n = q \) for all \( n \in \mathbb{N} \). This is both a Cauchy sequence and a convergent sequence in \( \mathbb{Q} \). We denote by \( q \) the corresponding real number (the equivalence class determined by \( q \)). As we have already mentioned, the map \( f : \mathbb{Q} \to \mathbb{R} \) given by \( f(q) = q \) is a ring homomorphism, which allows us to say that the field \( \mathbb{R} \) is an extension of the field \( \mathbb{Q} \). In Analysis we learn that any real number can be approximated by a rational number, to an arbitrary small error, i.e., that “\( \mathbb{Q} \) is dense in \( \mathbb{R} \)”. In our approach to the real number, this idea is formalized precisely as follows:

**Proposition 4.3.8.** If \( x \) and \( \varepsilon \) are real numbers and \( \varepsilon > 0 \), then there exists a rational \( q \in \mathbb{Q} \) such that \( |x - q| < \varepsilon \).

**Proof.** We start by choosing representatives for \( x \) and \( \varepsilon \), i.e., Cauchy sequences \( x = (x_1, x_2, \ldots) \) and \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots) \) in \( \mathbb{Q} \). Since \( \varepsilon > 0 \), there exists a rational number \( r > 0 \) such that \( \varepsilon_n \geq r \) for all \( n \geq N_1 \), where \( N_1 \in \mathbb{N} \).

We can now obtain the sought “rational number” \( q \) by replacing \( x \) by a constant sequence whose general term is one of the terms of \( x \) of high enough order. In fact, since \( x \) is a Cauchy sequence, there exists \( N_2 \in \mathbb{N} \) such that

\[
 n, m \geq N_2 \implies |x_n - x_m| < \frac{r}{2}.
\]

If we let \( q = x_{N_2} \), a rational number, we have that \( \frac{r}{2} < x_n - q \leq \frac{r}{2} \) for \( n \geq \max\{N_1, N_2\} \). Hence, we conclude that \(-r < x - q < r\), and this implies that \( |x - q| < \varepsilon \). \( \square \)

The basic properties of the real numbers, which form the foundations of Analysis, are usually introduced in an axiomatic form. The usual minimal set of axioms that one considers includes only the axioms that \( \mathbb{R} \) is an ordered field and a completeness axiom, such as the Least Upper Bound Axiom or Supremum Axiom. This is invoked, for example, to show that every Cauchy sequence in \( \mathbb{R} \) is convergent, contrary to what happens in \( \mathbb{Q} \).

Our approach to the real numbers in a constructive (as opposed to an axiomatic) approach. We have already showed that \( \mathbb{R} \) is an ordered field and it remains to show that the “Supremum Axiom” is another consequence of our definition. For this, one shows directly that every Cauchy sequence in \( \mathbb{R} \) is convergent:

**Theorem 4.3.9.** Every Cauchy sequence in \( \mathbb{R} \) is convergent\(^7\).

\(^7\)For this reason, we say that \( \mathbb{R} \) is a complete field.
4.3. REAL AND COMPLEX NUMBERS

We leave the proof as an exercise. Using this result one can show the “Supremum Axiom” holds in $\mathbb{R}$.

**Corollary 4.3.10 (Supremum Axiom).** Every non-empty subset of $\mathbb{R}$ having an upper bound must have a least upper bound (a supremum).

**Proof.** Assume $A \subset \mathbb{R}$ is a non-empty set having an upper bound: this means there exists $M \in \mathbb{R}$ such that $x \leq M$, for all $x \in A$. One defines a sequence in $\mathbb{R}$, by successive bisections, a common method in Analysis. We let $x_1 = M$.

Since $A \neq \emptyset$, there exist $a \in A$ and we let $a_1 := a$. Obviously, $a_1 \leq x_1$, so we can set $a_2 := \frac{a_1 + x_1}{2}$. The number $a_2$ splits the interval $[a_1, x_1]$ into 2 subintervals of equal length. There are two possibilities:

(i) There exists $x \in A$ such that $x > a_2$. In this case we let $x_2 := x_1$;

(ii) One has $x \leq a_2$ for all $x \in A$. In this case, we let $x_2 := a_2$.

We repeat this procedure, obtaining a sequence of real numbers which is easily shown to be a Cauchy sequence. By Theorem 4.3.9 this sequence converges and one shows easily that its limit is a least upper bound of $A$. $\square$

This finishes the proof that the real numbers can be constructed from the rational numbers (and hence, from the integers) and that its properties are a consequence of the axioms for the integers that were discussed in Chapter ??.

One can construct the complex numbers from the real numbers with not so much difficulty: since $\mathbb{R}$ is an ordered field, it is obvious that the polynomial $x^2 + 1$ is irreducible in $\mathbb{R}[x]$. Therefore the ring

$$\mathbb{C} = \frac{\mathbb{R}[x]}{(x^2 + 1)}$$

is a field, called the FIELD OF COMPLEX NUMBERS. The imaginary unit $i$ is, by definition, the equivalence class of the polynomial $x$. It clearly satisfies the identity $i^2 = -1$ in $\mathbb{C}$. We will not discuss in detail the properties of $\mathbb{C}$, but one can show that $\mathbb{C}$ is also a complete field.

**Exercises.**

1. Let $A$ be an ordered field. Show that any convergent sequence in $A$ is a Cauchy sequence and that any Cauchy sequence is bounded.
2. Show that if one sets \( x_1 := 1 \) and \( x_{n+1} := f(x_n) \), where \( f \) is the function in Example 4.3.2, then \( |x_n - x_m| \leq \frac{1}{2^n-2}|x_2 - x_1| \).

3. Show that the set of all Cauchy sequences in \( \mathbb{Q} \) is a subring of the ring of all sequences in \( \mathbb{Q} \).

4. Show that the set of all sequences in \( \mathbb{Q} \) which converge to 0 is an ideal in the ring of all Cauchy sequences in \( \mathbb{Q} \).

5. Let \( x \) be a Cauchy sequence in \( \mathbb{Q} \). Show that the following statements are equivalent:

   (a) \( x \) does not converge to 0;
   (b) there exists a rational number \( \varepsilon > 0 \) and subsequence \( x_{n_k} \) such that \( |x_{n_k}| \geq \varepsilon \) for \( k \) large enough;
   (c) there exists a rational number \( d > 0 \) such that \( |x_n| \geq d \) for \( n \) large enough.

6. Let \( \underline{x}, \underline{y} \in \mathbb{R} \).
   
   (a) Show that if \( \underline{x}, \underline{y} \in \mathbb{R}^+ \), then \( \underline{x} + \underline{y} \in \mathbb{R}^+ \) and \( \underline{x} \underline{y} \in \mathbb{R}^+ \).
   (b) Show that either \( \underline{x} \in \mathbb{R}^+ \), \( \underline{x} = 0 \), or \( -\underline{x} \in \mathbb{R}^+ \), but none of these hold simultaneously.

7. Give a proof of Theorem 4.3.9 and finish the proof of Corollary 4.3.10.

8. Show that the order of the real numbers is unique, i.e., show that if \( \mathbb{R} \) is an ordered field then \( x \in \mathbb{R}^+ \) if and only if there exists \( y \in \mathbb{R} - \{0\} \) such that \( x = y^2 \).

9. Show that \( \mathbb{R} \) is not countable. Conclude that \( \mathbb{R} \) is a transcendental extension of \( \mathbb{Q} \) (and a vector space of infinite dimension over \( \mathbb{Q} \)).

10. Show that if \( x \) is a real number and \( 0 \leq x < 1 \), then there exists a sequence \( a_1, a_2, \ldots \) such that \( 0 \leq a_n \leq 9 \) for all \( n \in \mathbb{N} \) and

    \[
    x = \sum_{n=1}^{\infty} \frac{a_n}{10^n}
    \]

11. Show that \( \mathbb{C} \) is a complete field.
4.4 Isomorphism Theorems for Groups

Let $G$ and $H$ be groups and let $K \subseteq G$ be a normal subgroup. It is natural to ask what is the relationship between the group homomorphisms $\phi : G \to H$ and the group homomorphisms $\tilde{\phi} : G/K \to H$.

On the one hand, since the quotient map $\pi : G \to G/K$, given by $\pi(x) = x = xK$ is a group homomorphism, it is obvious that for any group homomorphism $\phi : G/K \to H$ the composition $\phi := \tilde{\phi} \circ \pi : G \to H$ is a group homomorphism and that $\tilde{\phi}(x) = \phi(x)$.

On the other hand, given a group homomorphism $\phi : G \to H$ one may wonder if there exists some homomorphism $\tilde{\phi} : G/K \to H$, such that $\tilde{\phi}(x) = \phi(x)$. This is illustrated by the following commutative diagram, where the dot arrow is used to indicate that we seek the existence of the corresponding homomorphism:

\[
\begin{array}{ccc}
G & \xrightarrow{\phi} & H \\
\downarrow{\pi} & & \\
G/K & \xleftarrow{\tilde{\phi}} & \bullet
\end{array}
\]

Let us assume that there exists a homomorphism $\tilde{\phi} : G/K \to H$ such that $\tilde{\phi}(x) = \phi(x)$. If $x \in K$ then $x = K$ is the identity in $G/K$, so $\tilde{\phi}(x)$ is the identity in $H$. Then $\phi(x) = \tilde{\phi}(x)$ is also the identity in $H$ so that $x \in N(\phi)$. In other words, a necessary condition for the existence of $\tilde{\phi} : G/K \to H$ is that $K \subseteq N(\phi)$. Actually, this condition is also sufficient:

**Proposition 4.4.1.** Let $K \subseteq G$ be a normal subgroup. Given a group homomorphism $\phi : G \to H$ there exists a group homomorphism $\tilde{\phi} : G/K \to H$ such that $\tilde{\phi}(x) = \phi(x)$ if and only if $K \subseteq N(\phi)$. Moreover, every group homomorphism $\tilde{\phi} : G/K \to H$ arises in this way.

**Proof.** We have already observed that if $\tilde{\phi} : G/K \to H$ is a group homomorphism then $\phi = \tilde{\phi} \circ \pi$ is a group homomorphism $\phi : G \to H$ with kernel $N(\phi) \supseteq K$.

For the converse, suppose that $\phi : G \to H$ is a group homomorphism with $K \subseteq N(\phi)$. Then we define $\tilde{\phi} : G/K \to H$ by setting:

$$\tilde{\phi}(xK) := \phi(x).$$

This map is well defined because if $x' = xk$, with $k \in K$, then $\phi(x') = \phi(xk) = \phi(x)\phi(k) = \phi(x)$. Also, $\tilde{\phi}$ is a group homomorphism because:

$$\tilde{\phi}(x) \cdot \tilde{\phi}(x') = \phi(x)\phi(x') = \phi(x \cdot x') = \tilde{\phi}(x \cdot x').$$
Examples 4.4.2.

1. Let $G = \mathbb{Z}$, $H = \mathbb{Z}_n$, and $K = \langle k \rangle$. Denoting the quotient map $\pi : \mathbb{Z} \to \mathbb{Z}_m$ by $\pi_m$, we consider the homomorphism $\phi = \pi_n : \mathbb{Z} \to \mathbb{Z}_n$ given by $\pi_n(x) = x \in \mathbb{Z}_n$.

Since the kernel of $\phi$ is $N(\phi) = \langle n \rangle$, there exists a homomorphism $\tilde{\phi} : \mathbb{Z}_k \to \mathbb{Z}_n$ such that $\tilde{\phi}(\pi_k(x)) = \pi_n(x)$ if and only if $\langle k \rangle \subseteq \langle n \rangle$, i.e., if and only if $n|k$.

We can write $\tilde{\phi}(x) = x$, but note, in general, $\tilde{\phi}$ is not the identity. For example, if $k = 4$ and $n = 2$, we have that $\tilde{\phi}(0) = \tilde{\phi}(2) = 0$, and $\tilde{\phi}(1) = \tilde{\phi}(3) = 1$.

2. Let $H = \{1, i, -1, -i\}$ and let $\phi : \mathbb{Z} \to H$ be the group homomorphism $\phi(n) = i^n$. Its kernel is $N(\phi) = \langle 4 \rangle$, so taking $K = N$, we conclude that there exists a group homomorphism $\tilde{\phi} : \mathbb{Z}_4 \to H$ such that $\tilde{\phi}(1) = i$, $\tilde{\phi}(2) = -1$, and $\tilde{\phi}(3) = -i$, its is clear that $\tilde{\phi}$ is an isomorphism.

3. Let $G = \mathbb{Z}$, $H = \mathbb{Z}_{210}$, $K = \langle k \rangle$, and consider the homomorphism $\phi : \mathbb{Z} \to \mathbb{Z}_{210}$ given by $\phi(x) = 36x$. The kernel of $\phi$ is 

$N(\phi) = \{x \in \mathbb{Z} : 210|36x\} = \{x \in \mathbb{Z} : 35|6x\} = \langle 35 \rangle$.

We conclude that there exists a group homomorphism $\tilde{\phi} : \mathbb{Z}_k \to \mathbb{Z}_{210}$ such that $\tilde{\phi}(x) = 36x$, if and only if $\langle k \rangle \subseteq \langle 35 \rangle$, i.e., if and only if $35|k$. In particular, $\phi : \mathbb{Z}_{70} \to \mathbb{Z}_{210}$, given by $\phi(x) = 36x$, is a well defined group homomorphism.

Obviously, if $\phi$ and $\tilde{\phi}$ are as in Proposition 4.4.1, then they have the same image, so $\tilde{\phi}$ is surjective if and only if $\phi$ is surjective. The injectivity of $\tilde{\phi}$ is discussed in the following:

**Proposition 4.4.3.** Let $\phi : G \to H$ be a group homomorphism with kernel $N(\phi) \supseteq K$, for some normal subgroup $K \subseteq G$. Denote by $\pi : G \to G/K$ the quotient map and $\tilde{\phi} : G/K \to H$ the induced homomorphism. Then the kernel of $\tilde{\phi}$ is $N(\phi)/K = \pi(N(\phi))$. In particular, $\tilde{\phi}$ is injective if and only $K = N(\phi)$. 
4.4. ISOMORPHISM THEOREMS FOR GROUPS

Proof. Let \( e \) be the identity of \( H \). Then:

\[
N(\tilde{\phi}) = \{ x \in G/K : \tilde{\phi}(x) = e \} \\
= \{ x \in G/K : \phi(x) = e \} \\
= \{ x \in G/K : x \in N(\phi) \} = \pi(N(\phi)) = N(\phi)/K.
\]

Examples 4.4.4.

1. In Example 4.4.2.3, where \( G = \mathbb{Z} \), \( H = \mathbb{Z}_{210} \), \( K = \langle 70 \rangle \), and \( \phi : \mathbb{Z} \to \mathbb{Z}_{210} \) is given by \( \phi(x) = 36x \), we saw that the kernel of \( \phi \) is \( N(\phi) = \langle 35 \rangle \). Hence, it follows that the kernel of the induced homomorphism \( \tilde{\phi} : \mathbb{Z}_{70} \to \mathbb{Z}_{210} \) is \( N(\phi)/K = \pi_{70}(\langle 35 \rangle) = \langle 35 \rangle = \{ 35, 0 \} \).

2. If in the previous example we let instead \( K = \langle 35 \rangle \), we conclude that the group homomorphism

\[
\tilde{\phi} : \mathbb{Z}_{45} \to \mathbb{Z}_{210}, \quad x \mapsto 36x
\]

is injective.

If \( \phi : G \to H \) is a surjective group homomorphism and \( K = N(\phi) \) the previous proposition reduces to an important basic result in Group Theory which we shall use repeatedly:

**Theorem 4.4.5 (First Isomorphism Theorem).** If \( \phi : G \to H \) is a surjective group homomorphism and \( K = N(\phi) \), then \( G/N(\phi) \) and \( H \) are isomorphic: there exists a group isomorphism \( \tilde{\phi} : G/N \to H \) given by \( \tilde{\phi}(\bar{x}) = \phi(x) \), for all \( x \in G \):

\[
\begin{array}{ccc}
G & \xrightarrow{\phi} & H \\
\downarrow{\pi} & & \downarrow{\phi} \\
G/N & \xrightarrow{\tilde{\phi}} & H
\end{array}
\]

The First Isomorphism Theorem allows one to establish that groups of a very distinct nature are actually isomorphic. Note that even when a homomorphism \( \phi : G \to H \) is not surjective we can still apply the theorem, replacing \( H \) by \( \phi(G) \), and concluding that there is an isomorphism:

\[
\phi(G) \simeq G/N(\phi).
\]
Examples 4.4.6.

1. We saw in Example 4.4.2.2 that the multiplicative group $H = \{1, i, -1, -i\}$ is isomorphic to the additive group $\mathbb{Z}_4$. More generally, consider the multiplicative group $R_n$ consisting of the $n$th-roots of the unit $R_n = \{\alpha^k : k \in \mathbb{Z}\} = \langle \alpha \rangle$, where $\alpha = e^{2\pi i/n}$. The group homomorphism $\phi : \mathbb{Z} \to R_n$ defined by $\phi(k) = \alpha^k$ is surjective and its kernel is $N(\phi) = \{k \in \mathbb{Z} : \alpha^k = 1\} = \langle n \rangle$. Hence, $R_n$ is isomorphic to the additive group $\mathbb{Z}_n$.

2. Let $\phi : S_n \to \mathbb{Z}_2$ be the surjective homomorphism given by $\phi(\rho) = \text{sgn}(\rho)$. Its kernel (by definition) is the alternating group $A_n$. Hence we conclude that $S_n/A_n$ is isomorphic to $\mathbb{Z}_2$.

3. Let $\phi : \mathbb{Z} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$ be the group homomorphism $\phi(x) = (\pi_m(x), \pi_n(x))$. Its kernel is given by:

$$N(\phi) = \{x \in \mathbb{Z} : x \equiv 0 \pmod{m} \text{ and } x \equiv 0 \pmod{n}\} = \{x \in \mathbb{Z} : m|\text{gcd}(x,n) = \text{lcm}(m,n)\}.$$  

It follows that the homomorphism $\tilde{\phi} : \mathbb{Z}_{\text{lcm}(m,n)} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$, $x \mapsto (\tilde{x}, \tilde{x})$ is injective, and that $\mathbb{Z}_{\text{lcm}(m,n)} \cong \tilde{\phi}(\mathbb{Z}_{\text{lcm}(m,n)})$.

In particular, when $m$ and $n$ are relatively prime numbers, so that we have $\text{lcm}(m,n) = mn$, it follows that $\mathbb{Z}_{mn}$ and $\mathbb{Z}_m \oplus \mathbb{Z}_n$ are isomorphic, since both groups have the same number of elements. We leave it to the exercises to show that this map is actually the only isomorphism of rings between $\mathbb{Z}_{mn}$ and $\mathbb{Z}_m \oplus \mathbb{Z}_n$, and to determine its inverse.

The example above concerning the roots of the unit is much more general that it looks at first sight. In fact, if $G$ is a multiplicative group with identity $e$ and $\alpha \in G$, we know that the group generated by $\alpha$ is $\langle \alpha \rangle = \{\alpha^k : k \in \mathbb{Z}\}$. The group homomorphism $\phi : \mathbb{Z} \to \langle \alpha \rangle$ given by $\phi(k) = \alpha^k$ is always surjective and its kernel is $N(\phi) = \{k \in \mathbb{Z} : \alpha^k = e\}$. Since $N(\phi)$ is a subgroup of $\mathbb{Z}$, we have that $N(\phi) = \langle n \rangle$, where $n \geq 0$. There are two possibilities for $n$:

(i) $n = 0 \iff N(\phi) = \{0\}$: this means that $\phi$ is injective, so we conclude that $\langle \alpha \rangle \cong \mathbb{Z}$, and $\langle \alpha \rangle$ is an infinite group. The element $\alpha$ has infinite order.

(ii) $n > 0 \iff N(\phi) \neq \{0\}$: then $n$ is the smallest positive integer in $N(\phi)$, i.e., the smallest positive solution of the equation $\alpha^k = e$. In this case, $\langle \alpha \rangle \cong \mathbb{Z}/(n) = \mathbb{Z}_n$, and $\langle \alpha \rangle$ had $n$ elements. Therefore, $\alpha$ has order $n$, and the order of an element $\alpha$ is precisely the smallest natural number such that $\alpha^k = e$. 

4.4. ISOMORPHISM THEOREMS FOR GROUPS

Examples 4.4.7.

1. Consider the permutation $\varepsilon$ in $S_3$. We have $\varepsilon^1 = \varepsilon$, $\varepsilon^2 = \delta$ and $\varepsilon^3 = I$. Hence, $\varepsilon$ has order 3, and $\langle \varepsilon \rangle = \{\varepsilon, \delta, I\} = A_3 \cong \mathbb{Z}_3$.

2. Recall that $D_5$ is the symmetry group of a regular pentagon. The order of a non-trivial rotation $r \in D_5$ is 5, so that $\langle r \rangle \cong \mathbb{Z}_5$. More generally, the group $D_n$ has a subgroup $H \cong \mathbb{Z}_n$, and this is a normal subgroup in $D_n$ (why?).

Definition 4.4.8. A group $G$ is called a **cyclic group** if there exists some $g \in G$ such that $\langle g \rangle = G$. The element $g$ is called a **generator** of $G$.

Examples 4.4.9.

1. The group $\mathbb{Z}$ is cyclic, with generators 1 and $-1$.

2. $A_3$ is a cyclic group: we can take $g = \varepsilon$ or $g = \delta$.

3. The group $\{1, i, -1, -i\}$ is cyclic: we can take either $g = i$ or $g = -i$.

4. The groups $\mathbb{Z}_n$ are cyclic: any element of $\mathbb{Z}_n^*$ is a generator.

5. The $\mathbb{Z}_2 \oplus \mathbb{Z}_4$ is not cyclic (why?).

The next result identifies all cyclic groups and it is just summarizes what we saw above:

**Corollary 4.4.10** (Classification of cyclic groups). If $G$ is a cyclic group then either:

(i) $G$ is infinite and $G \cong \mathbb{Z}$, or

(ii) $G$ is finite with $n$ elements and $G \cong \mathbb{Z}_n$.

Also, applying Lagrange’s Theorem, we obtain the classification of all groups of prime order:

**Corollary 4.4.11** (Classification of groups of prime order). If $G$ is a finite group of order $p$, with $p$ prime, then $G \cong \mathbb{Z}_p$.

There are also other isomorphism theorems for groups, which are less used than the First Isomorphism Theorem, but which are still useful. They are actually consequences of the First Isomorphism Theorem.
Theorem 4.4.12 (Second Isomorphism Theorem). Let $N$ and $H$ be subgroups of $G$, and assume that $N$ is normal in $G$. Then $HN$ is a subgroup of $G$, $N$ is normal in $HN$, $H \cap N$ is normal in $H$, and
\[
\frac{HN}{N} \cong \frac{H}{H \cap N}.
\]
Proof. Exercise 10 in Section 4.1 shows that if $H$ and $N$ are subgroups of $G$, then $HN$ is a subgroup of $G$ if and only if $HN = NH$. Since $N$ is normal in $G$, we have that $HN = NH$, and we conclude that $HN$ is a subgroup of $G$.

Now take the quotient homomorphism $\pi : G \to G/N$ and restrict it to $H$: one obtains a group homomorphism $\phi : H \to G/N$ given by
\[
\phi(x) = \pi(x) = x, \quad (x \in H).
\]
The kernel of $\phi$ is $N(\phi) = \{x \in H : x \in N\} = H \cap N$. On the other hand, its image $\phi(H)$ is a subgroup of $G/N$, so that $\phi(H) = K/N$, where $K$ is a subgroup of $G$ which contains both $H$ and $N$. Therefore $HN \subseteq K$. But also any element in $K$ is equivalent to an element in $H$, i.e., if $k \in K$ then there exists $h \in H$ and $n \in N$ such that $k = hn$. Hence we have that $K = HN$.

Now the First Isomorphism Theorem applied to $\phi$ gives:
\[
\frac{HN}{N} \cong \frac{H}{H \cap N}.
\]
\[\square\]

Notice, as a special case of the Second Isomorphism Theorem, that if $H \cap N = \{e\}$ then $HN/N \cong H$.

Theorem 4.4.13 (Third Isomorphism Theorem). If $K \subseteq H \subseteq G$ are normal subgroups of $G$, then $K$ is a normal subgroup of $H$, $H/K$ is a normal subgroup of $G/K$, and
\[
\frac{G/K}{H/K} \cong \frac{G}{H}.
\]
Proof. Assume that $K \subseteq H \subseteq G$, where $K$ and $H$ are normal in $G$. Let $\pi_K : G \to G/K$ and $\pi_H : G \to G/H$ be the usual quotient homomorphisms, so that $\pi_K$ has kernel $K$ and $\pi_H$ has kernel $H$. Consider the diagram:
\[
\begin{array}{ccc}
G & \xrightarrow{\phi=\pi_H} & G/H \\
\pi=\pi_K \downarrow & & \downarrow \\
G/K & & \\
\end{array}
\]
The existence of \( \tilde{\phi} \) is a consequence of the assumption \( K \subseteq H \) (see Proposition 4.4.1). The homomorphism \( \tilde{\phi} \) is clearly surjective, since \( \pi_H \) is surjective. Applying Proposition 4.4.3, the kernel of \( \tilde{\phi} \) is the group \( H/K \). Now the First Isomorphism Theorem applied to the homomorphism \( \tilde{\phi} : G/K \to G/H \) yields the conclusion of the theorem.

This result states that the quotients of quotients of \( G \) are isomorphic to quotients of \( G \). Notice, by the way, that this is just a generalization of the remarks in Example 4.4.2.1.

Examples 4.4.14.

1. Let \( G = \mathbb{Z} \), \( H = \langle 3 \rangle \), and \( K = \langle 6 \rangle \). Obviously \( K \subset H \), and since both \( K \) and \( H \) are normal subgroups of \( \mathbb{Z} \), we find
   \[
   \frac{G}{K} = \mathbb{Z}/\langle 6 \rangle = \mathbb{Z}_6, \\
   \frac{H}{K} = \langle 3 \rangle/\langle 6 \rangle = \langle 2 \rangle \subset \mathbb{Z}_6, \text{ and} \\
   \frac{G}{H} = \mathbb{Z}/\langle 3 \rangle = \mathbb{Z}_3.
   \]
   According to the Third Isomorphism Theorem, we conclude that \( \mathbb{Z}_6/\langle 2 \rangle \) and \( \mathbb{Z}_3 \) are isomorphic.

2. The previous example is a special instance of the following general fact: if \( n \mid m \) then \( \mathbb{Z}_m/\langle n \rangle \cong \mathbb{Z}_n \). In fact, let \( G = \mathbb{Z} \), \( H = \langle n \rangle \), and \( K = \langle m \rangle \), so that \( K \subset H \) are normal subgroups in \( \mathbb{Z} \). Then \( G/K = \mathbb{Z}_m \), \( G/H = \mathbb{Z}_n \), and \( H/K = \langle n \rangle \subset \mathbb{Z}_m \), so that the Third Isomorphism Theorem yields our claim.

Exercises.

1. Let \( H = \langle g \rangle = \{ g^n : n \in \mathbb{Z} \} \) be a cyclic group. Show that
   (a) if \( H \) is infinite, its only generators are \( g \) and \( g^{-1} \);
   (b) if \( H \) has \( m \) elements, the order of \( g^n \) is \( \frac{m}{\gcd(n,m)} \), where \( d = \gcd(n,m) \);
   (c) if \( H \) has \( m \) elements, \( g^n \) is a generator of \( H \) if and only if \( \gcd(n,m) = 1 \).

2. Assume that \( g_1 \) and \( g_2 \) belong to an abelian group \( G \) and have order, respectively, \( n \) and \( m \). Show that the order of \( g_1 g_2 \) divides \( \text{lcm}(n,m) \). Conclude that the subset formed by all elements of finite order is a subgroup of \( G \).

3. Decide which of the groups \( \mathbb{Z}_4, \mathbb{Z}_2 \oplus \mathbb{Z}_2, \mathbb{Z}_8, \mathbb{Z}_4 \oplus \mathbb{Z}_2 \) and \( \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \) are isomorphic.

4. Let \( G = \mathbb{Z} \oplus \mathbb{Z} \) and \( N = \{ (n,n) : n \in \mathbb{Z} \} \). Show that \( G/N \cong \mathbb{Z} \).
5. Consider the multiplicative group \( H = \langle e^{\pi i/4} \rangle \subset \mathbb{C} \) formed by the complex solutions of \( z^8 = 1 \).

(a) Determine all the generators and subgroups of \( H \);
(b) Determine all the automorphisms \( \phi : H \to H \).

6. Determine all the generators and subgroups of \( H = \langle (123456) \rangle \subset S_6 \).

7. Show that \( \text{Aut}(\mathbb{Z}_n) \simeq \mathbb{Z}_n^* \).

8. Show that \( \mathbb{Z}_{mn} \) and \( \mathbb{Z}_m \oplus \mathbb{Z}_n \) are isomorphic groups/rings if and only if \( m \) and \( n \) are relatively prime.

9. Assume that \( G \) is a finite group, \( H \) is a normal subgroup of \( G \), \( K \) is a subgroup of \( G \), \( G = HK \), and \( G/H \) is isomorphic to \( K \). Show that \( H \cap K = \{e\} \).

10. Show that if \( n > 1 \), then \( \mathbb{Z}_{p^n} \) is not isomorphic to \( \bigoplus_{k=1}^{n} \mathbb{Z}_p \).

11. Show that the quotient \( \mathbb{Z}_{40}/\langle 15 \rangle \) is isomorphic to \( \mathbb{Z}_n \) and find \( n \).

12. Assume that the only subgroups of \( G \) are the trivial groups \( \{1\} \) and \( G \). Show that \( G \) is a cyclic group of prime order.

13. Show that if \( G \) is an abelian group of order \( pq \), when \( p \) and \( q \) are both prime, then \( G \) is cyclic.

14. Classify the groups with \( 2p \) elements, where \( p > 2 \) is a prime. (Hint: Show that there exists an element \( x \) of order \( p \), and that every element of order \( p \) belongs to \( \langle x \rangle \)).

15. Classify the groups with 8 elements, by proceeding as follows:

(a) Show that if \( G \) is abelian, then it is isomorphic to one of the groups \( \mathbb{Z}_8 \), \( \mathbb{Z}_4 \oplus \mathbb{Z}_2 \), or \( \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \), and these are non isomorphic groups.

(b) Assuming that \( G \) is not abelian, show that:
   (i) \( G \) has an element \( x \) of order 4, and \( H = \langle x \rangle \) is normal in \( G \).
   (ii) Assuming \( y \notin H \), show that \( y^3 \in H \), where \( y^2 = 1 \) or \( y^3 = x^2 \).
   (iii) Finally show that \( yx \in Hy \), where \( yx = x^3y \) (note that the order of \( yxy^{-1} \) is the order of \( x \)).
   (iv) Compare the result with the tables of \( D_4 \) and \( H_8 \).

16. Let \( G \) and \( H \) be groups, with normal subgroups \( K \subset G \) and \( N \subset H \). Show that \( (G \times H)/(K \times N) \) is isomorphic to \( (G/K) \times (H/N) \).
4.5 Isomorphism Theorems for Rings

When \( A \) and \( B \) are rings, \( I \subseteq A \) is an ideal of \( A \), and \( \phi : A \rightarrow B \) is a homomorphism of rings, we may apply the results from the previous section to \( \phi \) as a homomorphism of the additive groups \((A,+)\) and \((B,+).\) In particular, we know that if the kernel of \( \phi \) contains \( I \), then there exists a group homomorphism \( \tilde{\phi} : A/I \rightarrow B \) such that we have a commutative diagram.

\[
\begin{array}{ccc}
A & \xrightarrow{\phi} & B \\
\downarrow{\pi} & & \downarrow{\phi} \\
A/I & \xrightarrow{\tilde{\phi}} & B
\end{array}
\]

However, we note that if \( \phi \) is a homomorphism of rings then we have

\[
\tilde{\phi}(\bar{x}) \cdot \tilde{\phi}(\bar{x'}) = \phi(x)\phi(x') = \phi(x \cdot x') = \tilde{\phi}(\bar{x \cdot x'}).
\]

Hence, the group homomorphism \( \tilde{\phi} \) is a ring homomorphism as long as the original homomorphism \( \phi \) is a ring homomorphism. This allows us to obtain immediately the analogues for rings of Propositions 4.4.1 and 4.4.3.

**Proposition 4.5.1.** Let \( A \) and \( B \) be rings, \( I \subseteq A \) an ideal in \( A \), and denote by \( \pi : A \rightarrow A/I \) the quotient map, \( \pi(x) = x + I \).

(i) The ring homomorphisms \( \tilde{\phi} : A/I \rightarrow B \) are the maps of the form \( \tilde{\phi}(\pi(x)) = \phi(x) \), where \( \phi : A \rightarrow B \) is a ring homomorphism with kernel \( N(\phi) \supseteq I \).

(ii) If \( \phi : A \rightarrow B \) is a ring homomorphism with kernel \( N(\phi) \supseteq I \), then the induced ring homomorphism \( \tilde{\phi} : A/I \rightarrow B \) has kernel \( N(\phi)/I = \pi(N(\phi)) \). In particular, \( \tilde{\phi} \) is injective if and only if \( I = N(\phi) \).

Given a ring \( A \), a map \( \phi : \mathbb{Z} \rightarrow A \) is a homomorphism of the underlying additive groups if and only if \( h(n) = na \), for some fixed \( a \in A \). It is easy to check that \( \phi \) is a ring homomorphism if and only if additionally \( a = \phi(1) \) is a solution of \( x^2 = x \) in \( A \). We make use of this simple remark to look back at the examples of the previous section, now from the perspective of their ring structure.
Examples 4.5.2.

1. Let $A = \mathbb{Z}$, $B = \mathbb{Z}_n$, and $I = \langle k \rangle$. The quotient $\phi = \pi_n : \mathbb{Z} \rightarrow \mathbb{Z}_n$ is a homomorphism of rings. If $n | k$ then

$$
\begin{array}{c}
\mathbb{Z} \\
\downarrow \phi = \pi_n \\
\mathbb{Z}_n \\
\downarrow \\
\mathbb{Z}_k \\
\end{array}
$$

where $\tilde{\phi} : \mathbb{Z}_k \rightarrow \mathbb{Z}_n$, given by $\tilde{\phi}(\pi_k(x)) = \pi_n(x)$ is a homomorphism of rings.

2. Let $A = \mathbb{Z}$, $B = \mathbb{Z}_{210}$ and $I = \langle k \rangle$. The equation $x^2 = x$ has (non-obvious) solutions in $\mathbb{Z}_{210}$, such as $x = 21$. Hence, the homomorphism $\phi : \mathbb{Z} \rightarrow \mathbb{Z}_{210}$ given by $\phi(x) = \pi_{210}(21x)$ is a ring homomorphism. It is easy to check that its kernel is $N(\phi) = \langle 10 \rangle$. We conclude that there exists a ring homomorphism $\tilde{\phi} : \mathbb{Z}_k \rightarrow \mathbb{Z}_{210}$ such that $\tilde{\phi}(\pi_k(x)) = 21x$ if and only if $10 | k$. Obviously, $\tilde{\phi}$ is injective when $k = 10$. In fact, when $k = 10$ we have that $\phi(\mathbb{Z}) = \tilde{\phi}(\mathbb{Z}_{10}) = \langle 21 \rangle$ is unitary subring of $\mathbb{Z}_{210}$, isomorphic to $\mathbb{Z}_{10}$.

$$
\begin{array}{c}
\mathbb{Z} \\
\downarrow \phi \\
\mathbb{Z}_{210} \\
\downarrow \\
\mathbb{Z}_{10} \\
\end{array}
$$

3. Let $A = \mathbb{Q}[x]$, $B = \mathbb{Q}$, and denote by $\phi : \mathbb{Q}[x] \rightarrow \mathbb{Q}$ the evaluation map $\phi(p(x)) = p(1)$. By the Polynomial Remainder Theorem, we know that $\phi$ is a homomorphism of rings with kernel $N(\phi) = \langle x - 1 \rangle$. If $I = \langle m(x) \rangle$ is the ideal of $\mathbb{Q}[x]$ generated by the polynomial $m(x)$, it follows that there exists a homomorphism of rings $\tilde{\phi} : \mathbb{Q}[x]/I \rightarrow \mathbb{Q}$, defined by $\tilde{\phi}(\overline{p(x)}) = p(1)$, if and only if $(x - 1)m(x)$, i.e, if and only if $p(1) = 0$.

Using Proposition 4.5.1 we obtain immediately the First Isomorphism Theorem for rings:

**Theorem 4.5.3** (First Isomorphism Theorem for rings). If $\phi : A \rightarrow B$ is a surjective ring homomorphism and $I$ is the kernel of $\phi$, the the rings $A/I$ and $B$ are isomorphic. In particular, there is an isomorphism of rings $\tilde{\phi} : A/I \rightarrow B$ such that $\tilde{\phi}(a) = \phi(a)$, for all $a \in A$.

The other isomorphism theorem for rings follow by a simple application of the First Isomorphism Theorem for rings, exactly in the same manner as we did in the case of groups. We state them and leave the easy proofs for the exercises:
4.5. ISOMORPHISM THEOREMS FOR RINGS

Theorem 4.5.4 (Second Isomorphism Theorem for rings). Let $A$ be a ring, $I$ an ideal of $A$, and $B$ a subring of $A$. Then $I + B$ is a subring of $A$, $I$ is an ideal in $I + B$, $I \cap B$ is an ideal in $B$, and there is an isomorphism of rings

$$\frac{I + B}{I} \cong \frac{B}{I \cap B}.$$  

Theorem 4.5.5 (Third Isomorphism Theorem for rings). Let $A$ be a ring, $I,J$ ideals of $A$ with $I \subseteq J$. Then $I$ is an ideal of $J$, $J/I$ is an ideal of $A/I$, and there is an isomorphism of rings $A/I \cong J/I \cong A/J$. 

Examples 4.5.6.

1. Let us check that when $n$ and $m$ are relatively prime then the rings $\mathbb{Z}_{nm}$ and $\mathbb{Z}_n \oplus \mathbb{Z}_m$ are isomorphic. Just like in the case of groups, it is enough to observe that $\phi = \pi_{nm} : \mathbb{Z} \to \mathbb{Z}_n \oplus \mathbb{Z}_m$ given by $\phi(k) = (\pi_n(k), \pi_m(k))$ is a homomorphism of rings. Hence, by the First Isomorphism Theorem for rings, the isomorphism of groups $\tilde{\phi} : \mathbb{Z}_{nm} \to \mathbb{Z}_n \oplus \mathbb{Z}_m$ in Example 4.4.6.3 is also an isomorphism of rings.

2. Let $A = \mathbb{Z}$ and assume that $n \mid m$. If we let $I = \langle m \rangle$ and $J = \langle n \rangle$, then $J \supseteq I$ and $I$ and $J$ are ideals in $\mathbb{Z}$. In this case, $A/I = \mathbb{Z}_m$, $A/J = \mathbb{Z}_n$, and $J/I = \langle n \rangle \subseteq \mathbb{Z}_m$. We conclude from the Third Isomorphism Theorem that the rings $\mathbb{Z}_m/\langle n \rangle$ and $\mathbb{Z}_n$ are isomorphic. In particular, the quotient rings formed from the rings $\mathbb{Z}_m$ are always rings $\mathbb{Z}_n$.

3. Let $\alpha \in \mathbb{C}$ be algebraic over $\mathbb{Q}$, $\alpha \notin \mathbb{Q}$. Denote by $m(x)$ its minimal polynomial. Recall that the evaluation map $\phi : \mathbb{Q}[x] \to \mathbb{C}$ given by $\phi(p(x)) = p(\alpha)$ is a homomorphism of rings, with kernel $N(\phi) = \langle m(x) \rangle$ and that $\mathbb{Q}[\alpha] = \phi(\mathbb{Q}[x])$. We conclude by the First Isomorphism Theorem for rings that

$$\mathbb{Q}[\alpha] \simeq \frac{\mathbb{Q}[x]}{\langle m(x) \rangle}.$$ 

Notice that since $m(x)$ is an irreducible polynomial, we have that $\mathbb{Q}[x]/\langle m(x) \rangle$ is a field. Hence (why?):

$$\mathbb{Q}[\alpha] \simeq \frac{\mathbb{Q}[x]}{\langle m(x) \rangle} \simeq \mathbb{Q}(\alpha).$$
We now illustrate the Isomorphism Theorems for rings with two applications.

First, we can use the result of Example 4.5.6.1 to compute the Euler function \( \varphi : \mathbb{N} \to \mathbb{N} \). Recall that this function is defined as \( \varphi(n) = |\mathbb{Z}_n^*| \), i.e., \( \varphi(n) \) is the number of invertible elements in the ring \( \mathbb{Z}_n \), or also, the number of integers \( 1 \leq k \leq n \) which are relatively prime to \( n \).

**Lemma 4.5.7.** If \( n_1, \ldots, n_k \) are relatively prime, then

\[
\varphi(n_1 \cdots n_k) = \varphi(n_1) \cdots \varphi(n_k).
\]

**Proof.** We give a proof only for the case \( k = 2 \). The cases \( k > 2 \) can be obtained by a simple induction procedure.

Recall that if \( A \) and \( B \) are rings with units, then \((A \oplus B)^* = A^* \times B^*\). Hence, if \( C \simeq A \oplus B \), and the rings are finite, it is obvious that \( |C^*| = |A^*||B^*| \).

We apply this to \( A = \mathbb{Z}_n \), \( B = \mathbb{Z}_m \), and \( C = \mathbb{Z}_{nm} \), assuming that \( n \) and \( m \) are relative primes. Since \( \mathbb{Z}_{nm} \simeq \mathbb{Z}_n \oplus \mathbb{Z}_m \), we conclude immediately that:

\[
\varphi(nm) = \varphi(n) \varphi(m).
\]

The next theorem shows that one can compute \( \varphi(n) \) in a straightforward fashion, as long as one knows the prime factors of \( n \):

**Theorem 4.5.8.** If \( n = \prod_{i=1}^{k} p_i^{e_i} \) is the prime factorization of \( n \) then

\[
\varphi(n) = n \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right).
\]

**Proof.** The lemma above shows that if \( n \) is a natural number with prime factorization \( n = \prod_{i=1}^{k} p_i^{e_i} \), then

\[
\varphi(n) = \prod_{i=1}^{k} \varphi(p_i^{e_i}).
\]

Now if \( p \) is a prime number and \( m \) is a natural number, one finds easily \( \varphi(p^m) \): the elements of \( \mathbb{Z}_p^m \) which are not invertible are just the elements of the ideal \( \langle p \rangle \) in \( \mathbb{Z}_p^m \). This ideal has exactly \( p^m/p = p^{m-1} \) elements (why?). Therefore \( \mathbb{Z}_p^* = \mathbb{Z}_p^m - \langle p \rangle \) has \( p^m - p^{m-1} = p^{m-1}(1 - \frac{1}{p}) \) elements. In other words:

\[
\varphi(p_i^{e_i}) = p_i^{e_i} - p_i^{e_i-1} = p_i^{e_i}(1 - \frac{1}{p_i}).
\]
From this it follows that:

\[
\varphi(n) = \prod_{i=1}^{k} \varphi(p_i^{e_i}) = \prod_{i=1}^{k} p_i^{e_i} \left( 1 - \frac{1}{p_i} \right) = n \prod_{i=1}^{k} \left( 1 - \frac{1}{p_i} \right).
\]

\[\square\]

**Example 4.5.9.**

The prime factors of 9000 are 2, 3 and 5. Therefore:

\[
\varphi(9000) = 9000 \left( 1 - \frac{1}{2} \right) \left( 1 - \frac{1}{3} \right) \left( 1 - \frac{1}{5} \right) = 2400.
\]

As another application of the Isomorphism Theorems for rings, we will now show that the fields \(\mathbb{Z}_p\) and \(\mathbb{Q}\) are, in some sense, the smallest fields with a given characteristic. In other words, we will show that any field necessarily contains a subfield isomorphic to either \(\mathbb{Z}_p\) or \(\mathbb{Q}\). In what follows, we will denote by \(K\) a field and denote the identity by 1.

**Definition 4.5.10.** We say that \(K\) is a **primitive field** if its does not contain any strictly smaller subfield (i.e., \(\neq K\)).

Obviously there exist primitive fields (e.g., \(\mathbb{Z}_2\)) and non-primitive fields (e.g., \(\mathbb{R}\)). Moreover, any field \(K\) has exactly one primitive subfield: the intersection of all the subfields of \(K\) is necessarily a primitive field primitive. One calls it the **primitive subfield** of \(K\).

**Example 4.5.11.**

Obviously, \(\mathbb{Q}\) is the primitive subfield of \(\mathbb{R}\) and of \(\mathbb{C}\). Similarly, \(\mathbb{Q}\) is also the primitive subfield of \(\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}\).

Our next result identifies all the possible primitive fields:

**Theorem 4.5.12.** Let \(m\) be the characteristic of \(K\), where \(m = 0\) or \(m = p\), for a prime number \(p\). Then:

(i) If \(m = 0\), the primitive subfield of \(K\) is isomorphic to \(\mathbb{Q}\);

(ii) Se \(m = p\), the primitive subfield of \(K\) is isomorphic to \(\mathbb{Z}_p\).
Proof. We give the proof in the case $m = p$, leaving the case $m = 0$ for the exercises.

Consider the homomorphism $\phi : \mathbb{Z} \to K$ defined by $\phi(n) = n1$. It is easy to check that any subfield of $K$ must contain 1 and, hence, also $\phi(\mathbb{Z})$. It was shown in Chapter ??, that $\phi(\mathbb{Z})$ is isomorphic to $\mathbb{Z}_p$, and so it is a field. Since $\phi(\mathbb{Z})$ is primitive, it must be the primitive subfield of $K$.

A field $K$ is always a vector space over its primitive subfield, which may have finite or infinite dimension. If $K$ is a finite field, its characteristic is necessary a prime $p > 0$, so that its primitive subfield $J$ has $p$ elements and is isomorphic to $\mathbb{Z}_p$. The dimension of $K$ as a vector space over $J$ is finite, for otherwise $K$ would be infinite. Therefore, there exists a natural number $n$ such that $K$ is isomorphic to the vector space $J^n$. Hence, we have proved:

**Theorem 4.5.13.** Any finite field $K$ has $p^n$ elements, where $p$ is a prime, namely the characteristic of $K$.

We know that there are finite fields with exactly $p$ elements, namely $\mathbb{Z}_p$. One can show that, if $p$ is a prime and $n$ is a natural number, there exist fields with $p^n$ elements, and all fields with $p^n$ elements are isomorphic. Hence, up to isomorphism, there exists exactly one field with $p^n$ elements, called the **Galois field** of order $p^n$, denoted $CG(p^n)$. We will not prove these statements now, but we remark that if $p(x) \in \mathbb{Z}_p[x]$ is an irreducible polynomial of degree $n$, then $K = \mathbb{Z}_p[x]/\langle p(x) \rangle$ is a field with $p^n$ elements. Therefore, it must be the Galois field $CG(p^n)$. In this way, one can reduce the existence of Galois fields to the existence of irreducible polynomials of arbitrary degree $n$ in $\mathbb{Z}_p$.

**Exercises.**

1. Prove the Isomorphism Theorems for rings.

2. Let $A$ be a ring with unit which has $n$ elements. Show that:

   (a) $A \simeq \mathbb{Z}_n$ as **rings** if and only if $A$ has characteristic $n$.

   (b) $A \simeq \mathbb{Z}_n$ as **rings** if and only if $(A, +) \simeq (\mathbb{Z}_n, +)$ as **groups**.

---

8After Évariste Galois (1811-1832). Galois, was a main contributor to a major mathematical discovery of the 19th Century, the Theory of Groups. Galois is also a tragic figure of the History of Mathematics, for he died at the young age of 21, in a duel because of a women of dubious reputation. We will discuss in Chapter 7 the theory of Galois.

9See also Exercise 8 in this section.
4.5. ISOMORPHISM THEOREMS FOR RINGS

3. The statement that $\mathbb{Z}_n \oplus \mathbb{Z}_m \cong \mathbb{Z}_{nm}$, if $\gcd(n, m) = 1$, expresses the Chinese Remainder Theorem in terms of isomorphisms of rings. How does one expresses the Fundamental Theorem of Arithmetic in terms of isomorphisms of rings?

4. Assume that $n, m \in \mathbb{N}$, $d = \gcd(n, m)$ and $k = \text{lcm}(n, m)$. Show that $\mathbb{Z}_n \oplus \mathbb{Z}_m \cong \mathbb{Z}_d \oplus \mathbb{Z}_k$.

5. Assuming that $n$ and $m$ are relatively prime, show that:
   
   (a) There exists exactly one isomorphism of rings $\phi: \mathbb{Z}_{mn} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$, and
   
   (b) The inverse isomorphism $\phi^{-1}: \mathbb{Z}_m \oplus \mathbb{Z}_n \to \mathbb{Z}_{mn}$ is given by $\phi^{-1}(x, y) = \phi_m(x) + \phi_n(y)$, where $\phi_k: \mathbb{Z}_k \to \mathbb{Z}_{mn}$ are injective homomorphisms such that $\phi_k(\pi_k(x)) = \pi_{mn}(a_k x)$. Find the integers $a_k$ and explain the relationship between $\phi^{-1}$ and the Chinese Remainder Theorem.

6. Consider homomorphisms of rings $\phi: \mathbb{Z}_n \to \mathbb{Z}_{210}$.

   (a) For which values of $n$ are there injective homomorphisms $\phi$?
   
   (b) For which values of $n$ are there surjective homomorphisms $\phi$?

7. Solve the equation $\varphi(m) = 6$, where $\varphi$ is the Euler function, using the following steps:

   (a) Show that the prime factors of $m$ must be 2, 3, or 7.
   
   (b) Show that if $7|m$, then $m = 7$ or $m = 14$.
   
   (c) Show that if 7 is not a factor of $m$, then $3|m$ and $9|m$.
   
   (d) Determine all solutions of of $\varphi(m) = 6$.

8. Assume that $K \subset L$ are fields, $u \in L$ is algebraic over $K$, and $m(x)$ is the minimal polynomial of $u$ in $K[x]$. Show that $K[u], K(u)$ and $K[x]/\langle m(x) \rangle$ are isomorphic fields.

9. Let $I \subset A$ be an ideal of $A$. Is it true that $A$ is necessarily isomorphic to the ring $I \oplus A/I$? On the other hand, show that if $A$ is isomorphic to $I \oplus J$, then $J$ is isomorphic to $A/I$.

10. Let $K$ be a field, and assume that $p(x) = q(x)d(x)$ holds in $K[x]$. Show that $K[x]/\langle p(x) \rangle$ is isomorphic to $K[x]/\langle q(x) \rangle \oplus K[x]/\langle d(x) \rangle$ if and only if $\gcd(q(x), d(x)) = 1$.

11. Consider $p(x) = (x^2 + x + 1)(x^3 + x + 1) \in \mathbb{Z}_2[x]$. How many invertible elements are there in $\mathbb{Z}_2[x]/\langle p(x) \rangle$? And in $\mathbb{Z}_2[x]/\langle p(x)^2 \rangle$?

12. Finish the proof of Theorem 4.5.12.
13. Let $K$ be a primitive field, let $L$ and $M$ be extensions of $K$, and let $\phi : L \to M$ be a non-zero homomorphism. Show that:

(a) For all $a \in K$: $\phi(a) = a$.

(b) If $p(x) \in K[x]$, $b \in L$ and $p(b) = 0$, then $p(\phi(b)) = 0$, i.e., $\phi$ takes roots of $p(x)$ into roots of $p(x)$.

(c) $\mathbb{Q}[x]/\langle x^3 - 2 \rangle$ is not isomorphic to $\mathbb{Q}[x]/\langle x^3 - 3 \rangle$.

14. Show that any ordered field is an extension of $\mathbb{Q}$, i.e., $\mathbb{Q}$ is the smallest ordered field.

15. Show that any complete ordered field is an extension of the reals, i.e., $\mathbb{R}$ is the smallest complete ordered field.

4.6 Free Groups, Generators and Relations

If $X$ is a subset of a group $G$, the subgroup generated by $X$ is the intersection of all subgroups of $G$ which contain $X$ and is denoted by $\langle X \rangle$. Similarly to the case of ideals in rings, if $X = \{x_1, x_2, \ldots, x_n\}$ is a finite set, we also write $\langle X \rangle = \langle x_1, x_2, \ldots, x_n \rangle$.

Example 4.6.1.

If $G = S_3$ then we see immediately that:

- $\langle (12) \rangle = \{I, (12)\}$
- $\langle (123) \rangle = \langle (321) \rangle = \{I, (123), (321)\}$

On the other hand, $\langle (12), (123) \rangle = S_3$. In fact, for any subset $X \subset S_3$ containing a transposition and a 3-cycle we have $\langle X \rangle = S_3$.

These examples make it clear that a subgroup can have many different sets of generators.

The set $X$ is called a generating set of the group $G$ if and only if $\langle X \rangle = G$. This condition is equivalent to say that any element of $G$ can be written, in multiplicative notation, as a product of positive and negative powers of elements of $X$:

$$g = x_1^{n_1} \cdots x_k^{n_k}, \quad x_i \in X, n_i \in \mathbb{Z}.$$ 

If the group $G$ has a finite generating set, i.e., if $G = \langle x_1, x_2, \ldots, x_k \rangle$, then $G$ is called a finitely generated group.
4.6. FREE GROUPS, GENERATORS AND RELATIONS

In the case of an abelian group we use the additive notation. So, for example, if \( G \) is a finitely generated abelian group with generators \( x_1, \ldots, x_k \), then we can write any element of \( G \) as:

\[
g = n_1 x_1 + n_2 x_2 + \cdots + n_k x_k, \quad n_i \in \mathbb{Z}.
\]

Examples 4.6.2.

1. A cyclic group \( G = \langle \alpha \rangle \) is, by definition, finitely generated with one generator. Any element \( g \in G \) is of the form \( g = \alpha^n \), possibly for multiple values of \( n \).

2. Any finite group is finitely generated, since we can always take \( X = G \) as a generating set.

3. The group \( G = \mathbb{Z} \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_4 \) is neither cyclic nor finite, but it is finitely generated: a generating set is \( X = \{(1,0,0),(0,1,0),(0,0,1)\} \).

4. The abelian group \( \mathbb{Z}^k := \oplus_{i=1}^k \mathbb{Z} \) is finitely generated. A generating set is given by the \( n \) vectors of the usual canonical basis of \( \mathbb{R}^k \), \( e_1, e_2, \cdots, e_k \), where \( e_i \) has all components zero but the \( i \)-th component which is 1. Any element of \( g \in \mathbb{Z}^k \) can be written uniquely as:

\[
g = \sum_{i=1}^k n_i e_i, \quad n_i \in \mathbb{Z}.
\]

One can also define the normal subgroup of \( G \) generated by a subset \( X \subset G \) as the intersection of all normal subgroups which contain \( X \). In the abelian case, of course, this is the same as the subgroup generated by \( X \), but it the non-abelian the two subgroups are, in general, distinct. For example, in \( G = S_3 \) if we let \( X = \{\sigma\} \), where \( \sigma \) is any transposition, then the subgroup generated by \( X \) is \( H = \{1, \sigma\} \), but the normal subgroup generated by \( X \) is \( S_3 \) itself.

If \( X \) is a generating set for a group \( G \), in general, there are multiple products of elements of \( X \) which yield the identity \( 1 \in G \). For example:

(a) For all \( x \in X \), we have \( xx^{-1} = 1 \);

(b) If \( G \) is cyclic of order \( m \), and \( X = \{x\} \) is a generator, then \( x^m = 1 \).

One calls expressions formed by products of elements of \( G \) which equal the identity a relation. There are trivial relations, as in example (a), which are consequences of the groups axioms, and non-trivial relations, as in example (b), which depend on the specific group \( G \) and generating set \( X \). We will make these concepts more precise later.
CHAPTER 4. QUOTIENTS AND ISOMORPHISMS

Many groups can be completely described, in a very succinct form, by specifying a set of generators $X$ and relations between these generators which we write (in multiplicative notation) as:

$$x_1^{c_1}x_2^{c_2} \cdots x_k^{c_k} = 1.$$

**Examples 4.6.3.**

1. The cyclic group $G$ of order $n$ is completely described by specifying the generator $X = \{\alpha\}$, and the relation $\alpha^n = 1$.

2. The group $S_3$ is generated by $X = \{\alpha, \delta\}$. Its multiplicative table can be obtained from the relations $\alpha^2 = 1$, $\delta^3 = 1$, and $\delta \alpha = \alpha \delta^{-1}$. The last relation can also be written as $\alpha \delta \alpha \delta = 1$.

3. The group $H_8$ is generated by $X = \{i, j\}$, and is completely described by the relations $i^2j^2 = i^4 = 1$ and $iji = j$.

4. The group $D_n$ of symmetries of a regular polygon of $n$ sides is generated by two elements $\sigma$ and a $\rho$ satisfying the relations

$$\sigma^2 = 1, \quad \rho^n = 1, \quad \sigma \rho \sigma \rho = 1.$$

The element $\rho$ represents a rotation by $2\pi/n$, while the element $\sigma$ represents a reflection on a symmetry axis of the polygon.

In order to make easier the comparison of two distinct groups $G$ and $H$, using a single set of generators $X$, we will also say that $G$ is generated by a set $X$ if there exists an injective map $\iota : X \rightarrow G$ such that $G$ is generated by the set $\iota(X)$, in the sense we used before. When the map $\iota$ is clear from the context, in order to simplify the notation, we will use the same symbol to denote both an element $x \in X$ and the corresponding element $\iota(x) \in G$.

**Example 4.6.4.**

According to these conventions, the groups $S_3$, $H_8$, and $\mathbb{Z} \oplus \mathbb{Z}$ are all generated by $X = \{x_1, x_2\}$.

Assume now that $G$ is generated by $X = \{x_1, x_2, \ldots, x_n\}$, and let $H$ be an arbitrary group. A moment of thought shows that:

- Any homomorphism $\phi : G \rightarrow H$ is uniquely determined on all elements of $G$, by the values that $\phi$ takes on the generators $x_k$, but
4.6. FREE GROUPS, GENERATORS AND RELATIONS

• The values $y_k = \phi(x_k)$ are not arbitrary, because whatever relations hold for $x_1, \ldots, x_n$ in $G$ they must also hold for the $y_1, \ldots, y_n$ in $H$.

Example 4.6.5.

Any group homomorphism $\phi : S_3 \to H$ is uniquely determined by its values $\rho = \phi(\alpha)$ and $\xi = \phi(\delta)$. However, the elements $\rho, \xi \in H$ cannot be arbitrary, and must satisfy the same relations that $\alpha$ and $\delta$ satisfy, namely: $\rho^2 = \xi^3 = 1$ and $\xi\rho\xi = 1$.

Still from an informal point of view, a group $G$ generated by $X$ is free from relations between its generators, if there always exist a homomorphism $\phi : G \to H$, whatever the values $\phi(x), x \in X$. We make these ideas more precise as follows:

Definition 4.6.6. For a set $X$, we call $F$ a free group (respectively, free abelian group) on $X$, if $F$ is a group (respectively, abelian group) and there exists a map $\iota : X \to F$ such that the following property holds: for any group (respectively, abelian group) $H$ and any map $\phi : X \to H$ there exists a unique group homomorphism $\tilde{\phi} : F \to H$ such that the following diagram commutes:

\[
\begin{array}{ccc}
X & \xrightarrow{\iota} & F \\
\downarrow\phi & & \downarrow\tilde{\phi} \\
H & \xrightarrow{\phi} & H
\end{array}
\]

Example 4.6.7.

If $X = \{x_1, \ldots, x_n\}$ is a finite set, we consider the group $\mathbb{Z}^n = \bigoplus_{i=1}^n \mathbb{Z}$ and the map $\iota : X \to \mathbb{Z}^n$ defined by $\iota(x_i) = e_i, i = 1, \ldots, n$. If $H$ is any abelian group and $\phi : X \to H$ is any map, then we have a group homomorphism $\tilde{\phi} : \mathbb{Z}^n \to H$ given by:

$$\tilde{\phi}(k_1, \ldots, k_n) = k_1\phi(x_1) + \cdots + k_n\phi(x_n).$$

This makes the following diagram commute:

\[
\begin{array}{ccc}
\{x_1, \ldots, x_n\} & \xrightarrow{\iota} & \mathbb{Z}^n \\
\downarrow\phi & & \downarrow\tilde{\phi} \\
H & \xrightarrow{\phi} & H
\end{array}
\]

It is easy to see that $\tilde{\phi} : \mathbb{Z}^n \to H$ is the only group homomorphism with this property. Hence, $\mathbb{Z}^n$ is a free abelian group on $X = \{x_1, \ldots, x_n\}$. 
CHAPTER 4. QUOTIENTS AND ISOMORPHISMS

We leave it as an exercise to check that the definition of free group $F$ on the set $X$ implies that $\iota : X \to F$ is injective and $F$ is generated by $X$. Notice that, if the group $H$ is also generated by $X$, so there exists a map $\phi : X \to H$ such that $\langle \phi(X) \rangle = H$, then the homomorphism $\tilde{\phi}$ is onto. In particular, by the First Isomorphism Theorem, the groups $F/N$ and $H$ are isomorphic, where $N$ denotes the kernel of $\tilde{\phi}$. This shows the relevance of the free group $F$ for the classification of groups: Any group generated by $X$ is a group quotient of the free group $F$.

Example 4.6.8.

Any abelian group generated by $X = \{x_1, \ldots, x_n\}$ is a group quotient of the group $\mathbb{Z}^n$. We will explore this remark later, in order to obtain a classification of all finitely generated abelian groups.

We will see now that for a set $X$, there exists (up to isomorphism) exactly one free non-abelian group and one free abelian group generated by $X$, except when $X = \{x_1\}$ consists of a single element (where both groups are abelian and isomorphic to $\mathbb{Z}$).

We start by showing that the free group on $X$ is unique, up to isomorphism.

Proposition 4.6.9. Let $F$ and $F'$ be free groups on a set $X$, relative to maps $\iota : X \to L$ and $\iota' : X \to L'$, respectively. Both in the abelian and in the non-abelian case, there exists a unique isomorphism $\psi : F \to F'$ making the following diagram commute:

$$
\begin{array}{ccc}
F & \xrightarrow{\psi} & F' \\
\downarrow & & \downarrow \\
X & \xrightarrow{\iota} & L \\
\downarrow & & \downarrow \\
F' & \xrightarrow{\iota'} & L' \\
\end{array}
$$

Proof. From Definition 4.6.6 if we let $H = F'$ and $\phi = \iota'$, we conclude that there exists a unique homomorphism $\psi : F \to F'$ which makes the diagram in the statement commutative. Similarly, interchanging the roles of $\iota$ and $\iota'$, we conclude that there exists a unique homomorphism $\psi' : F' \to F$ which
makes the following diagram commutative:

\[ \begin{array}{ccc}
F' & \xrightarrow{\psi'} & F' \\
\downarrow & & \downarrow \\
X & \xrightarrow{\psi} & F
\end{array} \]

It follows that the following diagrams are also commutative:

\[ \begin{array}{ccc}
F & \xrightarrow{\psi \circ \psi'} & F \\
\downarrow & & \downarrow \\
X & \xrightarrow{\psi \circ \psi'} & F
\end{array} \]

\[ \begin{array}{ccc}
F' & \xrightarrow{\psi' \circ \psi} & F' \\
\downarrow & & \downarrow \\
X & \xrightarrow{\psi' \circ \psi} & F'
\end{array} \]

Finally, observe that if in the last two diagrams we replace \( \psi \circ \psi' \) and \( \psi' \circ \psi \) by the identity homomorphisms we also obtain commutative diagrams. The uniqueness in the defining property of a free group, allows us to conclude that \( \psi \circ \psi' = \text{id}_F \) and \( \psi' \circ \psi = \text{id}_{F'} \). Hence, the homomorphism \( \psi \) has a unique inverse and so it is an isomorphism.

Let us turn now to the existence of a free abelian group generated by \( X \). We already know that this group exists if \( X \) is finite with \( n \) elements and it is isomorphic to \( \mathbb{Z}^n \). For infinite sets we need the following:

**Definition 4.6.10.** Let \( \{G_i\}_{i \in I} \) be a family of groups.

(i) The **direct product** of the \( G_i \)'s, denoted by \( \prod_{i \in I} G_i \), is the group with supporting set the cartesian product \( \prod_{i \in I} G_i \) of the groups and group operation defined as follows: if \( g = (g_i)_{i \in I} \) and \( h = (h_i)_{i \in I} \) are elements of \( \prod_{i \in I} G_i \), then \( gh \equiv (g_i h_i)_{i \in I} \in \prod_{i \in I} G_i \).

(ii) The **direct sum** of the \( G_i \)'s, denoted \( \bigoplus_{i \in I} G_i \), is the subgroup of the direct product \( \prod_{i \in I} G_i \) consisting of elements \( (g_i)_{i \in I} \in \prod_{i \in I} G_i \) where only a finite number of \( g_i \)'s is different from the identity (in \( G_i \)).

Obviously, when the set of indices is finite the direct sum and direct product coincide, and are the same as the product defined in Chapter 1. We leave the proof of the following proposition for the exercises:

---

10 See the Definition ?? in the Appendix.
CHAPTER 4. QUOTIENTS AND ISOMORPHISMS

Proposition 4.6.11. If $X$ is any set the direct sum $F = \bigoplus_{x \in X} \mathbb{Z}$ is a free abelian group generated by $X$ relative to the map $\iota : X \to F$ which to each $x_0 \in X$ associates the element $(g_x)_{x \in X} \in L$, with $g_{x_0} = 1$ and $g_x = 0$, if $x \neq x_0$.

Notice that the map $\iota$ in this Proposition is injective. For that reason it is called the canonical injection. If we identify each element $x_i \in X$ with its image $\iota(x_i) \in \bigoplus_{x \in X} \mathbb{Z}$ then $X$ can be seen as a subset of $\bigoplus_{x \in X} \mathbb{Z}$. Moreover, we can express every element $g \neq 0$ of $\bigoplus_{x \in X} \mathbb{Z}$ in additive notation in the form

$$g = n_1x_{i_1} + n_2x_{i_2} + \cdots + n_kx_{i_k},$$

where the indices $i_1, \ldots, i_k$ are all distinct and $n_1, \ldots, n_k$ are non-zero integers. This expression is unique, up to the order of the factors, and every expression of this form represents an element of $\bigoplus_{x \in X} \mathbb{Z}$.

Let us explain now how to construct the non-abelian free group on a set $X$ with more than one element. First we index the elements of $X$, so that $X = \{x_i : i \in I\}$, and then we form an “appropriate product” of all free groups $F_i$ in the sets $X_i = \{x_i\}$. The product that we need is called the free product of groups, and is discussed in the following proposition.

Proposition 4.6.12. Let $\{G_i\}_{i \in I}$ be a family of groups. There exists a group $\prod_{i \in I} G_i$ and group homomorphisms $\phi_i : G_i \to \prod_{i \in I} G_i$ with the following property: given any group $H$ and group homomorphisms $\psi_i : G_i \to H$, there exists a unique homomorphism of groups $\psi : \prod_{i \in I} G_i \to H$ such that the following diagram commutes for all $i \in I$:

$$
\begin{array}{ccc}
G_i & \xrightarrow{\phi_i} & \prod_{i \in I} G_i \\
\downarrow{\psi_i} & & \downarrow{\psi} \\
H & & H
\end{array}
$$

Proof. Let $\{G_i\}_{i \in I}$ be a family of groups. We call a word in the $G_i$’s any finite sequence $(g_1, \ldots, g_n)$, where each $g_k$ belongs to some $G_i$. The natural number $n$ is called the word length, and we conventioneer that the empty word, denoted 1, has length zero. A reduced word is any word $(g_1, \ldots, g_n)$ which satisfies the following two properties:

(a) none of the $g_k$ is the identity element of a group $G_i$;

(b) none of the successive terms in the word belong to the same group $G_i$;
We denote by $\prod_{i \in I}^* G_i$ the set of all reduced words, including the empty word $1$.

In the set $\prod_{i \in I}^* G_i$, we define a binary operation as follows. Let $g = (g_1, \ldots, g_n)$ and $h = (h_1, \ldots, h_m)$ be two reduced words with $n \leq m$. Let $0 \leq N \leq n$ be the smallest integer such that $N < k \leq n$ and both $g_k$ and $h_{n-k+1}$ belong to the same group $G_i$, while $g_k h_{n-k+1}$ is the identity in $G_i$. Then the product $gh$ is the reduced word defined by

$$
gh = \begin{cases} 
(g_1, \ldots, g_N, h_{n-N+1}, \ldots, h_m) & \text{if } N > 0 \text{ and } g_N, h_{n-N+1} \text{ do not belong to the same group}, \\
(g_1, \ldots, g_{N-1}, g_N h_{n-N+1}, h_{n-N+2}, \ldots, h_m) & \text{if } N > 0, \text{ and } g_N, h_{n-N+1} \text{ belong to the same group}, \\
(h_{n-N+1}, \ldots, h_m) & \text{if } N = 0 \text{ and } n < m, \\
1 & \text{if } N = 0 \text{ and } n = m.
\end{cases}
$$

By definition the product of a reduced word $g$ with the empty word $1$ is $g1 = 1g = g$. It is easy to check that this operation defines a group structure in $\prod_{i \in I}^* G_i$, with identity the empty word $1$, and where the inverse of the reduced word $g = (g_1, \ldots, g_n)$ is the reduced word $g^{-1} = (g_n^{-1}, \ldots, g_1^{-1})$.

Now let $\phi_i : G_i \to \prod_{i \in I}^* G_i$ be the map which to an element $g \in G_i$, with $g \neq e$ associates the reduced word $(g)$, and to $e$ associates $1$. Its is obvious that $\phi_i$ is a group homomorphism.

Finally, given any group $H$ and group homomorphisms $\psi_i : G_i \to H$, we define a group homomorphism $\psi : \prod_{i \in I}^* G_i \to H$ as follows: $\psi(1) = e$ (the identity in $H$) and $\psi(g_1, \ldots, g_n) := \psi_{i_1}(g_1) \cdots \psi_{i_n}(g_n)$, if $g_k \in G_{i_k}$. One checks immediately that this is the unique group homomorphism which makes the following diagram commute for all $i \in I$:

$$
\begin{array}{c}
G_i \\
\phi_i \\
\downarrow \psi_i \quad \downarrow \psi \\
\prod_{i \in I}^* G_i \\
\downarrow \psi \\
H
\end{array}
$$

The group $\prod_{i \in I}^* G_i$ is called the **free product** of the groups $G_i$. From now on, we will use the multiplicative notation to write the word $(g_1, \ldots, g_n)$ in the form $g_1 \cdots g_n$. 

\[\square\]
Our next example illustrates the difference between the free product and the direct product of groups.

**Example 4.6.13.**

Let $G = \{1, g\}$ and $H = \{1, h\}$ be two cyclic groups of order 2. An element in the free product $G \ast H$ can be written as a finite alternating sequence of products of $g$ and $h$. For example, the following expressions are elements in the free product:

$$g, h, gh, hg, ghg, hgh, ghgh, \ldots$$

Notice that $gh \neq hg$ and that both these elements have infinite order! By contrast, the direct product $G \times H$ is an abelian of order 4!

Given any set $X = \{x_i : i \in I\}$ we consider the free product $F \equiv \prod_{i \in I} F_i$ of all free groups $F_i$ in the sets $\{x_i\}$ (recall that $F_i \simeq \mathbb{Z}$). We have an injective map $\iota : X \to F$ which associates to the element $x_i$ the reduced word $(x_i)$.

**Proposition 4.6.14.** If $X = \{x_i : i \in I\}$, the free product $F = \prod_{i \in I} F_i$ is a free group generated by $X$ relative to the map $\iota : X \to F$.

**Proof.** We need to show that for any group $H$ and map $\phi : X \to H$ there exists a unique homomorphism of groups $\tilde{\phi} : F \to H$ so that we have a commutative diagram:

$$
\begin{array}{ccc}
X & \xrightarrow{\iota} & F \\
\downarrow{\phi} & & \downarrow{\tilde{\phi}} \\
& H & 
\end{array}
$$

The elements of $F$ are reduced words of the form

$$x_1^{k_1} \cdots x_n^{k_n},$$

where the $k_1, \ldots, k_n$ are non-zero integers. It is easy to see that $\tilde{\phi}$ must be defined by

$$\tilde{\phi}(x_1^{k_1}, \ldots, x_n^{k_n}) = \phi(x_1)^{k_1} \cdots \phi(x_n)^{k_n}.$$

Any element in the free group generated by the set $X = \{x_i : i \in I\}$ can be written in the reduced form

$$x_1^{k_1} \cdots x_n^{k_n},$$
and two elements written in this form can be multiplied in the obvious form, by juxtaposition and elimination, using multiplication in the groups $G_i$ and cancelation of units.

There is a relationship between the non-abelian and the abelian free groups in a set $X$. To see this, let us denote by $(G,G) \subset G$ the smallest subgroup of $G$ which contains all commutator elements:

$$(g,h) \equiv g^{-1}h^{-1}gh, \quad g,h \in G.$$ 

It is easy to see that $(G,G)$ is a normal subgroup of $G$ and that the quotient $G/(G,G)$ is abelian. We will study this commutator group in Chapter 5.

We leave the proof of the following proposition as an exercise:

**Proposition 4.6.15.** If $F$ is a free group on $X$ relative to $\iota : X \to F$, then $F/(F,F)$ is the free abelian group on $X$ relative to the map $\bar{\iota} : X \to F/(F,F)$ given by $\bar{\iota} = \pi \circ \iota$, where $\pi : F \to F/(F,F)$ is the quotient map.

The existence of free groups allows us to formalize the notion of relation, and to clarify the difference between trivial and non-trivial relations. Moreover, we can define precisely what we mean by complete set of relations. In what follows, $H$ will denote a group generated by $X \subset H$ and $F$ denotes the free group on $X$ relative to a map $\iota : X \to F$. If $H$ is abelian, we will assume naturally that $F$ is the abelian free group. The map $i : X \hookrightarrow H$ denote the canonical inclusion ($i(x) = x$). As we have already observed:

**Proposition 4.6.16.** There exists a surjective homomorphism $\phi : F \to H$.

This is summarized by the following diagram:

\[
\begin{array}{ccc}
X & \xrightarrow{\iota} & F \\
\downarrow{\iota} & & \downarrow{\phi} \\
H & \xrightarrow{i} & H
\end{array}
\]

**Definition 4.6.17.** A non-trivial relation of $H$ is any element $r \in \ker \phi$ distinct from the identity $1 \in F$.

Notice that $r$ takes the form $\iota(x_{i_1})^{n_1}\iota(x_{i_2})^{n_2} \cdots \iota(x_{i_k})^{n_k}$, where $x_j \in X$. Moreover, to say that $r \in \ker \phi$ is equivalent to say that

$$x_{i_1}^{n_1}x_{i_2}^{n_2} \cdots x_{i_k}^{n_k} = 1,$$

and so we will continue to say that the last identity is a “relation”. 

Given a set of relations \( R = \{ r_i \}_{i \in I} \), we say that a relation \( r \) is a consequence of the relations \( r_i \)'s, if \( r \) belongs to the normal subgroup of \( F \) generated by the \( r_i \)'s. Moreover, we say the set \( R \) is a complete set of relations if the kernel of \( \phi \) is the normal subgroup of \( F \) generated by \( R \).

Given a complete set of relations \( R = \{ r_i \}_{i \in I} \) the group \( H \) is completely determined, up to isomorphism, by the set of generators \( X \) and by the set \( R \). In fact, \( H \) is isomorphic to quotient of \( F \) by the normal subgroup generated by \( R \). One calls the pair \( (X, R) \) a presentation of the group \( H \).

Two groups with the same presentation are obviously isomorphic. On the other hand, in general, a group admits many distinct presentations.

**Example 4.6.18.**

Consider the group \( H = \mathbb{Z}_2 \oplus \mathbb{Z}_3 \). It is easy to see that this group is generated by the element \( x_1 = (1, 1) \). The free group \( F \) generated by \( X = \{ x_1 \} \) is isomorphic to \( \mathbb{Z} \), where we identify \( x_1 \) with the element 1. There exists a unique surjective group homomorphism \( \phi : \mathbb{Z} \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_3 \). This homomorphism maps 1 to \( (1, 1) \) and the element 6 belongs to its kernel. In fact, 6 generates the kernel, so \( R = \{ 6 \} \) is a complete set of relations for \( H \), and the pair \( \{ \{ x_1 \}, R \} \) is a presentation of \( H \). Using the multiplicative notation we can write:

\[
\{ \{ x_1 \}, x_1^6 = 1 \}
\]

Now note that \( x_1 = (1, 0) \) and \( x_2 = (0, 1) \) is an equally valid choice of generators for \( \mathbb{Z}_2 \oplus \mathbb{Z}_3 \). In this case, we should take instead the free group \( F = \mathbb{Z} \ast \mathbb{Z} \), and the kernel of \( \phi : \mathbb{Z} \ast \mathbb{Z} \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_3 \) is the normal group generated by \( x_1^2, x_2^3 \) and \( x_1 x_2 x_1^{-1} x_2^{-1} \). Therefore, we obtain as a different presentation for \( H \):

\[
\{ \{ x_1, x_2 \}, x_1^2 = 1, x_2^3 = 1, x_1 x_2 x_1^{-1} x_2^{-1} = 1 \}
\]

Unfortunately, the characterization of a group by presentations does not solve the classification problem for groups. In the 1950's, Adyan and Rabin\(^{11}\) have shown, independently, that there is no algorithm to decide if two group presentations represent or not isomorphic groups.

However, in the case of finitely generated abelian groups one can indeed use presentations to classify them, as we will now sketch. We will see in Chapter 6, in the context of the theory of modules, that this procedure can be generalized to obtain a classification of finitely generated modules over a principal ideal domain \( D \) (the case of abelian groups corresponds to taking \( D = \mathbb{Z} \)).

Let us assume then that \( A \) is a finitely generated abelian group, with a set of generators \( X = \{x_1, x_2, \cdots, x_n\} \subseteq A \). Let \( \mathbb{Z}^n = \oplus_{k=1}^{n} \mathbb{Z} \) be the free abelian group on \( X \), with \( \iota : X \to \mathbb{Z}^n \) given by \( \iota(x_k) = e_k \). The generators \( n_i \) of the kernel of \( \phi : \mathbb{Z}^n \to A \) take the form \( n_i = (r_{i1}, r_{i2}, \cdots, r_{in}) \), where \( r_{ik} \in \mathbb{Z} \), and correspond to relations of the form 
\[
    r_{i1}x_1 + r_{i2}x_2 + \cdots + r_{in}x_n = 0.
\]

We can assemble these generators into a matrix \( R \) of dimension \( m \times n \) with entries \( r_{ik} \in \mathbb{Z} \). Hence, in practical terms, a presentation of the finitely generated abelian group \( A \) is simply a matrix \( R \in M_{m,n}(\mathbb{Z}) \).

As we have discussed before, we can replace \( X \) by any other set of generators of \( A \), which has the effect of changing the matrix \( R \). We leave as an exercise to check that:

**Proposition 4.6.19.** Let \( A \) be an abelian group generated by the set \( X = \{x_1, x_2, \cdots, x_n\} \), and let \( R \in M_{m,n}(\mathbb{Z}) \) be the matrix with the corresponding list of relations. Then:

(i) If \( S = (s_{kj}) \) is an invertible matrix in \( M_n(\mathbb{Z}) \), then the elements \( y_k = \sum_{k=1}^{n} s_{kj}x_j \), are also generators of \( A \), and

(ii) The matrix \( RS^{-1} \) gives the list of relations for the generators \( \{y_1, \ldots, y_n\} \).

The rows \( l_1, l_2, \cdots, l_m \) of the matrix \( R \) are the generators of kernel of \( \phi \) in \( \mathbb{Z}^n \). It should now be clear that if \( P = (p_{kj}) \) is an invertible matrix in \( M_m(\mathbb{Z}) \), then the elements \( l_1, l_2, \cdots, l_m \) can be replaced by elements \( l'_1, l'_2, \cdots, l'_m \), where \( l'_k = \sum_{j=1}^{n} p_{kj}l_j \). This amounts to having a new presentation of \( A \) with matrix \( PR \). In other words, we see that:

**Proposition 4.6.20.** Let \( A \) be an abelian group generated by the set \( X = \{x_1, x_2, \cdots, x_n\} \), and let \( R \in M_{m,n}(\mathbb{Z}) \) be the matrix with the corresponding list of relations. If \( P \) and \( Q \) are any invertible matrices, respectively in \( M_m(\mathbb{Z}) \) and in \( M_n(\mathbb{Z}) \), then \( PRQ \) is also the matrix of a presentation of \( A \).

In particular, we can choose \( P \) and \( Q \) to perform the usual elementary operations of Gauss elimination on the rows and columns of \( R \). Note, however, we can only use invertible operations in \( M_n(\mathbb{Z}) \), which allows us to:

- Switch rows or columns,

- Add a row (or column) a multiple of another row (or column),

- Multiply a row (or column) by \(-1\).
Examples 4.6.21.

1. Let $A$ be an abelian group generated by $X = \{x_1, x_2\}$. Assume that these generators satisfy the relations $6x_1 - 6x_2 = 0$ and $12x_1 + 20x_2 = 0$. We can perform the following sequence of operations on the resulting matrix:

\[
\begin{pmatrix}
6 & -6 \\
12 & 20
\end{pmatrix} \rightarrow \begin{pmatrix}
6 & -6 \\
30 & 2
\end{pmatrix} \rightarrow \begin{pmatrix}
96 & 0 \\
30 & 2
\end{pmatrix} \rightarrow \begin{pmatrix}
96 & 0 \\
0 & 2
\end{pmatrix}.
\]

We conclude that $A$ is an abelian group which has a set of generators $y_1$ and $y_2$, satisfying $2y_1 = 0$ and $96y_2 = 0$. Hence, $A \cong \mathbb{Z}_2 \oplus \mathbb{Z}_{96}$. More precisely, there exists a surjective homomorphism $\phi : \mathbb{Z}^2 \rightarrow A$ with kernel $N = \langle 2 \rangle \oplus \langle 96 \rangle$, so it follows that

\[A \cong \mathbb{Z}^2/N \cong \mathbb{Z}/\langle 2 \rangle \oplus \mathbb{Z}/\langle 96 \rangle = \mathbb{Z}_2 \oplus \mathbb{Z}_{96}.
\]

2. The group $\mathbb{Z}_6 \oplus \mathbb{Z}_{16}$ has generators $x_1 = (1, 0)$ and $x_2 = (0, 1)$, and these satisfy the relations $6x_1 = 0$ and $16x_2 = 0$. Again, we can perform the following operations:

\[
\begin{pmatrix}
6 & 0 \\
0 & 16
\end{pmatrix} \rightarrow \begin{pmatrix}
6 & 0 \\
6 & 16
\end{pmatrix} \rightarrow \begin{pmatrix}
6 & -12 \\
6 & 4
\end{pmatrix} \rightarrow \\
\begin{pmatrix}
18 & -12 \\
2 & 4
\end{pmatrix} \rightarrow \begin{pmatrix}
0 & -48 \\
2 & 4
\end{pmatrix} \rightarrow \begin{pmatrix}
2 & 0 \\
0 & 48
\end{pmatrix}
\]

to conclude that:

\[\mathbb{Z}_6 \oplus \mathbb{Z}_{16} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_{48}.
\]

In both these examples, the first aim was to find the greatest common divisor of all entries of the matrix. In fact, we have:

**Proposition 4.6.22.** Let $R \in M_n(\mathbb{Z})$ and $d = \gcd(R)$ be the greatest common divisor of all entries of $R$. There exist invertible matrices $P, Q \in M_n(\mathbb{Z})$ such that the matrix $R' = PRQ$ has $R'_{11} = \gcd(R') = d$.

The proof of this proposition is not hard, but we will see in Chapter 6 a much more general result so we omit it.

Note that once we have obtained a matrix $R$ with an entry equal to $\gcd(R)$, then we can use it to eliminate all other entries in the same row and column. In the examples above, where we had $2 \times 2$ matrices, the elimination stops after this step. For matrices of dimension $n \times n$, with $n > 2$, we can move the gcd to upper left corner, as in the previous proposition, and eliminate all entries in the first row and column. One starts all over
with now a matrix \( R' \) of dimension \((n-1) \times (n-1)\), formed by the elements 
\[ r_{ij}, \quad i > 1 \text{ and } j > 1, \]
where there exist, possibly, non-zero elements. The gcd of the entries of \( R' \) is a multiple of the gcd of the entries of \( R \), so this process yields a sequence of integers \( d'_1, d'_2, \ldots, d'_n \).

**Example 4.6.23.**

We illustrate the algorithm with a \( 3 \times 3 \) matrix.

\[
\begin{pmatrix}
3 & 0 & 0 \\
9 & 6 & 12 \\
12 & 6 & 24 \\
\end{pmatrix}
\rightarrow
\begin{pmatrix}
3 & 0 & 0 \\
0 & 6 & 12 \\
0 & 6 & 24 \\
\end{pmatrix}
\rightarrow
\begin{pmatrix}
3 & 0 & 0 \\
0 & 6 & 12 \\
0 & 0 & 12 \\
\end{pmatrix}
\rightarrow
\begin{pmatrix}
3 & 0 & 0 \\
0 & 6 & 0 \\
0 & 0 & 12 \\
\end{pmatrix}
\]

After this elimination process, the homomorphism \( \phi : \mathbb{Z}^n \to A \) is given by
\[
\phi(k_1, k_2, \ldots, k_n) = k_1y_1 + k_2y_2 + \cdots + k_ny_n, \]
and has kernel the subgroup \( N = \langle d'_1 \rangle \oplus \cdots \oplus \langle d'_n \rangle \). Hence, we conclude that 
\[
A \simeq \mathbb{Z}^n / N \simeq \mathbb{Z}_{d'_1} \oplus \cdots \oplus \mathbb{Z}_{d'_n}.
\]

It is possible that \( 1 = d'_1 = d'_2 = \cdots = d'_k \), for some \( k \leq n \). The corresponding quotients \( \mathbb{Z}_{d'_i} \) are trivial and we can ignore them. We can also have \( 0 = d'_j = d'_{j+1} = \cdots = d'_n \) for some \( j \leq n \). The corresponding quotients are \( \mathbb{Z}_{d'_j} \simeq \mathbb{Z} \). We conclude that:

**Theorem 4.6.24** (Classification of finitely generated abelian groups). If \( A \) is a finitely generated abelian group, then

\[
A \simeq \mathbb{Z}_{d_1} \oplus \cdots \oplus \mathbb{Z}_{d_m} \oplus \mathbb{Z}^r,
\]

where \( 1 < d_1 | d_2 | \cdots | d_n \).

The integer \( r \) is the number of integers \( d'_j \) which are zero and is called the **characteristic** of the abelian group \( A \). The integers \( d_1, \ldots, d_n \) are called **invariant factors** or **torsion coefficients** of the abelian group \( A \). These integers characterize the abelian group up to isomorphism. For this reason, we say that they form a **complete set of invariants** of an abelian group.

In general, for any abelian group \( A \), the subset of \( A \) formed by all elements of finite order is a subgroup called the **torsion subgroup** of \( A \), which we will denote by \( \text{Tors}(A) \). If \( \text{Tors}(A) \) is trivial, the group is said to be **torsion free**, and if \( \text{Tors}(A) = A \), then we say that \( A \) is a **torsion group**. In any case, the quotient \( A / \text{Tors}(A) \) is always a torsion free group.
Example 4.6.25.

In Examples 4.6.21 both groups are torsion. The group $A$ has invariant factors 2 and 96, while the invariant factors of $\mathbb{Z}_6 \oplus \mathbb{Z}_{16}$ are 2 and 48.

Exercises.

1. Let $G$ be a group and $X \subseteq G$. Show that the intersection of all subgroups (respectively, normal subgroups) of $G$ which contain $X$ is the smallest subgroup (respectively, normal subgroup) of $G$ that contains $X$.

2. Show directly from the definition of a free abelian group $F$ on a set $X$, that the image $\iota(X)$ is a set generating $F$.

3. Prove Proposition 3.6.11

4. Verify that if $A_1$ and $A_2$ are free abelian groups generated by finite sets, then $A_1$ and $A_2$ are isomorphic if and only if they have the same characteristic.

5. Show that a group with two generators $a$ and $b$ and relations

\[ aba^{-1}b^{-1} = 1, \quad a^n = 1, \quad b^m = 1, \]

is isomorphic to the group $\mathbb{Z}_m \oplus \mathbb{Z}_m$.

6. Show that a group with two generators $\sigma$ and $\rho$ and relations

\[ \sigma^2 = 1, \quad \rho^n = 1, \quad \rho \sigma = \sigma \rho^{-1}, \]

is isomorphic to the group $D_n$ of symmetries of a regular polygon with $n$ sides.

7. Show that a group with two generators $a$ and $b$ and relations

\[ a^4 = 1, \quad a^2 b^2 = 1, \quad abab^{-1} = 1, \]

is isomorphic to the group $H_8$.

8. Let $G$ be a group with two generators $a$ and $b$ satisfying the relation $a^3 b^{-2} = 1$, and let $H$ be a group with two generators $x$ and $y$ satisfying the relation $xy x^{-1} y^{-1} = 1$. Show that these two groups are isomorphic.

9. Give an example of two non-isomorphic abelian groups $A_1$ and $A_2$ such that $\text{Tors}(A_1) \cong \text{Tors}(A_2)$, and $A_1 / \text{Tors}(A_1) \cong A_2 / \text{Tors}(A_2)$.

10. Find the invariant factors of $\mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_6$ and $\mathbb{Z}_{16} \oplus \mathbb{Z}_3$? Are these two groups isomorphic?
11. Consider the groups in Examples 4.6.21. In each case, there exist generators \( \{y_1, y_2\} \) such that \( n_1y_1 = n_2y_2 = 0 \), where \( n_1, n_2 \) are the invariant factors of the group. What is the relation between these generators and the “original” generators \( x_1 \) and \( x_2 \)? Find the group homomorphisms \( \phi \) which allow one to conclude that these two groups are isomorphic to \( \mathbb{Z}_2 \oplus \mathbb{Z}_{96} \) and \( \mathbb{Z}_2 \oplus \mathbb{Z}_{48} \), respectively.

12. Proof Proposition 4.6.15

13. Let \( \{G_i : i \in I\} \) be a family of non-trivial groups with \( \#I > 1 \). Show that the non-abelian free product \( \prod_{i \in I}^* G_i \) has elements of infinite order and that its center is trivial.

14. Let \( G \) and \( H \) be groups. Show that if some element \( g \neq 1 \) in the free product \( G \ast H \) has finite order, then \( g \) is conjugate to an element of \( G \) or of \( H \).

15. Give proofs for Propositions 4.6.19 and 4.6.20

16. In the proof of Theorem 4.6.24 we have assumed that the number of generators of the kernel of \( \phi \) (denoted \( m \)) equals the number of generators of the group \( A \) (denoted \( n \)). Does this assumption imply any loss of generality?
Chapter 5

Finite Groups

5.1 Groups of Transformations

In this Chapter we study one of the most classic themes of Algebra, the structure of finite groups. We know already many examples of finite groups, such as the finite cycles groups $\mathbb{Z}_n$, the symmetric groups $S_n$ or the dihedral groups $D_n$. These groups are very distinct, although one can certainly find some connections between them. For example, $D_n$ contains a normal subgroup isomorphic to $\mathbb{Z}_n$ (the subgroup of rotations), and $S_3 \cong D_3$. Our aim now is, precisely, to study the finite groups in a more systematic way, trying to make such connections more evident.

We have already studied some of the properties of the symmetric group $S_n$, the group of bijections of the set $\{1, \ldots, n\}$. This group has a central role in the study of finite groups, since as we will see shortly, Cayley’s Theorem states that any finite group is isomorphic to a subgroup of $S_n$. More generally, we call a group of transformations of a set $X$ any subgroup of the group $S_X$ of bijections of $X$. Often, it is helpful to represent an abstract group as a group of transformations, since some properties of the group become more intuitive and geometric. In fact, the notion of a group of transformations is so natural that historically it preceded the abstract notion of a group: the great mathematicians of the 19th Century that many fundamental discoveries in Group Theory, such as Galois and Lie, only worked with transformation groups and did not know of the abstract

---

\footnote{Although this is a classic theme, one of the greatest achievements of modern days mathematics was the classification of finite simple groups, a class of groups we will study later. This classification occupies several hundreds of pages of mathematical literature. For a brief account see, e.g., the excellent article by R. Solomon, “On finite simple groups and their classification”, Notices of the American Mathematical Society \textbf{42}, 231–239 (1995)}
notion of a group, which would only be formalized later in the beginning of
the 20th Century.

The passage from an abstract group to a group of transformations
is done through the notion of group action, whose formal definition is as
follows:

**Definition 5.1.1.** An action of a group $G$ on a set $X$ is a map $G \times X \to X$,
written $(g, x) \mapsto gx$, which satisfy the following properties:

(i) $\forall x \in X, \ ex = x$;

(ii) $\forall g_1, g_2 \in G, \ \forall x \in X, \ g_1(g_2x) = (g_1 \cdot g_2)x$.

A slightly different, but equivalent, point of view is the following. Assume
that a group $G$ acts on a set $X$. For each $g \in G$ define a transformation
$T(g) : X \to X$ by setting $T(g)(x) \equiv gx$. Then conditions (i) and (ii) in
the definition of an action are equivalent to, respectively,

(i') $T(e) = I$ (identity transformation);

(ii') $\forall g_1, g_2 \in G, \ T(g_1 \cdot g_2) = T(g_1) \circ T(g_2)$.

Conversely, given a transformation $T(g) : X \to X$, for each $g \in G$, satisfying
(i') and (ii'), one obtains an action of $G$ on $X$ by the formula $gx \equiv T(g)(x)$.
Note, also, that each transformation $T(g)$ is bijective.

Hence, the map $g \mapsto T(g)$ is a group homomorphism from $G$ to the
group $S_X$ of bijections of the set $X$. We will call $T$ the associated homomorphism
of action. Therefore, an action of $G$ on $X$ realizes $G$ as a
group of transformations of $X$. An action is called effective if the associated homomorphism $T$ is injective, i.e., if the kernel of the homomorphism $T : G \to S_X$ is $\{e\}$. The kernel of the homomorphism $T$ is called the kernel of the action.

When $G$ acts in two distinct sets $X_1$ and $X_2$, with associated homomorphisms $T_1$ and $T_2$, we call the actions equivalent if there exists a bijection $\phi : X_1 \to X_2$ such that

\begin{equation}
\phi \circ T_1(g) = T_2(g) \circ \phi, \quad \forall g \in G.
\end{equation}

![Figure 5.1.2: Equivalent actions.](image)

This equation can also be expressed by the commutativity of the diagram:

$$
\begin{array}{ccc}
X_1 & \xrightarrow{\phi} & X_2 \\
\downarrow T_1(g) & & \downarrow T_2(g) \\
X_1 & \xrightarrow{\phi} & X_2
\end{array}
$$

Examples 5.1.2.

1. Consider the group $O(n)$ consisting of $n \times n$ orthogonal matrices, i.e., $AA^T = I$. There is an obvious action of $O(n)$ on $\mathbb{R}^n$ given by the action of a matrix on a vector: $(A, x) \mapsto Ax$. This is the usual way we view $O(n)$ as a group of transformations of $\mathbb{R}^n$. We saw in Chapter ?? that for each $g \in O(n)$ the transformation $T(g)$ is an isometry of $\mathbb{R}^n$ which fixes the origin. This action is effective (why?).

2. Similarly, consider the Euclidean group $E(n)$ formed by all pairs $(A, b)$, where $A \in O(n)$ and $b \in \mathbb{R}^n$, with binary operation $(A_1, b_1) \cdot (A_2, b_2) \equiv$
(A_1, A_2, b_2 + b_1). The group E(n) acts on \( \mathbb{R}^n \) by setting \(((A, b), x) \mapsto Ax + b\). We saw in Chapter ?? that the transformations \( T(g) \), where \( g \in E(n) \), are isometries of \( \mathbb{R}^n \) and that any isometry of \( \mathbb{R}^n \) is realized by a \( T(g) \). This is also an effective action.

3. A group \( G \) acts on itself by LEFT TRANSLATIONS \( : G \times G \to G, (g, x) \mapsto g \cdot x \). This action is effective (why?). We can also define an action of \( G \) on itself by RIGHT TRANSLATIONS by setting \( (g, x) \mapsto xg^{-1} \). These actions are equivalent: an equivalence is provided by the inverse map \( \phi : G \to G, x \mapsto x^{-1} \).

4. Another action of \( G \) on itself is the ACTION BY CONJUGATION, which is defined by:

\[(5.1.2) \quad (g, x) \mapsto gx \equiv gxg^{-1}, \quad g, x \in G.\]

It is easy to check that conditions (i) and (ii) in Definition 5.1.1 are satisfied and that they can be written as:

\[e \cdot x = x, \quad g_1(g_2 \cdot x) = g_1 \cdot g_2 x.\]

We leave it as an exercise to check that the kernel of this action is the center \( C(G) \) of the group \( G \).

5. Let \( G \) be a group and \( H \subset G \) a subgroup. Then \( G \) acts on the set of left cosets \( G/H \) by setting \( (g, xH) \mapsto (g \cdot x)H \). An element \( g \in G \) belongs to the kernel of this action if and only if:

\[(g \cdot x)H = xH, \quad \forall x \in G \iff g \in xHx^{-1}, \forall x \in G.\]

Hence the kernel of this action is

\[\bigcap_{x \in G} xHx^{-1}.\]

We leave it as an exercise to check that this is the largest normal subgroup of \( G \) which is contained in \( H \). Hence, the action is effective if and only if there is no non-trivial subgroup of \( H \) normal in \( G \).

We will explore the actions of a group on itself, as in the previous examples, to obtain information about the structure of the group. A simple application of this idea yields the following:

**Theorem 5.1.3** (Cayley). Let \( G \) be a finite group of order \( n \). Then \( G \) is isomorphic to a subgroup of \( S_n \).

**Proof.** Consider the action of \( G \) on itself by left translations. Since this action is effective, the associated homomorphism

\[T : G \to S_G \simeq S_n\]

is injective. \( \Box \)
5.1. GROUPS OF TRANSFORMATIONS

When $G$ acts on a set $X$ we obtain a partition of $X$ as follows. One defines an equivalence relation $\sim$ on $X$ where $x \sim y$ if there exists an element $g \in G$ such that $x = gy$. An equivalence class of $\sim$ is called a $G$-orbit. Hence, we obtain a partition of $X$ into $G$-orbits, where the $G$-orbit containing the element $x \in X$ is

$$O_x \equiv \{gx : g \in G\}.$$ 

The set of $G$-orbits is denoted by $X/G$. When $X$ is a finite set there exists a finite number of orbits $O_1, \ldots, O_n$, and each orbit has a finite number of elements. Therefore, counting elements, we obtain a class equation:

$$|X| = |O_1| + \cdots + |O_n|.$$ 

(5.1.3)

An action is called TRANSITIVE if it has only one orbit, namely $X$.

![Figure 5.1.3: A $G$-orbit of $x$.](image)

Examples 5.1.4.

1. The orbits of the action of the group $O(n)$ on $\mathbb{R}^n$ are the spheres $S_r = \{|x| = r : x \in \mathbb{R}^n\}$ ($r > 0$) and the origin $\{0\}$.

2. The action of $G$ on itself by (right or left) translations has only one orbit, hence it is transitive.

3. Let $H \subset G$ be a subgroup. The $H$-orbits of the action of $H$ on $G$ by left (respectively, right) translations are the right (respectively, left) cosets of $H$. In Chapter ?? we saw that the orbits of this action all have the same cardinal and that from this fact, and the class equation (5.1.3), it follows immediately Lagrange’s Theorem.
4. Consider the action of $G$ on itself by conjugation. The $G$-orbits are usually called conjugacy classes. We will denote the conjugacy class containing the element $x \in G$ by $Gx$. Two elements $x$ and $y$ are said to be conjugate if they belong to the same conjugacy class, i.e., if $y = gxg^{-1}$, for some $g \in G$. An element $x$ belongs to the center $C(G)$ if and only if $Gx = \{x\}$, so the center $C(G)$ is the union of all conjugacy classes containing a single element.

** Definition 5.1.5. ** If $G$ acts on $X$, the isotropy subgroup or stabilizer group of an element $x \in X$ is the subgroup $G_x \subset G$ given by

$$(5.1.4) \quad G_x = \{g \in G : gx = x\}.$$ 

When two elements $x, y \in X$ belong to the same orbit, the corresponding isotropy subgroups are conjugate. In fact, if $y = gx$ for some $g \in G$, then

$$h \in G_x \iff hx = x, \iff hg^{-1}y = g^{-1}y, \iff ghg^{-1}y = y \iff ghg^{-1} \in G_y.$$ 

Hence, the isotropy subgroups of $x$ and of $y = gx$ are related by

$$G_y = gG_xg^{-1}.$$ 

Moreover, the orbits are determined by the isotropy subgroups, as shown by the following proposition:

** Proposition 5.1.6. ** Assume a group $G$ acts on a set $X$. For any $x \in X$, the map $\phi : G/G_x \to O_x$, $hG_x \mapsto hx$, gives an equivalence between the $G$-action on $G/G_x$ and the $G$-action on the orbit through $x$.

We leave the proof for the exercises. In particular, for an action on a finite set we conclude that:

** Corollary 5.1.7. ** If $G$ acts on a finite set $X$ with orbits $O_{x_1}, \ldots, O_{x_n}$ then

$$(5.1.5) \quad |X| = \sum_{i=1}^n |G/G_{x_i}| = \sum_{i=1}^n [G : G_{x_i}].$$ 

The actions that we have described are sometimes called left actions, because they satisfy (ii) in Definition 5.1.1. One can also consider right actions where one writes the action as $X \times G \to X$, $(x, g) \mapsto xg$, and where (ii) is replaced by

$$\forall g_1, g_2 \in G, \forall x \in X, (xg_1)g_2 = x(g_1 \cdot g_2).$$
Unless otherwise mentioned, we will always consider left actions. You should check how to modify the examples in this section in order to obtain right actions.

**Exercises.**

1. Show that the kernel of the action by conjugation of \( G \) on itself is the center \( C(G) \).

2. Let \( H \) be a subgroup of \( G \). Show that \( \bigcap_{x \in G} xHx^{-1} \) is the largest normal subgroup of \( G \) which is contained in \( H \).

3. Give a proof of Proposition 5.1.6.

4. An action of a group \( G \) on a set \( X \) is called a **free action** if any \( e \neq g \in G \) acts without fixed points, i.e., if the isotropy subgroups \( G_x \) are trivial for all \( x \in X \). Show that a free action is effective and determined which of the actions in Examples 5.1.2 are free.

5. One says that a group \( G \) acts by automorphisms on a group \( K \) if there exists an action of \( G \) on \( K \) such that, for each \( g \in G \), the map \( k \mapsto gk \) is an automorphism of \( K \). Assuming that \( G \) acts by automorphisms on \( K \), show that:
   
   (a) the binary operation \( (g_1,k_1)*(g_2,k_2) = (g_1g_2,k_1(g_1k_2)) \) defines a group \((G \ltimes K,*)\), called the **semidirect product** of \( G \) by \( K \), which is denoted by \( G \ltimes K \);
   
   (b) the maps \( G \to G \ltimes K : g \mapsto (g,e) \) \( K \to G \ltimes K : k \mapsto (e,k) \) are monomorphisms. The image of second monomorphism is a normal subgroup of \( G \ltimes K \).

6. Consider the action of \( O(n) \) on \((\mathbb{R}^n,+)\). Show that this action is by automorphisms, and describe the semidirect product \( O(n) \ltimes \mathbb{R}^n \).

7. Determine the partition of the symmetric group \( S_n \) into conjugacy classes. (Hint: Consider first the case \( n = 3 \).)

### 5.2 The Sylow Theorems

If \( G \) is a finite group, then the order of a subgroup divides the order of the group. When \( G \) is a cyclic group, then for each divisor \( d \) of \( |G| \) there exists exactly one subgroup of \( G \) of order \( d \). In general, for a finite group \( G \), it is natural to ask:
• Given a factor $d$ of $|G|$, is there a subgroup of $G$ of order $d$?

In this section we will explore the action by conjugation of the group $G$ on itself, and the corresponding class equation (5.1.3), to obtain some answers to this question.

Let $x \in G$. The isotropy subgroup of $x$ for the action by conjugation is precisely:

$$\{g \in G : g \cdot x \cdot g^{-1} = x\} = \{g \in G : g \cdot x = x \cdot g\},$$

i.e., the set of all elements of $G$ which commute with $x$. This group is usually called the centralizer of $x$, and will be denoted by $C_G(x)$. Recall that $G^G_x$ denotes the $G$-orbit of $x$, i.e., the conjugacy class which contains $x$. By Proposition 5.1.6 $G^G_x$ is isomorphic to $G/C_G(x)$ and we conclude that:

$$|G^G_x| = [G : C_G(x)].$$

Moreover, it follows that:

**Proposition 5.2.1.** For any finite group $G$ one has:

$$|G| = |C(G)| + \sum_{i=1}^{n} [G : C_G(x_i)], \tag{5.2.1}$$

where $x_1, \ldots, x_n$ are representatives of each of the conjugacy classes of $G$ with more than one element.

**Proof.** As we have noted before, the center $C(G)$ is the union of all the conjugacy classes of $G$ with exactly one element. Equation (5.2.1) now follows from the class equation (5.1.5). \qed

As we shall see now, formula (5.2.1) is specially useful to exclude the existence of subgroups of certain orders. For example, as a first elementary application we show the existence of a family of groups whose centers are non-trivial.

**Proposition 5.2.2.** If $|G| = p^m$, where $p$ is a prime, then the center of $G$ has order $p^k$, where $k \geq 1$.

**Proof.** By Lagrange’s Theorem the order of the center of $C(G)$ divides the order of $G$. Hence, from the class equation (5.2.1), one obtains

$$p^m = p^k + \sum_{i=1}^{n} [G : C_G(x_i)].$$
Since \( x_i \) does not belong to the center, \( C_G(x_i) \neq G \) and we conclude that \( [G : C_G(x_i)] = p^{m_i} \), where \( m_i \geq 1 \). Hence:

\[
p^k = p^m - \sum_{i=1}^{n} p^{m_i}, \quad m, m_i \geq 1
\]

and we must have \( k \geq 1 \).

**Corollary 5.2.3.** Every group \( G \) with \( |G| = p^2 \) is abelian.

**Proof.** The previous proposition implies that \( |C(G)| = p \) or \( p^2 \). We leave it as an exercise to check that the case \( |C(G)| = p \) is not possible.

The next result gives a partial answer to the question raised at the beginning of this section in the case of abelian groups.

**Theorem 5.2.4 (Cauchy).** If \( G \) is a finite abelian group and \( p \) is a prime factor of \( |G| \), then \( G \) has an element \( g \) of order \( p \).

**Proof.** We use induction on the order \( |G| \) of \( G \). If \( |G| = p \) then the result is obvious. Assume that \( |G| > p \) and assume the result holds for all abelian groups of order less than \( |G| \). Choose an element \( e \neq a \in G \). Two cases are possible:

(i) \( a \) is an element of order divisible by \( p \). In this case, the cyclic group \( \langle a \rangle \) has an element \( g \) of order \( p \), so the theorem holds.

(ii) \( p \) does not divide the order of \( a \). In this case, the group \( G/\langle a \rangle \) has order divisible by \( p \). By the induction hypothesis, this group has an element \( b\langle a \rangle \) of order \( p \). The order \( s \) of \( b \) is divisible by \( p \), since \( \langle a \rangle = b^s\langle a \rangle = (b\langle a \rangle)^s \). Hence, the cyclic subgroup \( \langle b \rangle \) contains an element \( g \) of order \( p \).

Cauchy’s Theorem holds for finite abelian groups. We will see later that these groups can be classified. This classification will clarify completely what are the possible subgroups of a finite abelian group. We now turn to non-abelian groups for which we have the following generalization of Cauchy’s Theorem:

---

The results that follow are due to the Norwegian mathematician Ludvig Sylow (1832-1918) and appeared in the paper “Théorèmes sur les groupes de substitutions”, *Math. Ann.*, 5 (1872).
CHAPTER 5. FINITE GROUPS

Theorem 5.2.5 (Sylow I). Let $G$ be a finite group. If a prime power $p^k$ is a factor of the order of $|G|$, then there exists a subgroup $H$ of $G$ of order $p^k$.

Proof. We shall use again induce on the order $|G|$. Again, starting from the class equation (5.2.1):

$$|G| = |C(G)| + \sum_{i=1}^{n} [G : C_G(x_i)].$$

we observe that:

(i) If $p \nmid |C(G)|$, then for some $i \in \{1, \ldots, n\}$ we have that $p \nmid [G : C_G(x_i)]$. It follows that $C_G(x_i)$ is a subgroup of $G$ whose order is less than $|G|$ and divisible by $p^k$. By the induction hypothesis, there exists a subgroup $H$ of $C_G(x_i)$ of order $p^k$.

(ii) If $p \mid |C(G)|$, by Cauchy’s Theorem there exists an element $g \in C(G)$ of order $p$. If $k = 1$ we are done. If not, the subgroup $\langle g \rangle$ is normal in $G$ and the group $G/\langle g \rangle$ has order smaller than $|G|$ and his divisible by $p^{k-1}$. By induction, $G/\langle g \rangle$ contains a subgroup of order $p^{k-1}$. This subgroup takes the form $H/\langle g \rangle$, where $H$ is a subgroup of $G$, and we have:

$$|H| = [H : \langle g \rangle] |\langle g \rangle| = p^{k-1}p = p^k.$$

The previous theorem motivates the following definitions:

Definition 5.2.6. A group of order $p^k$ is called a $p$-GROUP (OF EXPONENT $k$). A $p$-subgroup $H \subset G$ where the exponent $k$ is maximal is called a SYLOW $p$-SUBGROUP.

The Sylow $p$-subgroups of a group $G$ are, in same sense, the analogues of the subgroups of a cyclic group, as shown by the following theorem:

Theorem 5.2.7 (Sylow II). Let $G$ be a finite group and assume $p \mid |G|$. Then:

(i) The Sylow $p$-subgroups of $G$ are unique up to conjugation.

(ii) The number of Sylow $p$-subgroups of $G$ is a divisor of the index of any Sylow $p$-subgroup and equals $\equiv 1 \pmod{p}$.

(iii) Any subgroup of $G$ of order $p^k$ is a subgroup of some Sylow $p$-subgroup of $G$. 

\qed
5.2. THE SYLOW THEOREMS

The proof of the first Sylow’s Theorem was based on the action of \( G \) on itself by conjugation. For the proof of the second Sylow’s Theorem we will use the action of \( G \) by conjugation of the set of its subgroups: if \( H \subset G \) is a subgroup, then \( gHg^{-1}, \ g \in G \), is a subgroup of \( G \) and \( |gHg^{-1}| = |H| \). For this action, the isotropy subgroup of \( H \subset G \) is precisely:

\[
N_G(H) \equiv \{ g \in G : gHg^{-1} = H \}.
\]

This subgroup is usually called the normalizer of \( H \) in \( G \). Note that \( H \) is normal in \( N_G(H) \) and, in fact, we leave it as an exercise to check that \( N_G(H) \) is the largest subgroup of \( G \) containing \( H \) as a normal subgroup.

We can also restrict this action of \( G \) to an action on the set \( \Pi \) of Sylow \( p \)-subgroups of \( G \).

**Lemma 5.2.8.** Let \( P \) be a Sylow \( p \)-subgroup of \( G \), and \( H \subset N_G(P) \) a subgroup of order \( p^k \). Then \( H \subset P \).

**Proof of Lemma 5.2.8.** Since \( P \) is a normal subgroup of \( N_G(P) \), we have a homomorphism \( \pi : N_G(P) \rightarrow N_G(P)/P \). Since \( H \) is a subgroup of \( N_G(P) \), it follows that \( \pi(H) \) is a subgroup of \( N_G(P)/P \) of order a power of \( p \). Since \( P \) is a Sylow \( p \)-subgroup of \( G \), it is also a Sylow \( p \)-subgroup of \( N_G(P) \) and we conclude that \( p \nmid |N_G(P)/P| \). But then we must have \( \pi(H) = \{e\} \), in other words \( H \subset P \).

**Proof of Sylow II.** Consider the action by conjugation of \( G \) on the set \( \Pi \) of Sylow \( p \)-subgroups of \( G \). Denote by \( O_P \) the orbit of a Sylow \( p \)-subgroup \( P \). Then:

1. \(|O_P| \equiv 1 \pmod{p} \): Consider the action of the group \( P \) on \( O_P \), obtained by restriction of the action of \( G \) (note that the \( G \)-action on \( O_P \) is transitive by definition, but not the \( P \)-action on \( O_P \)). The \( P \)-orbits which do not contain \( P \) have more than one element, since if \( \{\tilde{P}\} \) is a \( P \)-orbit distinct from \( P \), then \( \tilde{P} \) is a Sylow \( p \)-subgroup distinct from \( P \) and \( P \subset N_G(\tilde{P}) \), contradicting Lemma 5.2.8. On the other hand, all the \( P \)-orbits have cardinality a power of \( p \). Hence, \(|O_P| = 1 + \sum_i p^{k_i} \).

2. \( O_P = \Pi \): Assume that \( \tilde{P} \in \Pi - O_P \). The same argument as above applied to the action of \( \tilde{P} \) on \( O_P \), shows that \(|O_P| \equiv 0 \pmod{p} \), which contradicts (1).

Part (i) of Sylow II is equivalent to (2).

On the other hand, to show that (ii) holds, we note that (2) implies that

\[
|\Pi| = |G/N_G(P)| = [G : N_G(P)].
\]
Since $P \subset N_G(P) \subset G$, we conclude that:

$$[G : P] = [G : N_G(P)][N_G(P) : P],$$

so the number of $p$-subgroups is a divisor of $[G : P]$. By (a) and (b), we also have $|\Pi| \equiv 1 \pmod{p}$.

Finally, to show that (iii) holds, note that if $H \subset G$ is a subgroup of order $p^k$, then the orbits of the action by conjugation of $H$ on $\Pi$ have cardinality a power of $p$. But $|\Pi| \equiv 1 \pmod{p}$, so at least one of the orbits is of the form $\{\tilde{P}\}$. This means that $H \subset N_G(\tilde{P})$. By Lemma 5.2.8, $H \subset \tilde{P}$, as claimed in (iii).

Examples 5.2.9.

1. Consider the symmetric group $S_3 = \{I, \alpha, \beta, \gamma, \delta, \varepsilon\}$. Since $|S_3| = 6$, by Sylow I there exist Sylow $p$-subgroups of orders 2 and 3.

By Sylow II, the number of Sylow 3-subgroups must be 1 (mod 3) and a divisor of 2. Hence, there exists 1 subgroup of order 3. Obviously, this subgroup of $S_3$ is $P = A_3 = \{I, \delta, \varepsilon\}$.

Similarly, the number of Sylow 2-subgroups must be 1 (mod 2) and a divisor of 3. Hence, we can have either 1 or 3 subgroups of order 2. Obviously, there are 3 subgroups of order 2:

$P_1 = \{I, \alpha\}, \quad P_2 = \{I, \beta\}, \quad P_3 = \{I, \gamma\}$.

It is easy to check that these subgroups are all conjugate:

$P_1 = \delta P_2 \delta^{-1} = \varepsilon P_3 \varepsilon^{-1}$.

2. The subgroup $A_4$ of $S_4$ formed by the even permutations has order 12. By Sylow I, $A_4$ has Sylow $p$-subgroups of orders 3 and 4.

By Sylow II, the number of Sylow 3-subgroups can be 1 or 4. It is easy to find the following 4 Sylow 3-subgroups:

$P_1 = \{I, (123), (321)\}, \quad P_2 = \{I, (124), (421)\}$

$P_3 = \{I, (134), (431)\}, \quad P_4 = \{I, (234), (432)\}$.

We leave it as an exercise to check that all these subgroups are conjugate and to determine the Sylow 2-subgroups (which have order 4).

3. Let $G$ be a group of order $pq$, where $p < q$ are prime numbers. Assume that $p \nmid (q - 1)$. Then we claim that $G \simeq \mathbb{Z}_{pq}$.
To see this, observe that the number of Sylow p-subgroups of $G$ is $1 + kp$ and divides $pq$, so it is must be 1. On the other hand, the number of q-Sylow subgroups is $1 + lq$ and divides $pq$, so it must be 1. It follows that $G$ has two normal subgroups $H_1$, $H_2 \subset G$ of order $p$ and $q$, respectively. Obviously, $H_1 \cap H_2 = \{e\}$ and $G = H_1 H_2$, so we conclude that $G \simeq H_1 \oplus H_2 \simeq \mathbb{Z}_p \oplus \mathbb{Z}_q \simeq \mathbb{Z}_{pq}$. For example, any group of order 15 is isomorphic to $\mathbb{Z}_{15}$.

One can also show that when $p < q$ are prime numbers and $p \mid (q - 1)$ there are exactly two groups of order $pq$: the abelian group $\mathbb{Z}_{pq}$ and a non-abelian group with presentation:

$$\{a, b \mid a^p = 1, b^q = 1, ab = b^s a\},$$

where $s \in \mathbb{N}$ satisfies $s \neq 1 \pmod{q}$ and $s^p \equiv 1 \pmod{q}$ (see exercises).

Using the techniques developed so far, it is possible to classify all groups of order less or equal to 15, up to isomorphism. The list is as follows:

<table>
<thead>
<tr>
<th>Order</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>${e}$</td>
</tr>
<tr>
<td>2</td>
<td>$\mathbb{Z}_2$</td>
</tr>
<tr>
<td>3</td>
<td>$\mathbb{Z}_3$</td>
</tr>
<tr>
<td>4</td>
<td>$\mathbb{Z}_4, \mathbb{Z}_2 \oplus \mathbb{Z}_2$</td>
</tr>
<tr>
<td>5</td>
<td>$\mathbb{Z}_5$</td>
</tr>
<tr>
<td>6</td>
<td>$\mathbb{Z}_6, D_3$</td>
</tr>
<tr>
<td>7</td>
<td>$\mathbb{Z}_7$</td>
</tr>
<tr>
<td>8</td>
<td>$\mathbb{Z}_8, \mathbb{Z}_4 \oplus \mathbb{Z}_2, \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2, H_8, D_4$</td>
</tr>
<tr>
<td>9</td>
<td>$\mathbb{Z}_9, \mathbb{Z}_3 \oplus \mathbb{Z}_3$</td>
</tr>
<tr>
<td>10</td>
<td>$\mathbb{Z}_{10}, D_5$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>$\mathbb{Z}_{11}$</td>
</tr>
<tr>
<td>12</td>
<td>$\mathbb{Z}_{12}, \mathbb{Z}_2 \oplus \mathbb{Z}_6, A_4, D_6, E$</td>
</tr>
<tr>
<td>13</td>
<td>$\mathbb{Z}_{13}$</td>
</tr>
<tr>
<td>14</td>
<td>$\mathbb{Z}_{14}, D_7$</td>
</tr>
<tr>
<td>15</td>
<td>$\mathbb{Z}_{15}$</td>
</tr>
</tbody>
</table>

In this list, the group $E$ is a group of order 12 with presentation:

$$\{\{a, b\}, a^6 = 1, a^3 b^2 = 1, abab^{-1} = 1\}.$$ 

By contrast, there are 14 distinct groups of order 16, 51 distinct groups of order 32, etc. Although a finite group of order $n$ is a subgroup of $S_n$, there is no known formula for the number of distinct groups of order $n$. 
Exercises.

1. Show that if $G$ is a finite group and $|G| = p^2$, then $G$ is isomorphic to $\mathbb{Z}_{p^2}$ or to $\mathbb{Z}_p \oplus \mathbb{Z}_p$.

2. Show that, if $G$ is a finite abelian where all elements, excepting the identity $e$, have order $p$, then $|G| = p^n$ and $G \cong \mathbb{Z}_p \oplus \cdots \oplus \mathbb{Z}_p$.

3. Classify the finite groups of order $\leq 7$.

4. Given an example of a group $G$ where the converse to Lagrange’s Theorem fails, i.e., such that $|G|$ has a factor $d$ but $G$ has no subgroup of order $d$.
   **HINT:** Consider $G = A_4$.

5. Show that the normalizer $N_G(H)$ of a subgroup $H \subset G$ is the largest subgroup of $G$ containing $H$ as a normal subgroup.

6. Determine the Sylow $p$-subgroups of the alternating group $A_4$ and their conjugacy relations.

7. Determine all the $p$-subgroups of the group of quaternions $\mathbb{H}_8$.

8. Determine the Sylow $p$-subgroups of the dihedral group $D_p$ when $p$ is a prime.

9. Let $\phi : G_1 \to G_2$ be a surjective group homomorphism between finite groups. Show that is $P \subset G_1$ is a Sylow $p$-subgroup, then $\phi(P)$ is a Sylow $p$-subgroup of $G_2$.

10. Show that if $P \subset G$ is a Sylow subgroup then $N_G(N_G(P)) = N_G(P)$.

11. Let $p$ and $q$ be distinct prime numbers such that $p|(q-1)$ and $s$ a natural number such that $s \not\equiv 1 \pmod{q}$ and $s^p \equiv 1 \pmod{q}$ (one can show that such an $s$ always exists). Show that:
   
   (a) The map $\alpha : \mathbb{Z}_q \to \mathbb{Z}_q$, $k \mapsto sk$, is an automorphism of $\mathbb{Z}_q$;
   
   (b) The map $\theta : \mathbb{Z}_p \times \mathbb{Z}_q \to \mathbb{Z}_q$, $(i, k) \mapsto \alpha^i(k)$, is an action of $\mathbb{Z}_p$ on $\mathbb{Z}_q$ by automorphisms;
   
   (c) The semi-direct product $\mathbb{Z}_p \rtimes \mathbb{Z}_q$ is a group with two generators $a$ and $b$ satisfying the relations: $a^p = 1, b^q = 1, ab = b^s a$. 
5.3 Nilpotent and Solvable Groups

The $p$-groups, as we saw in the last section, play a crucial role in the study of the structure of finite groups. A $p$-group is an example of a nilpotent group. In this section we will study this class of groups, as well as the largest class of solvable groups. These two classes of groups appear naturally when one looks at the failure of elements of a group in commuting.

Let $G$ be any group. The commutator of two elements $g_1, g_2 \in G$ is the element $g_1^{-1}g_2^{-1}g_1g_2 \in G$. We denote this element by $(g_1, g_2)$, so that:

$$g_1g_2 = g_2g_1 \cdot (g_1, g_2).$$

The element $(g_1, g_2)$ measures the failure of $g_1$ and $g_2$ in commuting. The next result gives some elementary properties of commutators whose verification is a simple exercise.

**Proposition 5.3.1 (Properties of commutators).** Let $g_1, g_2, g_3 \in G$ be any elements in a group. Then:

(i) $(g_1, g_2)^{-1} = (g_2, g_1)$;

(ii) $(g_1, g_2) = e$ if and only if $g_1$ and $g_2$ commute;

(iii) $g(g_1, g_2) = (g_1, g_2)$;

(iv) $(g_1 g_2, g_3) \cdot (g_2 g_3, g_1) \cdot (g_3 g_1, g_2) = e$;

(v) $g_1((g_1^{-1}, g_2), g_3) \cdot g_3((g_3^{-1}, g_1), g_2) \cdot g_2((g_2^{-1}, g_3), g_1) = e$;

(vi) If $\phi : G \to H$ is a group homomorphism, then $\phi((g_1, g_2)) = (\phi(g_1), \phi(g_2))$.

Let $A, B \subset G$ be subgroups. We will denote by $(A, B)$ the subgroup of $G$ generated by all commutators $(a, b)$, where $a \in A$ and $b \in B$. So, by definition, $(A, B)$ is the smallest subgroup of $G$ which contains all elements $(a, b)$, with $a \in A$, $b \in B$. Notice that since $(A, B)$ is a group, if $(a, b) \in (A, B)$ then $(b, a) = (a, b)^{-1} \in (A, B)$, so that $(A, B) = (B, A)$. Notice also that there can exist elements in $(A, B)$ which are not commutators. In general, the elements of $(A, B)$ are of the form

$$(a_1, b_1)^{\pm 1} \cdot (a_2, b_2)^{\pm 1} \cdots (a_s, b_s)^{\pm 1}, \quad a_i \in A, \ b_i \in B,$$

where $s \geq 1$.

Footnote: Often the commutator is also denoted by $[g_1, g_2]$, but we will reserve this notation for the commutator in the context of the so called Lie algebras.
Definition 5.3.2. The derived group $G$ is the subgroup $(G, G)$ of $G$. We denote this group by $D(G)$.\footnote{Some authors also denote the derived group of $G$ by $G'$.}

One also calls $D(G) = (G, G)$ the commutator group of $G$ but this name is a little bit misleading since, as we have already observed, there can exist elements in $D(G)$ which are not commutators.

Proposition 5.3.3 (Properties of the derived group). Let $G$, $G_1$ and $G_2$ be groups.

(i) If $\phi : G_1 \to G_2$ is a group homomorphism, then $\phi(D(G_1)) \subset D(G_2)$, and if $\phi$ is surjective, then $\phi(D(G_1)) = D(G_2)$.

(ii) $D(G)$ is a normal subgroup of $G$.

(iii) $G/D(G)$ is an abelian group and for every homomorphism $\phi : G \to A$ into an abelian group $A$ there exists a unique homomorphism $\tilde{\phi} : G/D(G) \to A$ making the following diagram commute:

\[
\begin{array}{ccc}
G & \xrightarrow{\phi} & A \\
\pi \downarrow & & \downarrow \phi \\
G/D(G) & \xrightarrow{\tilde{\phi}} & A
\end{array}
\]

Proof. (i) Obvious from property (vi) of the commutators.

(ii) For each $g \in G$, the map $h \mapsto ghg^{-1}$ is a group automorphism of $G$. Hence, by (i), it follows that $gD(G)g^{-1} \subset D(G)$, so $D(G)$ is normal in $G$.

(iii) Since $g \cdot h = h \cdot g \cdot (g, h)$ it is obvious that $G/D(G)$ is an abelian group. If $\phi : G \to A$ is a group homomorphism from $G$ into an abelian group $A$, we have

\[
\tilde{\phi} = \phi 
\]

Hence, we can define $\tilde{\phi} : G/D(G) \to A$ by setting $\tilde{\phi}(gD(G)) = \phi(g)$. It is very easy to check that $\tilde{\phi}$ is a group homomorphism. By construction, $\phi = \tilde{\phi} \circ \pi$, where $\pi : G \to G/D(G)$ is the quotient map.

\[\square\]

Note that properties (ii) and (iii) actually characterize the derived group.
5.3. NILPOTENT AND SOLVABLE GROUPS

Examples 5.3.4.

1. Obviously, a group $G$ is abelian if and only if its derived group is $D(G) = \{e\}$.

2. For the group $H_8 = \{1, i, j, k, -1, -i, -j, -k\}$ the derived group is $D(H_8) = \{1, -1\} \cong \mathbb{Z}_2$, since the commutators of elements in $H_8$ are either $1$ or $-1$. This group is a normal subgroup of $H_8$ and the quotient $H_8/D(H_8)$ is isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ (exercise).

3. For the symmetric group $S_3 = \{I, \alpha, \beta, \gamma, \delta, \epsilon\}$ the derived group is the alternating group $A_3 = \{I, \delta, \epsilon\}$. In fact, the commutator of any two permutations is necessarily an even permutation and one checks easily that, for example, $\delta = (\alpha, \gamma)$ and $\epsilon = (\gamma, \alpha)$, so all even permutations are commutators. The group $A_3$ is normal in $S_3$ and $S_3/A_3 \cong \mathbb{Z}_2$.

For a group $G$ one defines its lower central series $\{C^k(G)\}_{k \geq 0}$ by:

$$C^0(G) \equiv G, \quad C^{k+1}(G) \equiv (G, C^k(G)) \quad (k \geq 0).$$

One obtains a chain of groups:

$$G = C^0(G) \supset C^1(G) \supset \cdots \supset C^i(G) \supset \cdots$$

When there exists an $n$ such that $C^n(G) = C^{n+k}(G)$ for all $k \geq 0$, we say that the series stabilizes. Notice that this will be the case if $C^n(G) = C^{n+1}(G)$. For an infinite group, the series may never stabilize.

**Definition 5.3.5.** A group $G$ is called nilpotent if there exists some $n$ such that

$$C^n(G) = \{e\}.$$ 

The smallest such integer $n$ is called the nilpotency class of $G$.

Examples 5.3.6.

1. A group is nilpotent of class $\leq 1$ if and only if it is abelian.

2. The group $H_8$ is nilpotent of class 2, since we have $C^0(H_8) = H_8$, $C^1(H_8) = (H_8, H_8) = \mathbb{Z}_2$, $C^2(H_8) = (H_8, \mathbb{Z}_2) = \{e\}$.

3. The dihedral group $S_3$ is not nilpotent: one finds that $C^0(S_3) = S_3$, $C^1(S_3) = A_3$ and $C^2(A_3) = A_3$, so the lower central series stabilizes in $A_3$ at $k = 2$.

4. The subgroup of $GL(n, \mathbb{R})$ formed by the upper triangular matrices with 1s in the main diagonal is nilpotent of class $n - 1$ (exercise).

5. Any subgroup and any quotient of a nilpotent group is nilpotent (exercise).
In general, to try to check directly from the definition that a group is nilpotent, involves a considerably amount of computation. It is convenient to have alternative characterizations of a nilpotent group. For that, we will call a TOWER OF SUBGROUPS of $G$ a sequence

$$G = G^0 \supset G^1 \supset \cdots \supset G^m.$$ 

Also, by a NORMAL TOWER we mean a tower where $G^{k+1}$ is normal in $G^k$, for all $k$. In this case we will often write:

$$G = G^0 \triangleright G^1 \triangleright \cdots \triangleright G^m.$$ 

Finally, by an ABELIAN TOWER we mean a normal tower where $G^k/G^{k+1}$ is abelian, for all $k$.

**Proposition 5.3.7.** The following statements are equivalent:

(i) $G$ is nilpotent of class $\leq n$.

(ii) There exists a tower of subgroups $G = G^0 \supset G^1 \supset \cdots \supset G^n = \{e\}$ where $G^{k+1} \supset (G, G^k)$.

(iii) There exists a subgroup $A$ of the center $C(G)$ such that $G/A$ is nilpotent of class $\leq n - 1$.

**Proof.** (i) $\Leftrightarrow$ (ii): If $G$ is nilpotent of class $\leq n$, then the tower $G^k \equiv C^k(G)$ satisfies (ii). Conversely, given a tower as in (ii), one shows easily by induction that $C^k(G) \subset G^k$: in fact, $C^0(G) = G = G^0$ and

$$C^k(G) \subset G^k \Rightarrow C^{k+1}(G) = (G, C^k(G)) \subset (G, G^k) \subset G^{k+1}.$$ 

Therefore, $C^n(G) \subset G^n = \{e\}$, so $G$ is nilpotent of class $\leq n$.

(i) $\Leftrightarrow$ (iii): If $G$ is nilpotent of class $\leq n$, we have $(G, C^{n-1}(G)) = C^n(G) = \{e\}$, hence $A = C^{n-1}(G)$ is a central subgroup. We leave as an exercise to check that $G/A$ is nilpotent of class $\leq n - 1$. Conversely, let $A$ be a subgroup of $C(G)$ such that $G/A$ is nilpotent of class $\leq n - 1$. The quotient map $\pi : G \to G/A$ is surjective, so $\pi(G, G) = (G/A, G/A)$. Iterating, we conclude that $\pi(C^{n-1}(G)) = C^{n-1}(G/A) = \{e\}$, so that $C^{n-1}(G) \subset A$. Since $A$ is central, it follows that:

$$C^n(G) = (G, C^{n-1}(G)) \subset (G, A) = \{e\}.$$ 

This shows that $G$ is nilpotent of class $\leq n$. 

$\square$
Another series associated with a group is its derived series \( \{D^k(G)\}_{k \in \mathbb{N}} \), which is defined inductively by

\[
D^0(G) \equiv G, \quad D^{k+1}(G) \equiv D(D^k(G)), \quad (k \geq 0).
\]

Note that \( D^0(G) = C^0(G) = G \) and \( D^1(G) = C^1(G) = (G,G) \). We leave it to the exercises to check that:

\[
D^k(G) \subseteq C^{2k-1}(G), \quad (k \geq 0).
\]  

Similarly, to what we did for the central series, we now define:

**Definition 5.3.8.** A group \( G \) is called solvable if for some \( n \),

\[
D^n(G) = \{e\}.
\]

The smallest such integer \( n \) is called the derived length of \( G \).

**Examples 5.3.9.**

1. A group is solvable of length \( \leq 1 \) if and only if it is abelian.
2. Every nilpotent group of class \( \leq 2^{n-1} \) is solvable of length \( \leq n \).
3. The group \( S_3 \) is solvable of class 2: its derived series has terms: \( D^0(S_3) = S_3 \), \( D^1(S_3) = A_3 \) and \( D^2(S_3) = \{e\} \).
4. The subgroup of \( GL(n, \mathbb{R}) \) formed by all invertible \( n \times n \) triangular matrices is solvable (what is its derived length?).
5. Any subgroup and any quotient of a solvable group is solvable.

We also have alternative characterizations of solvable groups. The proof of the following proposition is left for the exercises.

**Proposition 5.3.10.** The following statements are equivalent:

(i) \( G \) is solvable of length \( \leq n \).

(ii) There exists a tower \( G = G^0 \supset G^1 \supset \cdots \supset G^n = \{e\} \) where \( G^k \) is normal em \( G \) and \( G^k/G^{k+1} \) is abelian, for all \( k \geq 0 \).

(iii) There exists an abelian tower \( G = G^1 \rhd G^2 \rhd \cdots \rhd G^n = \{e\} \).

(iv) There exists an abelian normal subgroup \( A \subset G \) such that \( G/A \) is solvable of length \( \leq n - 1 \).
Example 5.3.11.

We saw above that the group $S_3$ is solvable (but not nilpotent). The group $S_4$ is also solvable, for it admits the following abelian tower of subgroups:

$$S_4 \triangleright A_4 \triangleright H \triangleright \{e\},$$

where $H = \{I, (12)(34), (13)(24), (14)(23)\}$.

Exercises.

1. If $\pi : G_1 \rightarrow G_2$ is a group homomorphism, show that $\pi(C^k(G_1)) = C^k(\pi(G_1))$ and $\pi(D^k(G_1)) = D^k(\pi(G_1))$.

2. Is the dihedral group $D_n$ nilpotent? Solvable? (your answer may depend on $n$).

3. Show that the subgroup of $GL(n, \mathbb{R})$ formed by the upper triangular matrices with 1s in the main diagonal is nilpotent of class $n - 1$.

4. Show that the subgroup of $GL(n, \mathbb{R})$ formed by the invertible upper triangular matrices is solvable. What is its derived length?

5. Let $G$ be a group. Show that:

   (a) If $H_1, H_2, H_3 \subset G$ are normal subgroups, then

   $$(H_1, (H_2, H_3)) \subset (H_3, (H_2, H_1)) \cdot (H_2, (H_1, H_3));$$

   (HINT: Use property (v) of the commutators.)

   (b) for all $m, n \in \mathbb{N}$

   $$\langle C^m(G), C^n(G) \rangle \subset C^{m+n}(G);$$

   (c) for all $n \in \mathbb{N}$

   $$D^n(G) \subset C^{2^{n-1}}(G).$$

6. Show that any subgroup and any quotient of a nilpotent (respectively, solvable) group is a nilpotent (respectively, solvable) group. What about direct products?

7. Show that if $G$ is nilpotent (respectively, solvable) of class (respectively, length) $\leq n$, then $G/C^{n-1}(G)$ (respectively, $G/D^{n-1}(G)$) is nilpotent (respectively, solvable) of class (respectively, length) $\leq n - 1$.

8. Give a proof of Proposition 5.3.10.
9. The upper central series of a group $G$ is the tower $\{C_k(G)\}_{k \in \mathbb{N}}$ defined inductively as follows: $C_1(G) = C(G)$ and $C_k(G)$ is the largest normal subgroup of $G$ such that $C_k(G)/C_{k-1}(G)$ is the center of $G/C_{k-1}(G)$. Show that:

(i) $C_k(G) = \{g \in G : (g, h) \in C_{k-1}(G), \forall h \in G\}$;

(ii) A group is nilpotent if and only if $G = C_n(G)$, for some $n$.

10. Show that a $p$-group is nilpotent.

11. Show that a finite group is nilpotent if and only if it is a direct product of $p$-subgroups.

(HINT: If $G$ is a nilpotent group, show that:

(a) If $H \subseteq G$ is a subgroup, then $H \subseteq N_G(H)$;

(b) Every Sylow subgroup $P \subset G$ is normal;

(c) $G$ is the direct product of its Sylow subgroups.)

5.4 Simple Groups

Contrary to what Example 5.3.11 may suggest, the symmetric groups $S_n$, for $n \geq 5$, are not solvable. They are related to a different class of groups, which in some sense is on the other extreme relative to the class of solvable groups. Notice that if one looks for an abelian tower of subgroups in some group $G$, then one may end up with a non-abelian subgroup which has no normal subgroups, besides the trivial ones. This motivates partially the following definition:

**Definition 5.4.1.** A group $G$ is called **simple** if it has no other normal subgroups besides the trivial subgroups $\{e\}$ and $G$.

In other words, the simple groups are the the groups for which there exists only one congruence equivalence relation.

**Examples 5.4.2.**

1. A subgroup of an abelian group is always a normal subgroup. Hence, an abelian group $G$ is simple if it is cyclic of prime order $p$, i.e., $\mathbb{Z}_p$.

2. If a solvable group $G$ is simple, then its derived group must be $D(G)$ must be trivial, so $G$ is abelian. Hence, the only simple solvable groups are the groups $\mathbb{Z}_p$ with $p$ prime.

3. If $G$ is a simple non-abelian group then $G = D(G)$. 
The most elementary, non-abelian, finite simple groups are the alternating groups $A_n$, with $n \geq 5$. Galois found that the fact that $A_n$ is simple for $n \geq 5$ is the reason why there is no general algebraic formula for the roots of algebraic equations of order greater or equal to 5. We will study these issues in Chapter 7.

**Theorem 5.4.3.** The alternating groups $A_n$ are simple if $n \geq 5$.

**Proof.** We will show that if $\{I\} \neq N \subset A_n$ is a normal subgroup then $N = A_n$. We divide the proof into 3 steps:

(i) *The group $A_n$ is generated by 3-cycles:* Any representation of $\pi \in A_n$ as product of transpositions has an even number of terms. On the other hand, any product of 2 distinct transpositions can be written as a product of 3-cycles (for example, $(12)(23) = (123)$ and $(12)(34) = (123)(234)$).

(ii) *If $N$ contains a 3-cycle, then $N = A_n$: *Assume, for example, that $(123) \in N$. Then conjugation by the element

$$
\delta = \begin{pmatrix}
1 & 2 & 3 & 4 & 5 & \cdots \\
 i & j & k & l & m & \cdots
\end{pmatrix}
$$

gives $\delta(123)\delta^{-1} = (ijk)$. This shows that the claim holds as long as $\delta \in A_n$. This can always be achieved, eventually by replacing $\delta$ by $\delta \to (lm)\delta$.

(iii) *$N$ contains a 3-cycle:* We choose an element $\alpha \neq I$ in $N$ which has the property that moves the smallest number of integers:

(PM) The number of integers $i$ such that $\alpha(i) = i$ is maximal among all elements of $N$.

We show that $\alpha$ must be a 3-cycle by contradiction. Suppose that $\alpha$ is not a 3-cycle. Since we cannot have $\alpha = (123l)$ (this permutation is odd), the expression of $\alpha$ as a product of disjoint cycles takes either of the following forms:

$$
\alpha = (123\ldots)(\ldots), \text{ ou } \alpha = (12)(34\ldots)\ldots
$$

where in the first case $\alpha$ permutes at least two more elements (for example, 4 and 5). One checks easily that if we let $\beta = (345)$, then $\gamma = (\alpha, \beta)$ belongs to $N$ and has the following properties:

(a) if $i > 5$ and $\alpha(i) = i$, then $\gamma(i) = i$;
(b) $\gamma(1) = 1$;
(c) in the second case, $\gamma(2) = 2$;
This shows that $\gamma$ is an element contradicting property (MP) of $\alpha$. Hence, $\alpha$ must be a 3-cycle, as claimed. 

**Corollary 5.4.4.** $S_n$ is not solvable if $n \geq 5$.

**Proof.** Assume that $S_n$ is solvable. The theorem would imply that $A_n \subset S_n$ is a solvable simple group, i.e., would be abelian, a contradiction.

The simple groups are, in some sense, indecomposable groups and that is the reason behind their name. One can decompose any finite group into “simple components”. In fact, if $G$ is a finite group then it has a normal tower

$$G = A^0 \triangleright A^1 \triangleright \ldots \triangleright A^m = \{e\},$$

where $A^k/A^{k+1}$ is a simple group, for all $k$: one chooses for $A^1$ any normal subgroup of $A^0 = G$ which is not contain in any normal subgroup of $A^0$, chooses for $A^2$ any normal subgroup of $A^1$ which is not contain in any normal subgroup of $A^1$, and so on.

**Definition 5.4.5.** A composition series of a group $G$ is a normal tower:

$$G = A^0 \triangleright A^1 \triangleright \ldots \triangleright A^m = \{e\},$$

where $A^k/A^{k+1}$ is a simple group.

Hence a finite group $G$ always admits a composition series. We will see below that the Jordan-Hölder Theorem states that composition series are “essentially unique”. In order to state it precisely, consider two normal towers of a group $G$:

$$G = A^0 \triangleright A^1 \triangleright \ldots \triangleright A^s,$$

$$G = B^0 \triangleright B^1 \triangleright \ldots \triangleright B^r.$$

One says that:

- the tower $\{A^i\}_{i=0}^s$ is a refinement of the tower $\{B^j\}_{j=0}^r$ if $r < s$ and for each $i$ there exists a $j_i$ such that $B^{j_i-1} \supseteq A^i \supseteq B^{j_i}$.

- The towers $\{A^i\}_{i=0}^s$ and $\{B^j\}_{j=0}^r$ are called equivalent if $r = s$ and there exists a permutation of the indices $i \mapsto i'$ such that

$$A^i/A^{i+1} \simeq B^{i'}/B^{i'+1}.$$
CHAPTER 5. FINITE GROUPS

Theorem 5.4.6 (Schreier). Let $G$ be a group. Any two normal towers of subgroups of $G$ which end with the trivial subgroup $\{e\}$ admit equivalent refinements.

Proof. Let $\{A^i\}_{i=0}^q$ and $\{B^j\}_{j=0}^r$ be two normal towers ending in the trivial group. Define:

$$\bar{A}^i j \equiv A^{i+1}(B^j \cap A^i).$$

Since $\bar{A}^i 0 = A^i$, $\bar{A}^i j$ is a normal subgroup of $\bar{A}^i j$, it follows that $\{\bar{A}^i j\}$ is a refinement of $\{A^i\}$. Similarly, defining

$$\bar{B}^i j \equiv B^{j+1}(A^i \cap B^j),$$

one obtains a refinement of $\{B^j\}$. The proof is completed by invoking the following lemma whose proof is left as an exercise:

Lemma 5.4.7 (Zassenhaus). If $G \supset S \triangleright S'$ and $G \supset T \triangleright T'$, then $S'(S \cap T')$ is normal in $S'(S \cap T)$, $T'(S' \cap T)$ is normal in $T'(S \cap T)$ and there is an isomorphism:

$$\frac{S'(S \cap T)}{S'(S \cap T')} \cong \frac{T'(S \cap T)}{T'(S' \cap T')}.$$

If in Zassenhaus’s Lemma we let $S = A^i$, $S' = A^{i+1}$, $T = B^j$ and $T' = B^{j+1}$, it follows that:

$$\frac{\bar{A}^i j}{A^{i+1}(A^i \cap B^j)} \cong \frac{B^{j+1}(A^i \cap B^j)}{B^{j+1}(A^{i+1} \cap B^j)} = \frac{\bar{B}^i j}{B^{j+1}}.$$

Hence, $\{\bar{A}^i j\}$ and $\{\bar{B}^i j\}$ are equivalent refinements. 

As a corollary of Schreier’s Theorem, we obtain:

Theorem 5.4.8 (Jordan-Hölder). Two composition series of a finite group $G$ are equivalent.

Proof. Notice that a composition series is precisely a normal tower of subgroups ending with the trivial group, and which do not admit any refinement. Therefore, by Schreier’s Theorem, any two such towers must be equivalent.

The Jordan-Hölder Theorem shows that the composition series is an invariant of a finite group, i.e., any two isomorphic finite groups have equivalent composition series. Hence, they can be used to decide if two groups...
are isomorphic. For example, two groups which have composition series of different lengths cannot be isomorphic. The length of a composition series is a numeric invariant of a finite group: any two isomorphic finite groups have the same length.

**Examples 5.4.9.**

1. The cyclic group $G = \langle a \rangle$ of order $p^m$ has composition series of length $m$:  
   $$G = \langle a \rangle \triangleright \langle a^p \rangle \triangleright \cdots \triangleright \langle a^{p^{m-1}} \rangle \triangleright \langle a^{p^m} \rangle = \{e\}.$$  

2. The group $\mathbb{H}_8 = \{1, i, j, k, -1, -i, -j, -k\}$ has the following composition series of length 3:  
   $$\mathbb{H}_8 \triangleright \{1, i, -1, -i\} \triangleright \{1, -1\} \triangleright \{1\},$$  
   $$\mathbb{H}_8 \triangleright \{1, j, -1, -j\} \triangleright \{1, -1\} \triangleright \{1\},$$  
   $$\mathbb{H}_8 \triangleright \{1, k, -1, -k\} \triangleright \{1, -1\} \triangleright \{1\}.$$  
   These composition series are all isomorphic: the quotient groups $G^i/G^{i+1}$ are all isomorphic to $\mathbb{Z}_2$.

3. By definition, a group is simple if has a unique trivial composition series:  
   $$G \triangleright \{e\}.$$  

4. A finite group $G$ is solvable if and only if for any of its composition series  
   $$G = G^0 \triangleright G^1 \triangleright \cdots \triangleright G^m = \{e\}$$  
   the factors $G^k/G^{k+1}$ are all cyclic groups of prime order, i.e., isomorphic to some $\mathbb{Z}_{p^k}$, with $p_k$ prime. In fact, if $G$ is solvable, all subgroups $G^k$ are solvable and also its quotients $G^k/G^{k+1}$ are solvable (cf. Exercise 6 in Section 5.3). But any solvable simple group is abelian, hence isomorphic to some $\mathbb{Z}_{p^k}$.

The results in this section suggest a program to classify all finite groups: one should first classify all finite simple groups and then classify all the possible composition towers made of this groups. The classification of all finite simple groups is one of the greatest achievements of modern days Mathematics. The classification theorem can be roughly stated as follows:

**Theorem 5.4.10** (Classification of finite simple groups). *Up to isomorphism there are 18 infinite families of simple groups and 26 sporadic simple groups.*
Examples of infinite families are the groups $\mathbb{Z}_p$, with $p$ prime, and the groups $A_n$, with $n \geq 5$. Another example is the projective linear group $\text{PL}(n, K)$ of some finite field $K$: it is obtained by taking the quotient of the group $\text{SL}(n, K)$ of $n \times n$ matrices of determinant 1 by its center. The sporadic simple groups are more mysterious groups. For example, the largest sporadic simple group has order:

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71 \sim 8 \cdot 10^{53}$$

and it is called the MONSTER GROUP.

**Exercises.**

1. Proof Zassenhaus’s Lemma.
   (Hint: Use the Second Isomorphism Theorem to show that each of the quotients that appear in the statement of the lemma are both isomorphic to $S \cap T / (S \cap T') (S' \cap T)$.)

2. Show that a cyclic $p$-group has a unique composition series.

3. Show that an abelian group has a composition series if and only if it is finite.

4. Determine the composition series of the following groups:
   (a) $\mathbb{Z}_6 \times \mathbb{Z}_5$;
   (b) $S_4$;
   (c) $G$ where $|G| = pq$ ($p$ and $q$ prime).

5. If a simple group $G$ has a subgroup of index $n > 1$, show that the order of $G$ divides $n!$. Use this to prove that $A_5$ has no subgroup of order 15.

6. Show that a group of order $pq^2$, where $p$ and $q$ are primes, is solvable.

7. Classify all groups of order 20.

8. Show that there exists no non-abelian simple group of order smaller than 30.

**5.5 Symmetry Groups**

We introduced in Chapter 1 the notion of symmetry group of a subset $\Omega \subset \mathbb{R}^n$. We will now apply the results we have obtained before on the structure of finite groups to the classification of symmetry groups.
Recall that the group of symmetries of $\mathbb{R}^n$ is, by definition, the euclidean group $E(n)$ consisting of all isometries of $\mathbb{R}^n$. We saw in Exercise 5.1.5, that this group is isomorphic to the semi-direct product $O(n) \ltimes \mathbb{R}^n$. In fact, by Theorem 1.8, an isometry $f \in E(n)$ takes the form

$$f(x) = Ax + b,$$

where $A \in O(n)$ represents an orthogonal transformation and $b \in \mathbb{R}^n$ a translation. Recall also that an orthogonal transformation is called a rotation if $\det A = 1$.

If $\Omega \subset \mathbb{R}^n$ its group of symmetries $G_\Omega \subset E(n)$ consists of all isometries which map $\Omega$ into itself:

$$G_\Omega = \{f \in E(n) : f(\Omega) = \Omega\}.$$

When $\Omega$ is bounded, it is obvious that group $G_\Omega$ contains only orthogonal transformations. Moreover, we have:

**Proposition 5.5.1.** If $G_\Omega$ is the group of symmetries of a bounded set $\Omega \subset \mathbb{R}^n$, then one of the following statements holds:

(i) $G_\Omega$ contains only rotations;

(ii) The rotations of $G_\Omega$ form an index 2 subgroup (hence, normal) in $G_\Omega$.

**Proof.** Since $\Omega$ is bounded, $G_\Omega \subset O(n)$ and $G_\Omega$ is formed by rotations iff $G_\Omega \subset SO(n)$. If $G_\Omega \not\subset SO(n)$, then $H = G_\Omega \cap SO(n)$ is the subgroup of rotations in $G_\Omega$, and coincides with the kernel of the epimorphism $\det : G_\Omega \to \{1, -1\}$. By the First Isomorphism Theorem, it follows that

$$[G_\Omega : H] = |G_\Omega/H| = |\{1, -1\}| = 2.$$

In the reminder of this section we will consider only finite symmetry groups. We will see that the the results we have obtained so far on the structure of finite groups allows us to completely classify the symmetry groups of planar ($n = 2$) and tridimensional ($n = 3$) bounded figures.

### 5.5.1 Symmetry groups of plane figures

Before we get to the classification of symmetry groups of plane bounded figures, let us recall some examples of plane figures whose symmetry groups we have encountered before.
Examples 5.5.2.

1. In Example 1.8.7.2 we saw that $D_3$, the group of symmetries of an equilateral triangle, contains 6 elements: 3 rotations ($I, \frac{2\pi}{3}, \frac{4\pi}{3}$) and 3 reflections across its lines of symmetry.

More generally, the dihedral group $D_n$, the symmetry group of a regular polygon with $n$ sides, has $2n$ elements: $n$ rotations and $n$ reflections across the lines of symmetry of the polygon. If we denote by $\rho$ a rotation of $\frac{2\pi}{n}$ and by $\sigma$ a reflection across one line of symmetry of the polygon, then:

$$D_n = \langle \rho, \sigma \rangle = \{I, \rho, \ldots, \rho^{n-1}, \sigma, \sigma\rho, \ldots, \sigma\rho^{n-1}\}.$$  

As we saw in Example 4.6.3.4, this group has the presentation

$$\{\{\rho, \sigma\}, \sigma^2 = I, \rho^n = I, \rho\sigma = \sigma\rho^{-1}\}.$$

2. Consider the group of symmetries of the sails of a windmill. It is a cyclic group of order 4, generated by a rotation $\rho$ by $\frac{\pi}{2}$:

$$C_4 = \{I, \rho, \rho^2, \rho^3\}.$$  

More generally one can consider a windmill with $n$ sails, leading to a cyclic group of symmetries with $n$ elements: $C_n = \{I, \rho, \ldots, \rho^{n-1}\}$

![Figure 5.5.1: Plane figure with cyclic group of symmetries.](image)

The symmetry groups that one finds in these examples exhaust all the possibilities since one has the following:

Theorem 5.5.3. A finite group of symmetries of a plane figure $\Omega \subset \mathbb{R}^2$ is isomorphic to either $C_n$ or $D_n$.

Proof. Let $\Omega \subset \mathbb{R}^2$ and let $G$ be its group of symmetries, which we assume to be finite. According to Proposition 5.5.1 we need to consider two cases:

\[\text{In the study of symmetries it is common to denote by } C_n \text{ the cyclic group of order } n, \text{ which we have denoted before by } \mathbb{Z}_n.\]
5.5. SYMMETRY GROUPS

(a) $G$ contains only rotations: Let $\rho \in G$ be a rotation by same angle $\theta_{\rho}$. We can assume that $\theta_{\rho}$ is the smallest possible among all rotations of $G$ (which exists, since $G$ is finite). Then $\{I, \rho, \rho^2, \ldots\} \subset G$. On the other hand, if $\tilde{\rho} \in G$ is a rotation by an angle $\theta_{\tilde{\rho}}$ which is not among the powers $\rho^n$, then there exists some positive integer $k$ such that:

$$k\theta_{\rho} < \theta_{\tilde{\rho}} < (k+1)\theta_{\rho}.$$ 

It follows that $\tilde{\rho}\rho^{-k}$ is a rotation by an angle smaller than $\rho$, a contradiction. Hence, $G = \{I, \rho, \ldots, \rho^{n-1}\} = C_n$.

(b) $G$ contains a reflection $\sigma$: By (a) the subgroup $H \subset G$ of proper rotations is of the form $H = \{I, \rho, \ldots, \rho^{n-1}\}$. Since $[G : H] = 2$, we find that $G = \{I, \rho, \ldots, \rho^{n-1}, \sigma, \sigma\rho, \ldots, \sigma\rho^{n-1}\}$. We leave as an exercise to check that in this case $G \cong D_n$.

If $n = p$ is prime, $|D_p| = 2p$ and the Sylow Theorems show that the subgroups of $D_p$ are:

(a) $p$ subgroups of order 2: $\{I, \sigma\}, \{I, \sigma\rho\}, \ldots, \{I, \sigma\rho^{p-1}\}$;

(b) 1 subgroup of order $p$: $C_p = \{I, \rho, \ldots, \rho^{p-1}\}$;

The subgroup of order $p$ is normal. Since $p$ is a prime, the subgroups of order 2 are conjugate via a rotation (exercise). Geometrically, this means that we can obtain any reflection from a fixed reflection by conjugating it by rotations. For example, the figure below illustrates in the case $p = 5$ how the reflection $\rho\sigma$ can be obtain from the reflection $\sigma$, conjugating by the rotation $\rho^2$.

The structure of the subgroups of the dihedral group $D_n$ when $n$ is not a prime is more complex and will not be discussed here.

\[\text{Figure 5.5.2: Symmetries of a pentagon.}\]
5.5.2 Symmetry groups of tridimensional figures

We consider now the symmetry groups $G$ of a tridimensional figure. We start by looking at the rotational symmetries, i.e., the case where $G \subset SO(3)$.

**Theorem 5.5.4.** A finite group of rotational symmetries of a figure $\Omega \subset \mathbb{R}^3$ is isomorphic to one of the following rotational symmetry groups:

(i) The symmetry group $C_n$ of a windmill with $n$ sails;

(ii) The symmetry group $D_n$ of a regular polygon with $n$ sides;

(iii) The rotational symmetry group $T$ of a regular tetrahedron;

(iv) The rotational symmetry group $O$ of a cube or a regular octahedron:

(v) The rotational symmetry group $I$ of a regular dodecahedron or a regular icosahedron:

**Proof.** Let $G \subset SO(3)$ be a finite subgroup. The idea of the proof consists in introducing an action of $G$ in a fine set $P$ and then exploring the class equation [5.1.3].
For the finite set $P$ we will take the set of poles of $G$: an element $p \in S^2 = \{ x \in \mathbb{R}^3 : |x| = 1 \}$ is called a pole if there exists a non-trivial rotation $g \in G$, such that $g \cdot p = p$. Therefore, $p$ is a pole fixed by $g \in G$ if and only if $p \in S^2 \cap L$, where $L$ the axis of rotation of $g$. In particular, to each $g \in G$ one associates two poles. Obviously, $P \neq \emptyset$ if $G$ is non-trivial.

The group $G$ acts on the set of its poles $P$: if $p \in P$ is such that $g \cdot p = p$ for some $g \in G$, then for any $h \in G$, we have that $(hgh^{-1})h \cdot p = h \cdot p$.

Since $hgh^{-1} \neq e$ of $g \neq e$, it follows that $h \cdot p \in P$.

The study of the action of $G$ on $P$ leads to the following lemma:

**Lemma 5.5.5.** The action of $G$ on the set of its poles $P$ satisfies

\[
\sum_i (1 - \frac{1}{r_{p_i}}) = 2 - \frac{2}{N},
\]

where the sum is over the orbits $O_i$ of the action, $p_i$ is a pole representing $O_i$, $r_{p_i}$ is the order of the isotropy subgroup $G_{p_i}$ and $N$ is the order of $G$.

**Proof of the Lemma.** For each pole $p \in P$, the isotropy subgroup $G_p$ consists of those elements $g \in G$ that fix $p$. Let $N = |G|$ and $r_p = |G_p|$. For each $g \in G - \{e\}$ there exist 2 poles associated to $g$, hence we find:

\[
2(N - 1) = \sum_{g \in G \atop g \neq e} 2 = \sum_{p \in P} (r_p - 1).
\]

Let us group together the elements of $P$ in terms of the orbits of $G$. If $O_i$ is an orbit, we choose a representative $p_i \in O_i$ and we write $n_i = |O_i|$. Then equation (5.5.2) yields

\[
\sum_i n_ir_{p_i} - |P| = 2(N - 1),
\]

where the sum is now over the orbits $O_i$ of the action. Since $O_i \simeq G/G_{p_i}$, it follows that $n_ir_{p_i} = N$, and we conclude:

\[
\sum_i N - |P| = 2(N - 1).
\]

On the other hand, the class equation (5.1.3) gives

\[
|P| = \sum_i |O_i| = \sum_i \frac{N}{r_{p_i}}.
\]

Replacing (5.5.3) in (5.5.4), we obtain (5.5.1). \qed
CHAPTER 5. FINITE GROUPS

Equation (5.5.1) puts restrictions over $G$ which eventually lead to its classification. A first remark is that there can be at most 3 orbits. Indeed, the right hand side of (5.5.1) is $< 2$, while each term on the left side is $\geq \frac{1}{2}$. It follows that we have the following possibilities:

(i) 1 orbit: We would have

$$1 - \frac{1}{r_1} = 2 - \frac{2}{N}.$$  

This equation has no solutions

(ii) 2 orbits: We obtain

$$\frac{1}{r_1} + \frac{1}{r_2} = \frac{2}{N}.$$  

The only solution is $r_1 = r_2 = N$. This means that we have 2 poles $p_1$ and $p_2$ which are fixed by all elements of $G$. Hence, we have $G = C_N$, the group of rotations around the axis that goes through $p_1$ and $p_2$.

(iii) 3 orbits: In this case we obtain

$$\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} - 1 = \frac{2}{N}.$$  

We can assume, without loss of generality, that $r_1 \leq r_2 \leq r_3$. It follows that $r_1 = 2$ and we obtain the following subgroups:

(a) $r_1 = r_2 = 2$, $N = 2r_3$. $|O_1| = |O_2| = \frac{N}{2}$, $|O_3| = 2$;  

(b) $r_1 = 2$, $r_2 = r_3 = 3$, $N = 12$. $|O_1| = 6$, $|O_2| = |O_3| = 4$;  

(c) $r_1 = 2$, $r_2 = 3$, $r_3 = 4$, $N = 24$. $|O_1| = 12$, $|O_2| = 8$, $|O_3| = 6$;  

(d) $r_1 = 2$, $r_2 = 3$, $r_3 = 5$, $N = 60$. $|O_1| = 30$, $|O_2| = 20$, $|O_3| = 12$.

We leave as an exercise to check that the cases (a), (b), (c) and (d) correspond, respectively, to the groups $G \simeq D_N$, $G \simeq T$, $G \simeq O$ and $G \simeq I$.

In case (a), the poles are the intersections of the symmetry lines of the regular polygon with the unit sphere and the intersection of the line through the center of the polygon and perpendicular to its plane with the unit sphere. In cases (b)-(d), the poles are the intersections of the symmetry axis of the regular polyhedra with the unit sphere.
5.5. SYMMETRY GROUPS

Proposition 5.5.1 now shows that the full symmetry groups take one of the following forms:

(a) If \(-I \in G\), then \(G = H \cup -H\), where \(H = \{I, \rho_1, \ldots, \rho_{n-1}\}\) the subgroup of rotations in \(G\) and \(-H \equiv \{-I, -\rho_1, \ldots, -\rho_{n-1}\}\).

(b) If \(-I \notin G\), then \(G = H \cup \tilde{H}\), where \(H\) is the subgroup of rotations in \(G\) and, if \(-\rho \in \tilde{H}\) then \(\rho\) has even order and \(\rho^2 \in H\).

Since we already know what \(H\) can be, by Theorem 5.5.4, this leads to the classification of the finite groups of symmetries of a tridimensional figure \(\Omega \subset \mathbb{R}^3\).

Podemos descrever, de forma mais explícita, The symmetry groups of the regular polyhedra can be described even more explicitly. We illustrate with the case of a regular dodecahedron (for the remaining regular polyhedra, see the exercises at the end of the section).

Example 5.5.6.

The group of symmetries \(I\) of a dodecahedron (or its dual, a regular icosahedron) has order \(60 = 2^2 \times 3 \times 5\). By the Sylow Theorems we obtain the following subgroups:

(i) Subgroups of order 5: the number of subgroups of order 5 divides 12 and is equal to 1 (mod 5). Hence, we can have either 1 or 6 subgroups of order 5. There are 6 subgroups each generated by a rotation by \(\frac{2\pi}{5}\) around a line that connects the centers of two parallel faces of the dodecahedron. These subgroups are precisely the isotropy subgroups of the orbit \(O_3\).

(ii) Subgroups of order 3: the number of subgroups of order 3 divides 20 and is equal to 1 (mod 3). We can have 1, 4 or 10 subgroups of order 3. There 10 subgroups each generated by a rotation by \(\frac{2\pi}{3}\) around a line through opposite vertices of the dodecahedron. These subgroups are precisely the isotropy subgroups of the orbit \(O_2\).

(iii) Subgroups of order 4: the number of subgroups of order 4 divides 15 and is equal to 1 (mod 2). We can have 1, 3, 5 or 15 subgroups. There are 15 subgroups of order 2 corresponding to rotations by \(\pi\) around the lines through the centers of parallel edges of the dodecahedron. These subgroups are precisely the isotropy subgroups of the orbit \(O_1\), and they give rise to 5 subgroups of order 4, formed by the rotations associated to 3 orthogonal edges of the dodecahedron (in the figure below, these are the edges parallel to the edges of the cube).

This list of rotations of order 2, 3 and 5, exhaust all the elements of the group \(I\), since:

\[
60 = |I| = 1 + 15 + 20 + 24.
\]

\(\text{order 2} \quad \text{order 3} \quad \text{order 5}\)
This description of the elements of $I$ also allows to show that $I$ is a simple group. In fact, if $H \subset G$ is a normal subgroup and $H$ contains an element of order $r$, then $H$ must contain all the elements of order $r$, since the Sylow subgroups are all conjugate and the subgroups of order 2 are not normal. Hence, the order of $H$ is a sum of terms in equation (5.5.5). However, there is no integer that is a sum of terms of (5.5.5) and that divides 60. Hence, we must have $H = I$, so this is a simple group.

Our study of the alternating group $A_n$ suggests that $I \simeq A_5$. To see that this is indeed the case, consider the 5 cubes inscribed in the dodecahedron (see figure). The action of $I$ in the vertices of the dodecahedron transforms the vertices of the cubes into vertices of the cubes (preserving their orientation). This gives an action of $I$ in a set of 5 elements:

$$T(R)(\text{cube}) = R(\text{cube})$$

Figure 5.5.3: One of the cubes inscribed in a dodecahedron.

Since $I$ is simple, this action is effective: $N(T) = \{e\}$. Since $I$ contains only rotations which preserve orientations, we have $\text{Im}(T) \subset A_5$. Finally, since $|I| = 60 = |A_5|$, we conclude that $I \simeq \text{Im}(T) = A_5$.

The classification of finite groups of symmetries of figures $\Omega \subset \mathbb{R}^n$, for $n > 3$, is only known for small values of $n$. A special important case is the groups generated by reflections in hyperplanes of $\mathbb{R}^n$, the so-called Coxeter groups. Its classification, obtained by Coxeter in 19349, is related to the classification of Lie algebras, and has many applications in several domains of Mathematics and Physics.

Exercises.

1. Complete the proof of Theorem 5.5.3.

2. Let $D_p$ be the group of symmetries of a regular polygon with $p$ sides. Show that if $p$ is a prime, all subgroups of $D_p$ of order 2 are conjugate via a rotation.

3. Complete the proof of Theorem 5.5.4.

4. Show that the action of the symmetry group $I$ on the set of 5 cubes inscribed in the dodecahedron is effective (see Example 5.5.6).

5. Show that $T \simeq A_4$.
   (HINT: Consider the action of $T$ on the vertices of the tetrahedron.)

6. Mostre que $O \simeq S_4$.
   (HINT: Consider the action of $O$ on the diagonals of the cube.)
Chapter 6

Modules

6.1 Modules over a ring

Let us recall the definition of a vector space over a field:

**Definition 6.1.1.** A vector space over a field $K$ is an abelian group $(V, +)$ with a binary operation $K \times V \rightarrow V$, written $(k, v) \mapsto kv$, which satisfies:

(i) $k(v_1 + v_2) = kv_1 + kv_2$, $k \in K, v_1, v_2 \in V$;

(ii) $(k + l)v = kv + lv$, $k, l \in K, v \in V$;

(iii) $k(lv) = (kl)v$, $k, l \in K, v \in V$;

(iv) $1v = v$, $v \in V$.

Now let $(G, +)$ be an abelian group, for which we use the additive notation. Recall that we have a binary operation which to the elements $n \in \mathbb{Z}$ and $g \in G$ associates the element $ng \in G$. This operation satisfies:

- $n(g_1 + g_2) = ng_1 + ng_2$, $n \in \mathbb{Z}, g_1, g_2 \in G$;
- $(n + m)g = ng + mg$, $n, m \in \mathbb{Z}, g \in G$;
- $n(mg) = (nm)g$, $n, m \in \mathbb{Z}, g \in G$;
- $1g = g$, $g \in G$.

These properties are formally analogous to the axioms in Definition 6.1.1 where we replace the elements of $V$ (“vectors”) by elements of the group $G$, and the elements of the field $K$ (“scalars”) by elements of the ring $\mathbb{Z}$. 

281
Next consider a linear transformation \( T : \mathbb{R}^n \rightarrow \mathbb{R}^n \). If \( p(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{R}[x] \) is any polynomial and \( \mathbf{v} \in \mathbb{R}^n \) is a vector, we let \( p(x) \cdot \mathbf{v} \in \mathbb{R}^n \) be defined by

\[
p(x) \cdot \mathbf{v} := a_0 \mathbf{v} + a_1 T(\mathbf{v}) + \cdots + a_n T^n(\mathbf{v}) = \sum_{k=1}^{n} a_k T^k(\mathbf{v}),
\]

where:

\[
T^0 = I, \quad T^k = T \circ T \circ \cdots \circ T \quad (k \text{ times}).
\]

For example, if \( T : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \) is the linear transformation whose matrix relative to the canonical basis \( \mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0), \mathbf{e}_3 = (0, 0, 1) \) is:

\[
A = \begin{pmatrix}
2 & 0 & 0 \\
0 & 3 & 1 \\
0 & 0 & 3
\end{pmatrix}.
\]

and we let \( p(x) = 3 + \frac{1}{2} x - 2 x^2 \) and \( \mathbf{v} = (1, 2, 1) \), then

\[
p(x) \cdot \mathbf{v} = 3(1, 2, 1) + \frac{1}{2} T(1, 2, 1) - 2 T^2(1, 2, 1) = (-4, -77/2, -27/2).
\]

It is easy to check that this operation satisfies the following properties:

- \( p(x) \cdot (\mathbf{v}_1 + \mathbf{v}_2) = p(x) \cdot \mathbf{v}_1 + p(x) \cdot \mathbf{v}_2; \)
- \( (p(x) + q(x)) \cdot \mathbf{v} = p(x) \cdot \mathbf{v} + q(x) \cdot \mathbf{v}; \)
- \( p(x) \cdot (q(x) \cdot \mathbf{v}) = (p(x)q(x)) \cdot \mathbf{v}; \)
- \( 1 \mathbf{v} = \mathbf{v}. \)

for any \( p(x), q(x) \in \mathbb{R}[x], \mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n \). In this example, the “vectors” are elements in \( \mathbb{R}^n \) and the “scalars” are elements of the polynomial ring \( \mathbb{R}[x] \).

Notice that in each of these three examples all the properties satisfied by the operation in question involve only the ring structure of the “scalars” (respectively, \( K \), \( \mathbb{Z} \) and \( \mathbb{R}[x] \)) and the additive structures of the “vectors”. There are other similar circumstances where one can find analogues properties, so it is natural to extend these concepts to an arbitrary ring. The resulting algebraic structure, called module, is the basic stone to study concepts of Linear Algebra, such as linear independence, linear transformations, dimension, basis, etc.

---

1From now, in order to simplify the notation, we will not use the convention of denoting an indeterminate by \( x \). Usually, the indeterminates will be denoted by \( x_1, \ldots, x_n \) (or \( x \) if \( n = 1 \); or \( x, y \) if \( n = 2 \); or \( x, y, z \) if \( n = 3 \)) and we will denote by bold the vectors (the elements of a vector space or, more generally, of a module).
Definition 6.1.2. A module $M$ over a ring $A$ or an $A$-module is an abelian group $(M,+)$, together with an operation of a unitary ring $A$ on $M$, written $(a,v) \mapsto av$, satisfying the following properties$^2$:

(i) $a(v_1 + v_2) = av_1 + av_2$, $a \in A, v_1, v_2 \in M$;
(ii) $(a_1 + a_2)v = a_1v + a_2v$, $a_1, a_2 \in A, v \in M$;
(iii) $a_1(a_2v) = (a_1a_2)v$, $a_1, a_2 \in A, v \in M$;
(iv) $1v = v$, $v \in M$.

To be precise, the modules we have just defined are “left modules”. We leave it to the reader the task of defining “right modules”. All the results in this chapter will be stated for left modules, but they are valid mutatis mutandis for right modules. For a commutative ring $A$ there is little sense in distinguishing between a left and a right module.

We will denote by $0_A$ and $0_M$ the identities in $(A,+)$ and $(M,+)$, since $(M,+)$ is an abelian group the element $nv \in M$, where $n \in \mathbb{Z}$ and $v \in M$, has the usual meaning. Similarly, we have the element $na \in A$, where $n \in \mathbb{Z}$ and $a \in A$. One checks easily the following elementary properties:

Proposition 6.1.3. For any $A$-module $M$:

(i) $a0_M = 0_M$, $a \in A$;
(ii) $0_A v = 0_M$, $v \in M$;
(iii) $(-a)v = -(av) = a(-v)$, $a \in A, v \in M$;
(iv) $n(av) = (na)v$, $n \in \mathbb{Z}, a \in A, v \in M$.

A submodule $N$ of an $A$-module $M$ is a subgroup of $(M,+)$ which is closed for multiplication by elements of $A$: if $a \in A$ and $v \in N$, then $av \in N$. A submodule is obviously an $A$-module.

Examples 6.1.4.

1. Vector spaces are the same thing as modules over a field. More generally, one can call a module over a division ring $D$ a vector space. In this case the submodules are the linear subspaces.

2. Abelian groups are the same thing as modules over $\mathbb{Z}$. In this case, the submodules coincide with the subgroups of $G$.

$^2$One can also consider modules over rings $A$ without an identity 1, where one omits axiom (iv). We shall only consider modules over unitary rings.
3. Generalizing the example above, a fixed linear transformation \( T : V \to V \) makes the vector space \( V \) into a module over the polynomial ring \( K[x] \): given \( p(x) = a_0 + a_1 x + \cdots + a_n x^n \in K[x] \) and \( v \in V \) then one sets:

\[
p(x) \cdot v := a_0 v + a_1 T(v) + \cdots + a_n T^n(v).
\]

In this example the submodules are the linear subspaces \( W \subset V \) which are invariant under the linear transformation \( T \).

4. If \( A \) is a ring and \( I \subset A \) is a (left) ideal then \( I \) is an \( A \)-module: if \( a \in A \) and \( b \in I \), then \( ab \in I \). Similarly, \( A/I \) is an \( A \)-module, since if \( a \in A \) and \( b + I \in A/I \), we can set:

\[
a(b + I) = ab + I.
\]

5. If \( A \) is a ring and \( B \subset A \) is a subring, then \( A \) is a \( B \)-module. In particular, the rings \( A[x_1, \ldots, x_n] \) and \( A[[x_1, \ldots, x_n]] \) are \( A \)-modules.

6. If \( G \) is an abelian group and \( \text{End}(G) \) is the ring of endomorphisms of \( G \). Then \( G \) is a \( \text{End}(G) \)-module for the operation

\[
\phi g \equiv \phi(g), \quad \phi \in \text{End}(G), g \in G.
\]

7. Let \( A \) and \( B \) be rings and \( \phi : A \to B \) a ring homomorphism. If \( M \) is a \( B \)-module, then we can make \( M \) into an \( A \)-module, denoted \( \phi^* M \), which has the same underlying supporting set but a new operation defined by

\[
a v \equiv \phi(a) v, \quad a \in A, v \in M.
\]

Notice that in all these examples by studying the structure of the module (e.g., the structure of its submodules) one can obtain information about the underlying objects: the subgroups of an abelian group, the invariant subspaces, etc. This will be a recurrent theme in this chapter.

**Definition 6.1.5.** A HOMOMORPHISM OF \( A \)-MODULES \( \phi : M_1 \to M_2 \) is a map between \( A \)-modules which satisfies:

(i) \( \phi(v_1 + v_2) = \phi(v_1) + \phi(v_2), \quad v_1, v_2 \in M; \)

(ii) \( \phi(a v) = a \phi(v), \quad a \in A, v \in M. \)

One defines in an obvious way monomorphisms, epimorphisms and isomorphisms of \( A \)-modules. We will also use interchangeably the terms \( A \)-linear transformations or linear map to denote a homomorphism of \( A \)-modules.

Note that if \( \phi : M_1 \to M_2 \) is a linear transformation, its kernel \( N(\phi) \) and its image \( \text{Im}(\phi) \) are submodules of \( M_1 \) and \( M_2 \).
Examples 6.1.6.

1. A homomorphism of \( \mathbb{Z} \)-modules is the same thing as a homomorphism of abelian groups.

2. If \( V_1 \) and \( V_2 \) are vector spaces over \( K \), the \( K \)-homomorphisms \( \phi : V_1 \to V_2 \) are the usual linear transformations.

If \( M \) is an \( A \)-module and \( N \subset M \) is a submodule, the inclusion \( \iota : N \to M \) is an \( A \)-linear map. The quotient \( M/N \) has a natural structure of an \( A \)-module for which the quotient map \( \pi : M \to M/N \) is an \( A \)-linear map: \( M/N \) is an abelian group and we can define a scalar multiplication by:

\[
    a(v + N) \equiv (av) + N.
\]

One checks easily that (i)-(iv) are satisfied. The module \( M/N \) is called the quotient module of \( M \) by \( N \).

If \( \{N_i\}_{i \in I} \) is a family of submodules of an \( A \)-module \( M \), then \( \bigcap_{i \in I} N_i \) is a submodule of \( M \). Hence, if \( S \subset M \) is a non-empty set, then the intersection of all submodules of \( M \) which contains \( S \) is a submodule \( \langle S \rangle \), called the module generated by \( S \). The elements of \( \langle S \rangle \) are of the form \( a_1v_1 + \cdots + a_rv_r \), where \( a_i \in A \) and \( v_i \in S \).

If \( \{N_i\}_{i \in I} \) is a family of submodules of an \( A \)-module \( M \), we denote by \( \sum_{i \in I} N_i \) the submodule generated by \( S = \bigcup_{i \in I} N_i \). If \( I = \{1, \ldots, m\} \) is finite, we write \( \sum_{i=1}^m N_i \) or \( N_1 + \cdots + N_m \). In general, the elements of \( \sum_{i \in I} N_i \) take the form \( v_{i_1} + \cdots + v_{i_m} \), \( v_{i_j} \in N_{i_j} \).

**Theorem 6.1.7** (Isomorphism Theorems for Modules).

(i) If \( \phi : M_1 \to M_2 \) is a homomorphism of \( A \)-modules, then there exists an isomorphism of \( A \)-modules: \( \text{Im}(\phi) \simeq M_1/N(\phi) \).

(ii) If \( N_1 \) and \( N_2 \) are submodules of an \( A \)-module \( M \), then there exists an isomorphism of \( A \)-modules:

\[
    \frac{N_1 + N_2}{N_2} \simeq \frac{N_1}{N_1 \cap N_2}.
\]

(iii) If \( N \) and \( P \) are submodules of an \( A \)-module \( M \) and \( M \supset N \supset P \), then \( P \) is a submodule of \( N \) and there is an isomorphism of \( A \)-modules:

\[
    M/N \simeq \frac{M/P}{N/P}.
\]

\(^3\text{Note that this does not hold for modules over rings without a unit. For this modules, the elements of } \langle S \rangle \text{ are of the form } \sum_i a_i v_i + \sum_j n_j \tilde{v}_j, \text{ where } a_i \in A, n_j \in \mathbb{Z} \text{ and } v_i, \tilde{v}_j \in S.\)

We leave the easy proofs of this theorems for the exercises.

Let \( \{M_i\}_{i \in I} \) be a family of \( A \)-modules. We define the \( A \)-module \( \prod_{i \in I} M_i \), called **direct product** of the family \( \{M_i\}_{i \in I} \), as follows. The underlying set of \( \prod_{i \in I} M_i \) is the cartesian product of the \( M_i \). If \( (v_i)_{i \in I}, (w_i)_{i \in I} \in \prod_{i \in I} M_i \), then \((v_i)_{i \in I} + (w_i)_{i \in I}\) denotes the element \((v_i + w_i)_{i \in I} \in \prod_{i \in I} M_i \), and if \( a \in A \), then \( a(v_i)_{i \in I} \) denotes the element \((av_i)_{i \in I} \in \prod_{i \in I} M_i \). One checks immediately that \( \prod_{i \in I} M_i \) becomes an \( A \)-module. If \( i \in I \), the **canonical projection** \( \pi_i : \prod_{i \in I} M_i \to M_i \) the \( A \)-linear map which maps \((v_i)_{i \in I} \in \prod_{i \in I} M_i \) to the element \( v_i \in M_i \).

The **direct sum** of a family of \( A \)-modules \( \{M_i\}_{i \in I} \), denoted \( \bigoplus_{i \in I} M_i \), is the submodule of the direct product \( \prod_{i \in I} M_i \) formed by the elements \((v_i)_{i \in I}\) where only a finite number of \( v_i \)'s are non-zero. If \( k \in I \), the **canonical injection** \( \iota_k : M_k \to \bigoplus_{i \in I} M_i \) is the \( A \)-linear map which maps \( v_k \in M_k \) to the element \((v_i)_{i \in I} \in \prod_{i \in I} M_i \) with all \( v_i = 0 \) for \( i \neq k \).

When \( I = \{1, \ldots, m\} \) is finite, the direct sum and direct product coincide. In this case we write \( \bigoplus_{i=1}^m M_i \) or \( M_1 \oplus \cdots \oplus M_m \).

**Proposition 6.1.8.** \( M \simeq M_1 \oplus \cdots \oplus M_m \) (as \( A \)-modules) if and only if there exist \( A \)-linear maps \( \pi_k : M \to M_k \) and \( \iota_k : M_k \to M \) such that:

1. \( \pi_k \circ \iota_k = id_{M_k}, \ k = 1, \ldots, m; \)
2. \( \pi_k \circ \iota_l = 0, \ k \neq l; \)
3. \( \iota_1 \circ \pi_1 + \cdots + \iota_m \circ \pi_m = id_M. \)

**Proof.** Assume that \( \phi : M \to M_1 \oplus \cdots \oplus M_m \) is an isomorphism. Then the compose of the canonical projections and injections with \( \phi \) and \( \phi^{-1} \) satisfy (i), (ii) and (iii).

Conversely, if there exist \( A \)-linear maps satisfying (i), (ii) and (iii), we can defined \( \phi : M \to M_1 \oplus \cdots \oplus M_m \) and \( \psi : M_1 \oplus \cdots \oplus M_m \to M \) by:

\[
\phi(x) = (\pi_k(x))_{k=1,\ldots,m},
\]

\[
\psi((x_k)_{k=1,\ldots,m}) = \iota_1(x_1) + \cdots + \iota_m(x_m).
\]

These are \( A \)-linear maps which, by (i), (ii) and (iii), satisfy \( \phi \circ \psi = id_{M_1 \oplus \cdots \oplus M_m} \) and \( \psi \circ \phi = id_M \). Hence, \( \phi \) and \( \psi \) establish an isomorphism of \( A \)-modules \( M \simeq M_1 \oplus \cdots \oplus M_m \).

If \( \{N_i\}_{i \in I} \) is a family of submodules of an \( A \)-module \( M \), we say that \( M \) is the **direct sum of the submodules** \( \{N_i\}_{i \in I} \), if the map \( \bigoplus_{i \in I} N_i \to M \), \( (v_i) \mapsto \sum_i v_i \), is an isomorphism. In this case, we write \( M = \bigoplus_{i \in I} N_i \). We have the following result which characterizes when an \( A \)-module is a direct sum of submodules. We leave the proof as an exercise:
Proposition 6.1.9. Let $M$ be an $A$-module and $\{M_i\}_{i \in I}$ a family of submodules. Then $M = \bigoplus_{i \in I} M_i$ if and only if:

(i) $M = \sum_{i \in I} M_i$;

(ii) $M_j \cap (M_{i_1} + \cdots + M_{i_k}) = \{0\}$ for $j \not\in \{i_1, \ldots, i_k\}$.

The notion of module is also relevant for studying other algebraic structures. One important example is the following:

Definition 6.1.10. Let $R$ be a ring with identity. An algebra over $R$ or $R$-algebra is a ring $A$ such that:

(i) $(A, +)$ is an $R$-module;

(ii) $k(ab) = (ka)b = a(kb)$ for all $k \in R$ and $a, b \in A$.

If $(A, +, \cdot)$ is a division ring, $A$ is called a division algebra.

The notions of subalgebra, homomorphism and isomorphism of $R$-algebras are more or less obvious and we leave it to the reader its definition. An algebra over a field $K$ which, as a vector space over $K$, has finite dimension is called a finite dimensional algebra over $K$.

Examples 6.1.11.

1. A field extension $k \subset K$ can be viewed as an algebra over $k$. For example, the fields $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ are algebras over each of the preceding fields. Similarly, the quaternions $\mathbb{H}$ is an algebra over $\mathbb{Q}$ and $\mathbb{R}$ (what about over $\mathbb{C}$?). These are all division algebras.

2. Let $R$ be a ring with identity. The set $A = M_n(R)$ of all matrices $n \times n$ with entries in $R$ is an algebra over $R$, which fails to be a division algebra. If $R = K$ is a field, $M_n(K)$ is an algebra over $K$ of finite dimension.

3. If $V$ is a vector space over a field $K$, the set $A = \text{End}_K(V)$ of all the linear maps $V \to V$ is an algebra over $K$. Note that $A$ is finite dimensional iff $V$ is finite dimensional: if $\dim V = n$ then $A$ is isomorphic to the algebra $M_n(K)$.

4. If $A$ is a commutative ring with identity, the polynomial ring $A[x_1, \ldots, x_n]$ and the power series ring $A[[x]]$ are algebras over $A$.

Algebras where the product is non-associative are also important but will not be discussed here.

---

4For example, Lie algebras and Jordan algebras are non-associative algebras.
Exercises.

1. Consider the $\mathbb{R}[x]$-module structure on $V = \mathbb{R}^n$ defined by the linear transformation $T(v_1, \ldots, v_n) = (v_n, v_1, \ldots, v_{n-1})$. Determine the elements $v \in \mathbb{R}^n$ such that $(x^2 - 1)v = 0$.

2. Let $\phi : M_1 \to M_2$ be a homomorphism of $A$-modules, and $N_i \subset M_i$ ($i = 1, 2$) submodules such that $\phi(N_1) \subset N_2$. Show that:
   
   (a) There exists one, and only one, homomorphism of $A$-modules $\tilde{\phi} : M_1/N_1 \to M_2/N_2$ for which the following diagram commutes:

   $\begin{array}{ccc}
   M_1 & \xrightarrow{\phi} & M_2 \\
   \pi_1 & & \pi_2 \\
   M_1/N_1 & \xrightarrow{\tilde{\phi}} & M_2/N_2
   \end{array}$

   (b) $\tilde{\phi}$ is an isomorphism if and only if $\text{Im}(\phi) + N_2 = M_2$ and $\phi^{-1}(N_2) \subset N_1$.

3. Let $\{N_i\}_{i \in I}$ be a family of $A$-modules. Show that:
   
   (a) Given an $A$-module $M$ and homomorphisms $\{\phi_i : N_i \to M\}_{i \in I}$, there exists a unique homomorphism $\phi : M \to \prod_{i \in I} N_i$ such that for all $k \in I$ the following diagram commutes:

   $\begin{array}{ccc}
   M & \xrightarrow{\phi} & \prod_{i \in I} N_i \\
   \phi_k & & \pi_k \\
   N_k & \xrightarrow{} & N_k
   \end{array}$

   (b) $\prod_{i \in I} N_i$ is determined up to isomorphism by this property.

4. Let $\{N_i\}_{i \in I}$ be a family of $A$-modules. Show that:
   
   (a) Given an $A$-module $M$ and homomorphisms $\{\phi_i : N_i \to M\}_{i \in I}$, there exists a unique homomorphism $\phi : \bigoplus_{i \in I} N_i \to M$ such that for all $k \in I$ the following diagram commutes:

   $\begin{array}{ccc}
   M & \xrightarrow{\phi} & \bigoplus_{i \in I} N_i \\
   \phi_k & & \iota_k \\
   N_k & \xrightarrow{} & N_k
   \end{array}$

   (b) $\bigoplus_{i \in I} N_i$ is determined up to isomorphism by this property.
5. Let $M$ be an $A$-module and $\{M_i\}_{i \in I}$ a family of submodules of $M$. Show that $M = \bigoplus_{i \in I} M_i$ if and only if the following two conditions hold:

(i) $M = \sum_{i \in I} M_i$;
(ii) $M_j \cap (M_{i_1} + \cdots + M_{i_k}) = \{0\}$ for all $j \notin \{i_1, \ldots, i_k\}.

6. A sequence of homomorphisms of $A$-modules:

$$M_0 \xrightarrow{\phi_1} M_1 \xrightarrow{\phi_2} M_2 \xrightarrow{\cdots} \xrightarrow{\phi_n} M_n,$$

is said to be exact if $\operatorname{Im}(\phi_i) = \operatorname{N}(\phi_{i+1})$, $i = 1, \ldots, n - 1$. Show that:

(a) If $N \subset M$ is a submodule, then one has an exact sequence:

$$0 \xrightarrow{} N \xrightarrow{\iota} M \xrightarrow{\pi} M/N \xrightarrow{} 0$$

(b) If $M_1$ and $M_2$ are $A$-modules, then one has an exact sequence:

$$0 \xrightarrow{} M_1 \xrightarrow{\iota_1} M_1 \oplus M_2 \xrightarrow{\pi_2} M_2 \xrightarrow{} 0$$

7. (Five Lemma) Consider the following commutative diagram of $A$-modules and linear transformations:

$$
\begin{array}{ccccccc}
M_1 & \xrightarrow{\phi_1} & M_2 & \xrightarrow{\phi_2} & M_3 & \xrightarrow{\phi_3} & M_4 & \xrightarrow{\phi_4} & M_5 \\
\downarrow{\phi_1} & & \downarrow{\phi_2} & & \downarrow{\phi_3} & & \downarrow{\phi_4} & & \downarrow{\phi_5} \\
N_1 & \xrightarrow{} & N_2 & \xrightarrow{} & N_3 & \xrightarrow{} & N_4 & \xrightarrow{} & N_5
\end{array}
$$

Show that, if the rows are exact and $\phi_1, \phi_2, \phi_4$ and $\phi_5$ are isomorphisms, then $\phi_3$ is an isomorphism.

8. If $M$ and $N$ are left $A$-modules, $\operatorname{Hom}_A(M, N)$ denotes the set of all $A$-linear transformations $\phi : M \to N$. Show that:

(a) $\operatorname{Hom}_A(M, N)$ is a $\mathbb{Z}$-module;
(b) $\operatorname{Hom}_A(M, A)$ is a right $A$-module;
(c) $\operatorname{End}_A(M) \equiv \operatorname{Hom}_A(M, M)$ is a $A$-algebra, provided $A$ is abelian.

9. Let $M$ be a left $A$-module. The dual of $M$ is the right $A$-module $M^* \equiv \operatorname{Hom}_A(M, A)$ (see previous exercise). Show that:

(a) if $\phi : M \to N$ is $A$-linear, there exists a dual linear transformation (of right $A$-modules) $\phi^* : N^* \to M^*$;
(b) $(\bigoplus_{i \in I} N_i)^* \cong \prod_{i \in I} N_i^*$;
(c) one can have $M \neq \{0\}$ and $M^* = \{0\}$. 
6.2 Linear Independence

Let $M$ be an $A$-module. A set $\emptyset \neq S \subset M$ is called linearly independent if for any finite family $\{v_1, \ldots, v_n\}$ of elements of $S$ and any $a_1, \ldots, a_n \in A$:

$$a_1v_1 + \cdots + a_nv_n = 0 \implies a_1 = \cdots = a_n = 0.$$ 

Otherwise, we call $S$ linearly dependent.

A set $S \subset M$ is called generating if $M = \langle S \rangle$. In this case any element $v \in M$ can be written as linear combination (in general, non-unique) of elements de $S$:

$$v = \sum_{i=1}^{m} a_iv_i, a_i \in A, v_i \in S.$$ 

An $A$-module is said to be of finite type if it has some finite generating subset.

A set $S \subset M$ which is both generating and linearly independent is called a basis of the $A$-module $M$. Given a basis $S$, any element $v \in M$ can be written in a unique way as a linear combination $\sum_{i=1}^{m} a_iv_i, a_i \in A, v_i \in S$.

An $A$-module is said to be free if it admits a basis.

Examples 6.2.1.

1. A vector space always contains a basis (exercise).

2. The abelian group $\mathbb{Z}_n$, viewed as a $\mathbb{Z}$-module, does not admit a basis. In fact, given $g \in \mathbb{Z}_n$, there exists always a $0 \neq m \in \mathbb{Z}$ such that $mg = 0$. Hence in $\mathbb{Z}_n$ every subset is linearly dependent.

3. The abelian group $\mathbb{Z}^m = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}$ is free. A basis is given by $S = \{e_1, \ldots, e_m\}$, where $e_i = (0, \ldots, 1, \ldots, 0)$.

4. If $D$ is an integral domain, the polynomial ring $D[x]$ is free. A basis is given by the monomials $S = \{1, x, \ldots, x^n, \ldots\}$.

5. Any ring $A$ with a unit is a free $A$-module with basis $\{1\}$. The submodules coincide with the ideals of $A$. In particular, a submodule may not be free, and even if it is free, it may have a basis with more than one element.

We say that an $A$-module $M$ is cyclic if it is generated by one element, i.e., if $M = \langle v \rangle$ for some $v \in M$. If $M = \langle v \rangle$ is cyclic, then we have an

\[\text{This is the analogue for modules of the notion of a free group.}\]

\[\text{This is the analogue for modules of the notion of a cyclic group.}\]
homomorphism of $A$-modules, $A \to M$, given by $a \mapsto av$. This homomorphism is surjective and, by the First Isomorphism Theorem, $M \simeq A/\text{ann } v$, where the annihilator of $v$ is the ideal

$$\text{ann } v := \{ a \in A : av = 0 \}.$$

If $\text{ann } v = \{0\}$, then we say that $v$ is a free element, since in this case $M = \langle v \rangle \simeq A$ is free. A non-free element is said to have torsion. The set of elements of $M$ with torsion is denoted by $\text{Tors}(M)$. When $A$ is a domain, $\text{Tors}(M) \subset M$ is a submodule of $M$.

If $X$ is any set and $A$ is a ring, let us associate to each $x \in X$ a copy of $A$ and form the free $A$-module $M = \bigoplus_{x \in X} A$, called the free $A$-module generated by the set $X$. We can represent the elements of $M$ as finite sums $a_1x_1 + \cdots + a_rx_r$, where $x_1, \ldots, x_r \in X$ and $a_i \in A$: by such a sum we mean the element $(a_x)_{x \in X} \in \bigoplus_{x \in X} A$, where $a_{x_1} = a_1, \ldots, a_{x_r} = a_r$ and $a_x = 0$ if $x \neq x_i (i = 1, \ldots, r)$.

The next proposition gives a characterization of free modules. In particular, notice that the free module generated by the set $X$ satisfies the same universal property that characterizes free groups.

**Proposition 6.2.2.** Let $A$ be a ring. For an $A$-module $M$, the following statements are equivalent:

(i) $M$ is free.

(ii) There exists a family $\{N_i\}_{i \in I}$ of cyclic submodules of $M$, with $N_i \simeq A$, such that $M \simeq \bigoplus_{i \in I} N_i$.

(iii) $M \simeq \bigoplus_{j \in J} A$ for some index set $J$.

(iv) There exists a set $X \neq \emptyset$ and a map $\iota : X \to M$ such that the following universal property holds: for any $A$-module $N$ and map $\phi : X \to N$ there exists a unique homomorphism of $A$-modules $\tilde{\phi} : M \to N$ such that the following diagram commutes:

\[
\begin{array}{ccc}
X & \xrightarrow{\iota} & M \\
\downarrow{} & & \downarrow{\tilde{\phi}} \\
\phi & \searrow & N
\end{array}
\]

Proof. We show that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv) $\Rightarrow$ (i).
(i) ⇒ (ii) Assume that \( M \) is free and let \( \{ e_i \}_{i \in I} \) be a basis of \( M \). Then, for each \( i \in I \), \( N_i := \langle e_i \rangle \) is a cyclic submodule of \( M \) isomorphic to \( A \). The map \( \phi : \bigoplus_{i \in I} N_i \to M \), \( (v_i)_{i \in I} \mapsto \sum_{i \in I} v_i \) is an isomorphism of \( A \)-modules.

(ii) ⇒ (iii) Obvious.

(iii) ⇒ (iv) Let \( \psi : \bigoplus_{j \in J} A \to M \) be an isomorphism of \( A \)-modules and \( e_k = (x_j)_{j \in J} \) the element of \( \bigoplus_{j \in J} A \), with \( x_k = 1 \) and \( x_j = 0 \), for \( j \neq k \). Also, let \( X = J \) and consider the map \( \iota : X \to M \) given by \( \iota(j) = \psi(e_j) \). If \( \phi : X \to N \) is any map into an \( A \)-module \( N \), we can define \( \tilde{\phi} : M \to N \) to be the linear transformation \( \psi((x_j)_{j \in J}) \mapsto \sum_j x_j \phi(j) \). Clearly, \( \tilde{\phi} \) makes the diagram above commute. Since \( \{ \psi(e_k) \} \) is a basis of \( M \), \( \tilde{\phi} \) is unique.

(iv) ⇒ (i) We leave as an exercise to check that \( \{ \iota(x) \}_{x \in X} \) is a basis of \( M \).

Let \( M \) be a free \( A \)-module which admits a finite basis \( \{ e_1, \ldots, e_n \} \). The proposition shows that \( M \cong \bigoplus_{i=1}^n A \equiv A^n \). Is it true that any other basis of \( M \) has the same number of elements? In other words, is it true that \( A^n \cong A^m \) implies \( n = m \)? Maybe somewhat surprising, the answer is no, as shown by an exercise at the end of this section.

On the other hand, if \( M \) is a free \( A \)-module which admits an infinite basis we have the following result:

**Proposition 6.2.3.** If an \( A \)-module \( M \) admits a basis with an infinite number of elements, then all the bases of \( M \) have the same cardinality.

**Proof.** Let \( \{ e_i \}_{i \in I} \) and \( \{ f_j \}_{j \in J} \) be bases of \( M \) and assume that \( I \) is infinite. Then:

(a) \( J \) is infinite: Assume, to the contrary, that \( J \) is finite, say \( J = \{ 1, \ldots, m \} \). Then there exist elements \( c_{ji} \in A \) with \( i_l \in I, \ l, j \in J \), such that \( f_j = \sum_{i=1}^m c_{ji} e_{i_l} \). But then \( \{ e_1, \ldots, e_m \} \) is a set generating \( M \). Hence, if \( e_0 \) is another basis element distinct from \( e_{i_l} \) there exist \( a_1, \ldots, a_m \in A \) such that

\[
e_0 = a_1 e_1 + \cdots + a_m e_m,
\]

contradicting the linear independence of the \( \{ e_i \}_{i \in I} \).

(b) \( \exists \phi : I \to \mathcal{P}_{\text{fin}}(J) \times \mathbb{N} \) injective

\[\text{Let } \psi : I \to \mathcal{P}_{\text{fin}}(J) \text{ be the map which to } i \in I \text{ associates } \{ j_1, \ldots, j_m \}, \text{ where the } j_1, \ldots, j_m \text{ are the (unique)}\]

\[\text{We denote by } \mathcal{P}_{\text{fin}}(J) \text{ the set consisting of the finite subsets of } J. \text{ It is shown in the Appendix, that if } J \text{ is infinite, then } \mathcal{P}_{\text{fin}}(J) \text{ has the same cardinality as } J.\]
indices in \( J \) such that:

\[
e_i = a_{j_1} f_{j_1} + \cdots + a_{j_m} f_{j_m} \quad (a_{j_l} \neq 0).
\]

The map \( \psi \) is not injective, but if \( P \in \mathcal{P}_{\text{fin}}(J) \), then \( \psi^{-1}(P) \) is finite (why?). Hence we can order \( \psi^{-1}(P) \). If \( i \in \psi^{-1}(P) \), set \( \phi(i) \equiv (P, \alpha) \), where \( \alpha \) is the ordinal number of \( i \) in \( \psi^{-1}(P) \). Since \( I \) is the disjoint union of the \( \psi^{-1}(P) \), the map \( \phi : I \rightarrow \mathcal{P}_{\text{fin}}(J) \times \mathbb{N} \) is injective.

(c) \( |I| = |J| \): Since \( J \) is infinite, by (b), we have

\[
|I| \leq |\mathcal{P}_{\text{fin}}(J) \times \mathbb{N}| = |\mathcal{P}_{\text{fin}}(J)| = |J|.
\]

Interchanging the roles of \( I \) and \( J \), we conclude that \( |J| \leq |I| \). By the Schröder-Bernstein Theorem, we see that \( |I| = |J| \).

The previous results motivate the following definition:

**Definition 6.2.4.** We say that a ring \( A \) has the INVARIANT DIMENSION PROPERTY if for any free \( A \)-module \( M \), all the bases of \( M \) have the same cardinality. In this case, the common cardinality of the bases of \( M \) is called the DIMENSION of \( M \), denoted \( \dim_A M \) or simply \( \dim M \).

We leave as an exercise to check that any division ring has the invariant dimension property, hence it makes sense to talk about the dimension of a vector space. We also have:

**Proposition 6.2.5.** A commutative ring has the invariant dimension property.

**Proof.** Let \( \{e_1, \ldots, e_n\} \) and \( \{f_1, \ldots, f_m\} \) be bases of a free \( A \)-module \( M \). Then there are elements \( b_{ji}, c_{ij} \in A \), \( i = 1, \ldots, n \), \( j = 1, \ldots, m \) such that

\[
f_j = \sum_i b_{ji} e_i, \quad e_i = \sum_j c_{ij} f_j.
\]

By substitution, we conclude that:

\[
f_j = \sum_{il} b_{ji} c_{il} f_l, \quad e_i = \sum_{jl} c_{ij} b_{jl} e_l.
\]

Since \( \{e_1, \ldots, e_n\} \) and \( \{f_1, \ldots, f_m\} \) are bases of \( M \), if we define the matrices

\[
B = (b_{ji})_{j=1,i=1}^{m,n} \quad \text{and} \quad C = (c_{ij})_{i=1,j=1}^{n,m},
\]

we conclude that \( BC = I_{m \times m} \) and \( CB = I_{n \times n} \).
If $A$ has zero characteristic, since $A$ is commutative, we have\footnote{The trace of a $n \times n$ matrix $A = (a_{ij})$ is defined by $\text{tr} A = \sum_{i=1}^{n} a_{ii} \in A$.}

\[ m = \text{tr}(I_{m\times m}) = \text{tr}(BC) = \text{tr}(CB) = \text{tr}(I_{n\times n}) = n. \]

The first and last equality are valid in characteristic zero. We leave the proof for non-zero characteristic as an exercise. \qed

**Exercises.**

1. Give an example of an $A$-module, non-isomorphic to $A$, where any set with 2 or more elements is linearly dependent.

2. Let $A$ be a commutative ring with unit and $M$ an $A$-module.
   (a) Show that, if $v \in \text{Tors}(M)$, then $\langle v \rangle \subset \text{Tors}(M)$;
   (b) Is $\text{Tors}(M)$ always a submodule of $M$?

3. Let $A$ be a commutative ring. Show that $\text{End}_A(A^n)$ is isomorphic to the ring $M_n(A)$ of $n \times n$ matrices with entries in $A$.

4. Let $M$ be an $A$-module, $X \neq \emptyset$ a set and $\iota : X \to M$ a map satisfying the following property: for any $A$-module $N$ and map $\phi : X \to N$ there exists a unique $A$-linear transformation $\tilde{\phi} : M \to N$ such that $\phi = \tilde{\phi} \circ \iota$. Show that \{\iota(x)\}_{x \in X} is a basis of $M$.

5. Let $V$ be a vector space over a division ring $D$. Show that $V$ has a basis.

6. Show that a division ring $D$ has the invariant dimension property.

7. Let $A$ be a commutative ring. Complete the proof of Proposition 6.2.5 by showing that:
   (a) if $B, C \in M_n(A)$ are $n\times n$ matrices, then $BC = I_{n\times n}$ implies $CB = I_{n\times n}$;
   (b) if $B$ is a $m \times n$ matrix, $C$ is a $n \times m$ matrix, $BC = I_{m\times m}$ and $CB = I_{n\times n}$, then $m = n$.

8. Let $\mathbb{R}^\infty = \bigoplus_{i=1}^{\infty} \mathbb{R}$ (direct sum of $\mathbb{R}$-modules) and let $A = \text{End}(\mathbb{R}^\infty)$ be the ring of all $\mathbb{R}$-linear transformations of $\mathbb{R}^\infty$. Show that $A \simeq A \oplus A$ (as $A$-modules), i.e., that $A$ has a basis with 2 elements, so $A$ does not have the invariant dimension property.

9. Show that any $A$-module is the quotient of a free $A$-module.
6.3 Tensor Products

In this section $A$ denotes a commutative ring. In particular, as we saw before, the free modules have the invariant dimension property.

If $M_1, \ldots, M_r, N$ are $A$-modules, an $A$-multilinear map is a map $\mu : M_1 \times \cdots \times M_r \to N$ which is $A$-linear in each entry:

$$\mu(v_1, \ldots, av_i' + bv_i'', \ldots, v_r) = a\mu(v_1, \ldots, v_i', \ldots, v_r) + b\mu(v_1, \ldots, v_i'', \ldots, v_r).$$

We will denote by $L(M_1, \ldots, M_r; N)$ the set of all $A$-multilinear maps. One checks immediately that $L(M_1, \ldots, M_r; N)$ is an $A$-module for the usual operations of addition and multiplication by scalars:

$$(\mu_1 + \mu_2)(v_1, \ldots, v_r) := \mu_1(v_1, \ldots, v_r) + \mu_2(v_1, \ldots, v_r),$$

$$(a\mu)(v_1, \ldots, v_r) := a\mu(v_1, \ldots, v_r).$$

If $M_1 = \cdots = M_r = M$ we write $L^r(M; N)$ instead of $L(M, \ldots, M; N)$.

**Proposition 6.3.1.** Let $M_1, \ldots, M_r$ be $A$-modules.

(i) There exists an $A$-module $\bigotimes_{i=1}^r M_i \equiv M_1 \otimes \cdots \otimes M_r$ and an $A$-multilinear map $\iota : M_1 \times \cdots \times M_r \to M_1 \otimes \cdots \otimes M_r$ satisfying the following universal property: for any $A$-module $N$ and $A$-multilinear map $\phi : M_1 \times \cdots \times M_r \to N$, there exists a unique $A$-linear map $\tilde{\phi} : M_1 \otimes \cdots \otimes M_r \to N$ which makes the following diagram commute:

$$\begin{array}{ccc}
M_1 \times \cdots \times M_r & \xrightarrow{\iota} & M_1 \otimes \cdots \otimes M_r \\
\phi \downarrow & & \downarrow \tilde{\phi} \\
N
\end{array}$$

(ii) The $A$-module $M_1 \otimes \cdots \otimes M_r$ is determined by the universal property in (i) up to isomorphism.

**Proof.** Let $L$ be the free $A$-module generated by the set $I = M_1 \times \cdots \times M_r$, i.e.,

$$L = \bigoplus_{i \in I} A,$$

*One can also define tensor products of modules over non-commutative rings. The definition is similar, but one needs to distinguish between left and right modules.*
where in this sum there exists a term for each \((v_1, \ldots, v_r) \in M_1 \times \cdots \times M_r\). Denoting by \(R\) the submodule of \(L\) generated by the elements of the form

\[
(6.3.1) \quad (v_1, \ldots, av'_i + bv''_i, \ldots, v_r) - a(v_1, \ldots, v'_i, \ldots, v_r) - b(v_1, \ldots, v''_i, \ldots, v_r),
\]

we let \(M_1 \otimes \cdots \otimes M_r\) denote the quotient module \(L/R\). We define \(\iota : M_1 \times \cdots \times M_r \rightarrow L\) with the quotient map \(L \rightarrow L/R\).

If \(\phi : M_1 \times \cdots \times M_r \rightarrow N\) is an \(A\)-multilinear map, then we induce a linear map \(\bar{\phi} : L \rightarrow N\) as follows: if \(\oplus_{i \in I} a_i (v_{i1}, \ldots, v_{ir}) \in L\), then

\[
\bar{\phi}(\oplus_{i \in I} a_i (v_{i1}, \ldots, v_{ir})) = \sum_{i \in I} a_i \phi(v_{i1}, \ldots, v_{ir}).
\]

This map is well-defined, since by definition of the direct sum only a finite number of the \(a_i\)'s is non-zero. Also, it is immediate to check that \(\bar{\phi}\) is \(A\)-linear.

Since \(\phi\) is \(A\)-multilinear, the linear map \(\bar{\phi}\) vanishes on elements of the form \((6.3.1)\), hence on the submodule \(R\). Passing to the quotient, we obtain an \(A\)-linear transformation \(\bar{\phi} : M_1 \otimes \cdots \otimes M_r \rightarrow N\) which, by definition, satisfies \(\phi = \bar{\phi} \circ \iota\).

Finally, let \((\bigotimes_{i=1}^r M_i)'\) be an \(A\)-module and \(\iota' : M_1 \times \cdots \times M_r \rightarrow (\bigotimes_{i=1}^r M_i)'\) an \(A\)-multilinear map satisfying the universal property expressed in (i). We then obtain commutative diagrams:

\[
\begin{array}{ccc}
M_1 \times \cdots \times M_r & \overset{\iota}{\longrightarrow} & \bigotimes_{i=1}^r M_i \\
\downarrow \iota' & & \downarrow \iota' \\
(\bigotimes_{i=1}^r M_i)' & \overset{\iota}{\longrightarrow} & \bigotimes_{i=1}^r M_i \\
\end{array}
\]

The compose of the \(A\)-linear maps \(\iota\) and \(\iota'\) make the following diagrams commute:

\[
\begin{array}{ccc}
M_1 \times \cdots \times M_r & \overset{\iota}{\longrightarrow} & \bigotimes_{i=1}^r M_i \\
\downarrow \iota \circ \iota' & & \downarrow \iota' \circ \iota' \\
(\bigotimes_{i=1}^r M_i)' & \overset{\iota}{\longrightarrow} & \bigotimes_{i=1}^r M_i \\
\end{array}
\]

\[
\begin{array}{ccc}
M_1 \times \cdots \times M_r & \overset{\iota'}{\longrightarrow} & (\bigotimes_{i=1}^r M_i)' \\
\downarrow \iota \circ \iota' & & \downarrow \iota' \circ \iota' \\
(\bigotimes_{i=1}^r M_i)' & \overset{\iota}{\longrightarrow} & (\bigotimes_{i=1}^r M_i)' \\
\end{array}
\]
6.3. TENSOR PRODUCTS

Since the identity transformations $\text{id}_{M_1 \otimes \cdots \otimes M_r}$ and $\text{id}_{(M_1 \otimes \cdots \otimes M_r)'}$ also make the diagrams commute, the uniqueness requirement in the universal property implies that $\tilde{\iota} \circ \tilde{\iota}' = \text{id}_{M_1 \otimes \cdots \otimes M_r}$ and $\tilde{\iota}' \circ \tilde{\iota} = \text{id}_{(M_1 \otimes \cdots \otimes M_r)'}$. Hence, these maps give an isomorphism of $A$-modules $\bigotimes_{i=1}^r M_i \simeq (\bigotimes_{i=1}^r M_i)'$.

The $A$-module $M_1 \otimes \cdots \otimes M_r$ is called the tensor product of the modules $M_1, \ldots, M_r$. If $(v_1, \ldots, v_r) \in M_1 \times \cdots \times M_r$, the image $\iota(v_1, \ldots, v_r) \in M_1 \otimes \cdots \otimes M_r$ is denoted by $v_1 \otimes \cdots \otimes v_r$. Notice that in this notation, the following identity holds:

$$v_1 \otimes \cdots \otimes (av'_i + bv''_i) \otimes \cdots \otimes v_r = a(v_1 \otimes \cdots \otimes v'_i \otimes \cdots \otimes v_r) + b(v_1 \otimes \cdots \otimes v''_i \otimes \cdots \otimes v_r).$$

An element of $M_1 \otimes \cdots \otimes M_r$ can be written as a linear combination of elements of the form $v_1 \otimes \cdots \otimes v_r$, since we saw above that the set of all such elements is a generating set. However, this representation is very far from being unique, since $\iota$ is not injective.

**Example 6.3.2.**

In the tensor product (over $\mathbb{Z}$) of $\mathbb{Z}_2$ with $\mathbb{Z}_4$, we have the following relations:

$$0 = 0(n \otimes m) = 0 \otimes m = n \otimes 0,$$
$$1 \otimes 2 = 2(1 \otimes 1) = 2 \otimes 1 = 0 \otimes 1 = 0,$$
$$1 \otimes 1 = 1 \otimes 1 + 1 \otimes 2 = 1 \otimes 3.$$

Hence, the only element in $\mathbb{Z}_2 \otimes \mathbb{Z}_4$ which is possibly non-zero is $1 \otimes 1 = 1 \otimes 3$. But since the bilinear map $\mathbb{Z}_2 \times \mathbb{Z}_4 \to \mathbb{Z}_2$, $(n, m) \mapsto nm$, maps $(1, 1) \mapsto 1 \neq 0$, we conclude that $1 \otimes 1 \neq 0$, and then it follows that $\mathbb{Z}_2 \otimes \mathbb{Z}_4 \simeq \mathbb{Z}_2$.

Using the definition of tensor product we obtain immediately:

**Proposition 6.3.3.** (Properties of $\otimes$) Given $A$-modules $M$, $N$, $P$ and $\{M_i\}_{i \in I}$, there exist isomorphisms of $A$-modules:

(i) $M \otimes N \otimes P \simeq (M \otimes N) \otimes P \simeq M \otimes (N \otimes P)$ such that $v \otimes w \otimes z \leftrightarrow (v \otimes w) \otimes z \leftrightarrow v \otimes (w \otimes z)$ for any elements $v \in M$, $w \in N$, and $z \in P$;

(ii) $M \otimes N \simeq N \otimes M$ such that $v \otimes w \leftrightarrow w \otimes v$, for any $v \in M$ and $w \in N$;

(iii) $\bigoplus_{i \in I} M_i \otimes N \simeq \bigoplus_{i \in I}(M_i \otimes N)$ such that $(v_i)_{i \in I} \otimes w \leftrightarrow (v_i \otimes w)_{i \in I}$, for any $v_i \in M_i$, $w \in N$. 

As we saw in the example above, in general the tensor product $M \otimes N$ involves a large number of relations between elements of the form $v \otimes w$. However, in the case of free modules, there are only the “obvious relations” among these elements, as shown by the following proposition:

**Proposition 6.3.4.** Let $M$ and $N$ be free $A$-modules, with bases $\{v_i\}_{i \in I}$ and $\{w_j\}_{j \in J}$. Then $M \otimes N$ is free, with basis $\{v_i \otimes w_j\}_{(i,j) \in I \times J}$.

**Proof.** Obviously, $\{v_i \otimes w_j\}_{(i,j) \in I \times J}$ is a generating set. In order to see that it is also linearly independent, consider the map $\phi : M \times N \to \bigoplus_{(i,j) \in I \times J} A$ defined by:
\[
\phi\left(\sum_{i \in I} a_i v_i, \sum_{j \in J} b_j w_j\right) = (a_i b_j)_{(i,j) \in I \times J}.
\]
Since $\phi$ is $A$-bilinear, there exists an $A$-linear map $\tilde{\phi} : M \otimes N \to \bigoplus_{(i,j) \in I \times J} A$ such that $\phi = \tilde{\phi} \circ I$.
\[
\tilde{\phi}(v_k \otimes w_l) = \phi(v_k, w_l) = (e_{kl})_{(i,j) \in I \times J}.
\]
where $(e_{kl})_{ij} = 1$, if $(k, l) = (i, j)$, and $(e_{kl})_{ij} = 0$, otherwise. The elements $(e_{kl})_{(i,j) \in I \times J}$ form a basis of $\bigoplus_{(i,j) \in I \times J} A$, hence, the set $\{v_i \otimes w_j\}_{(i,j) \in I \times J}$ is linearly independent. \(\square\)

**Corollary 6.3.5.** Let $M$ and $N$ be free $A$-modules with $\dim M = m$ and $\dim N = n$. Then $\dim(M \otimes N) = mn$.

If $\phi_i : M_i \to N_i$, $i = 1, \ldots, r$ are homomorphisms of $A$-modules, we have the homomorphism $T(\phi_1, \ldots, \phi_r) : M_1 \otimes \cdots \otimes M_r \to N_1 \otimes \cdots \otimes N_r$ defined as follows:
\[
T(\phi_1, \ldots, \phi_r)(v_1 \otimes \cdots \otimes v_r) = \phi_1(v_1) \otimes \cdots \otimes \phi_r(v_r).
\]
Since the right hand side defines a multilinear expression in the $v_1, \ldots, v_r$ the universal property of tensor product shows that this map is well defined.

In the next results we use the fact that $A$ is commutative to write $\text{Hom}_A(M, N)$, $\text{End}_A(M)$ and $M^*$ as left $A$-modules (see Exercises 8 and 9 in Section 6.1). Note that these are all free of finite dimension, provided $M$ and $N$ are free of finite dimension.

**Proposition 6.3.6.** Let $M_i$, $N_i$, $i = 1, \ldots, r$, be free $A$-modules of finite dimension. There exists an isomorphism:
\[
\text{Hom}_A(M_1, N_1) \otimes \cdots \otimes \text{Hom}_A(M_r, N_r) \simeq \text{Hom}_A(M_1 \otimes \cdots \otimes M_r, N_1 \otimes \cdots \otimes N_r),
\]
which maps the element $\phi_1 \otimes \cdots \otimes \phi_r$ to $T(\phi_1, \ldots, \phi_r)$. 


Proof. By the associativity property of the tensor product, it is enough to prove the case \( r = 2 \). So let \( M_1, M_2, N_1 \) and \( N_2 \) be free \( A \)-modules with bases \( \{ v'_1, \ldots, v'_{m_1} \}, \{ v''_1, \ldots, v''_{m_2} \}, \{ w'_1, \ldots, w'_{n_1} \} \) and \( \{ w''_1, \ldots, w''_{n_2} \} \), respectively. Define a basis \( \{ \phi_{ij} \} \) of \( \text{Hom}_A(M_1, N_1) \) and \( \{ \psi_{kl} \} \) of \( \text{Hom}_A(M_2, N_2) \) by the formulas:

\[
\phi_{ij}(v'_a) = \begin{cases} w'_j & \text{if } a = i, \\ 0 & \text{if } a \neq i, \end{cases} \quad \psi_{kl}(v''_b) = \begin{cases} w''_l & \text{if } b = k, \\ 0 & \text{if } b \neq k. \end{cases}
\]

By the previous proposition, \( \{ \phi_{ij} \otimes \psi_{kl} \} \) is a basis of \( \text{Hom}_A(M_1, N_1) \otimes \text{Hom}_A(M_2, N_2) \). On the other hand, we see that

\[
T(\phi_{ij}, \psi_{kl})(v'_a \otimes v''_b) = \begin{cases} w'_j \otimes w''_l & \text{if } (a, b) = (i, k), \\ 0 & \text{if } (a, b) \neq (i, k). \end{cases}
\]

Hence, \( \{ T(\phi_{ij}, \psi_{kl}) \} \) is a basis of \( \text{Hom}(M_1 \otimes N_1, M_2 \otimes N_2) \), and we conclude that there exists an isomorphism of \( A \)-modules such that \( \phi \otimes \psi \mapsto T(\phi, \psi) \).

This result shows that in the case of free \( A \)-modules of finite dimension we can write \( \phi_1 \otimes \cdots \otimes \phi_r \) instead of \( T(\phi_1, \ldots, \phi_r) \), without any ambiguity.

**Corollary 6.3.7.** Let \( M \) and \( N \) be free \( A \)-modules of finite dimension. There exist isomorphisms:

(i) \( \text{End}_A(M) \otimes \text{End}_A(N) \cong \text{End}_A(M \otimes N) \);

(ii) \( M^* \otimes N^* \cong (M \otimes N)^* \).

These isomorphisms together with the following isomorphism lead to a more concrete interpretation of the tensor product of free \( A \)-modules of finite dimension.

**Corollary 6.3.8.** Let \( M \) and \( N \) be free \( A \)-modules of finite dimension. There exists an isomorphism

\[
M^* \otimes N \cong \text{Hom}_A(M, N)
\]

mapping an element \( l \otimes w \) to the homomorphism \( \phi_{l,w} \) given by \( \nu \mapsto l(\nu)w \).
Proof. If \( \{v_1, \ldots, v_m\} \) is a basis for \( M \), let \( \{l_1, \ldots, l_m\} \) be the dual basis for \( M^* \) defined by:

\[
l_i(v_j) = \begin{cases} 
1 & \text{if } j = i, \\
0 & \text{if } j \neq i. 
\end{cases}
\]

Then, if \( \{w_1, \ldots, w_n\} \) is a basis for \( N \), the tensor products \( \{l_i \otimes w_k\} \) form a basis for \( M^* \otimes N \). On the other hand, the \( A \)-linear transformations \( \phi_{l_i, w_k} \in \text{Hom}_A(M, N) \) satisfy

\[
\phi_{l_i, w_k}(v_j) = l_i(v_j)w_k = \begin{cases} 
w_k & \text{if } j = i, \\
0 & \text{if } j \neq i,
\end{cases}
\]

hence the \( \{\phi_{l_i, w_k}\} \) form a basis of \( \text{Hom}_A(M, N) \), and there exists an isomorphism \( M^* \otimes N \simeq \text{Hom}_A(M, N) \), \( l \otimes w \mapsto \phi_{l, w} \).

This leads to the following interpretation of the tensor product of free modules of finite dimension, which is often used in Geometry to characterize tensor fields and differential forms:

**Corollary 6.3.9.** Let \( M \) and \( N \) be free \( A \)-modules of finite dimension. There exists an isomorphism

\[
M \otimes N \simeq L(M^*, N^*; A).
\]

Proof. The diagram in the universal property of tensor product:

\[
\begin{array}{ccc}
M \times N & \xrightarrow{i} & M \otimes N \\
\downarrow{\phi} & & \downarrow{\bar{\phi}} \\
A & \xrightarrow{\phi} & \text{Hom}_A(M, N) \\
\end{array}
\]

determines an isomorphism:

\[
L(M, N; A) \simeq (M \otimes N)^*, \quad \phi \mapsto \bar{\phi}.
\]

In general, this isomorphism does not characterize the tensor product, since one can have \( M \otimes N \neq \{0\} \) and \( (M \otimes N)^* = \{0\} \). However, when \( M \) and \( N \) are both free of finite dimension, we obtain:

\[
L(M^*, N^*; A) \simeq (M^* \otimes N^*)^* \simeq (M^*)^* \otimes (N^*)^* \simeq M \otimes N.
\]

\[\square\]
It is well known that any vector space $V$ over $\mathbb{R}$ can be seen as a vector space over $\mathbb{C}$. One can use tensor products to generalize this construction and extend the ring of scalars of a given module, as follows.

Let $A$ be a ring and $\tilde{A}$ an extension of $A$ (i.e., $A$ is a subring of $\tilde{A}$). We can view $\tilde{A}$ as an $A$-module: if $a \in A$ and $b \in \tilde{A}$, then the product $ab$ is defined using the multiplication in $\tilde{A}$. Hence, if $M$ is any $A$-module, we can form the $A$-module $M_{\tilde{A}} = \tilde{A} \otimes_A M$ \footnote{Whenever one works with more than one ring at the same time, it is convenient to indicate the ring as a subscript in the symbol of tensor product, so that it is clear in which ring one is taking the tensor product.} and define an operation of $\tilde{A}$ on $M_{\tilde{A}}$ by:

$$b(c \otimes v) := (bc) \otimes v.$$  

One checks easily that $M_{\tilde{A}}$ with this new scalar multiplication operation is an $\tilde{A}$-module. We say that $M_{\tilde{A}}$ is obtained from $M$ by extension of the ring of scalars. If $\phi : M \to N$ is an $A$-linear map, then we obtain an $\tilde{A}$-linear map $\tilde{\phi} : M_{\tilde{A}} \to N_{\tilde{A}}$ by setting:

$$\tilde{\phi}(c \otimes v) := c \otimes \tilde{\phi}(v).$$

**Proposition 6.3.10.** If $M$ is a free $A$-module, then $M_{\tilde{A}}$ is a free $\tilde{A}$-module and $\dim_A M = \dim_{\tilde{A}} M_{\tilde{A}}$.

**Proof.** If $M \simeq \bigoplus_{i \in I} A$, then

$$M_{\tilde{A}} = \tilde{A} \otimes_A M$$

$$\simeq \tilde{A} \otimes_A \left( \bigoplus_{i \in I} A \right)$$

$$\simeq \bigoplus_{i \in I} (\tilde{A} \otimes_A A) \simeq \bigoplus_{i \in I} \tilde{A},$$

where the last isomorphism is obtained from the isomorphism $\tilde{A} \to \tilde{A} \otimes_A A$, $a \mapsto a \otimes 1$.

**Examples 6.3.11.**

1. If $V$ is a vector space over $\mathbb{R}$, then $V_{\mathbb{C}} = V \otimes_{\mathbb{R}} \mathbb{C}$ (sometimes called the complexification of $V$) is a vector space over $\mathbb{C}$. If $\{v_1, \ldots, v_n\}$ is a basis of $V$ over $\mathbb{R}$, then $\{1 \otimes v_1, \ldots, 1 \otimes v_n\}$ is a basis of $V_{\mathbb{C}}$ over $\mathbb{C}$. Hence, if $V \simeq \mathbb{R}^n$ then $V_{\mathbb{C}} \simeq \mathbb{C}^n$.

2. If one extends the ring of scalars of the $\mathbb{Z}$-module $\mathbb{Z}$ to $\mathbb{Q}$, one obtains a $\mathbb{Q}$-module isomorphic a $\mathbb{Q}$.

3. On the other hand, if one extends the ring of scalars of the $\mathbb{Z}$-module $\mathbb{Z}_n$ to $\mathbb{Q}$, one obtains a trivial $\mathbb{Q}$-module (exercise).
There are many other constructions where tensor products, Hom and duality play a crucial role. A few of them are covered in the exercises.

**Exercises.**

1. Check the basic properties of tensor products given in Proposition 6.3.3.

2. Let $\rho_1 : G \to GL(V_1)$ and $\rho_2 : G \to GL(V_2)$ be representations of the same group $G$ in vector spaces $V_1$ and $V_2$. Show that there exists exactly one representation $\rho : G \to GL(V_1 \otimes V_2)$ satisfying the following property:
   
   
   $\rho(g)(v_1 \otimes v_2) = \rho_1(g)(v_1) \otimes \rho_2(g)(v_2)$.

3. Show that $\mathbb{Z}_m \otimes_{\mathbb{Z}} \mathbb{Z}_n \simeq \mathbb{Z}_q$. What is the expression of $q$ in terms of $m$ and $n$?

4. Show that $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}_n$ is trivial.

5. Show that if

   
   
   
   $0 \longrightarrow M_1 \longrightarrow M_2 \longrightarrow M_3 \longrightarrow 0$

   
   
   is a short exact sequence of $A$-modules (see Exercise 6, in Section 6.1) and $N$ is an $A$-module, then the sequence of $A$-modules

   
   
   
   $M_1 \otimes N \longrightarrow M_2 \otimes N \longrightarrow M_3 \otimes N \longrightarrow 0$

   
   
   is also exact. Give an example that shows that the first map in this sequence may fail to be injective.

6. Let $M$ be an $A$-module, and let $R$ be the submodule of $\bigotimes_{i=1}^r M$ generated by elements of the form

   
   $v_1 \otimes \cdots \otimes v_r, \quad v_i = v_j$ for some $i, j$ ($i \neq j$)

   
   Denote by $\bigwedge^r M$ the quotient module $\bigotimes_{i=1}^r M/R$, and by $v_1 \wedge \cdots \wedge v_r$ the image of $v_1 \otimes \cdots \otimes v_r$ in $\bigwedge^r M$. Show that:

   (i) the $A$-multilinear map $\iota : M \times \cdots \times M \to M \wedge \cdots \wedge M$, $(v_1, \ldots, v_r) \mapsto v_1 \wedge \cdots \wedge v_r$ is alternating, i.e.,

   
   
   $\iota(v_{\sigma(1)}, \ldots, v_{\sigma(r)}) = \text{sgn } \sigma \cdot \iota(v_1, \ldots, v_r), \quad \forall \sigma \in S_r$.

   (ii) if $\phi : M \times \cdots \times M \to N$ is any $A$-multilinear alternating map, there exists a unique $A$-linear map $\hat{\phi} : M \wedge \cdots \wedge M \to N$ making the following diagram commute:
6.3. TENSOR PRODUCTS

(iii) The $A$-module $M \wedge \cdots \wedge M$ is determined by the universal property in (ii) up to isomorphism.

(iv) If $M$ is free of finite dimension $n$, then $\bigwedge^r M$ is free of dimension $\binom{n}{r}$ if $1 \leq r \leq n$, and of dimension 0 if $r > n$.

(v) If $M$ is free of finite dimension, then $\bigwedge^r M^* \simeq A^r(M)$ (the module of all alternating $A$-multilinear maps $\varphi : M \times \cdots \times M \to A$).

7. Let $\{M_i\}_{i \in I}$ be a family of $A$-modules where $I$ is a partial ordered set which satisfies the following condition $\text{[1]}$:

$$\forall i, j \in I, \exists k \in I : i \leq k \text{ and } j \leq k.$$ 

Assume also that for any pair $i, j \in I$ with $i \leq j$ there exists an $A$-linear map $\phi^k_i : M_i \to M_j$ such that whenever $i \leq j \leq k$ one has:

$$\phi^k_j \circ \phi^k_i = \phi^k_i, \quad \phi^1_i = \text{id}.$$ 

Show that:

(a) there exists an $A$-module $M$ and $A$-linear maps $\phi_i : M_i \to M$ satisfying the following universal property: if $N$ is an $A$-module and $\varphi_i : M_i \to N$ are $A$-linear maps such that $\varphi_j \circ \phi^k_i = \varphi_i$, there exists a unique $A$-linear map $\varphi : M \to N$ making the following diagram commute:

(b) if $\phi_i(v) = 0$ for some $v \in M_i$, then there exists $j \geq i$ such that $\phi^k_j(v) = 0$;
(c) if $N$ is an $A$-module, then $\varprojlim (M_i \otimes N) = (\varprojlim M_i) \otimes N$;
(d) if $M_1 \subset M_2 \subset \cdots \subset M_i \subset \cdots$ are $A$-modules, find $\varprojlim M_i$.

8. Define inverse limit of a directed family of $A$-modules, and show that it is characterized by a universal property analogous to the direct limit where the sense of the arrows is inverted.

$\text{[1]}$ A partial ordered set which satisfies this condition is said to be directed or filtered.
6.4 Modules over Integral Domains

In this section we assume that rings are commutative with a unit and the
cancellation law holds, i.e., they are integral domains. In linear algebra this
property of the ring leads to:

**Proposition 6.4.1.** Let $M$ be a module over an integral domain $D$. Then $\text{Tors}(M)$ is a submodule of $M$.

**Proof.** Recall that $\text{Tors}(M) = \{v \in M : \text{existe } a \in D \text{ com } av = 0 \text{ and } a \neq 0\}$. Hence, if $v_1, v_2 \in \text{Tors}(M)$, then there exist $a_1, a_2 \in D$, non-zero, such that $a_1v_1 = 0$ and $a_2v_2 = 0$. If $d_1, d_2 \in D$, then

$$a_1a_2(d_1v_1 + d_2v_2) = a_2d_1a_1v_1 + a_1d_2a_2v_2 = 0,$$

with $a_1a_2 \neq 0$, for if $a_1a_2 = 0$ then the cancellation law would imply that $a_1 = 0$ or $a_2 = 0$. Hence, $d_1v_1 + d_2v_2 \in \text{Tors}(M)$. 

One calls $\text{Tors}(M)$ torsion submodule of $M$. If $M = \text{Tors}(M)$, then one says that $M$ is a torsion module. If $\text{Tors}(M) = 0$, i.e., if all elements of $M$ are free, then one says that $M$ as module free of torsion.

The next proposition gives elementary properties of the torsion submodule and its proof is left as an exercise.

**Proposition 6.4.2.** Let $D$ be an integral domain.

(i) If $\phi : M_1 \to M_2$ is a $D$-linear transformation, then

$$\phi(\text{Tors}(M_1)) \subset \text{Tors}(M_2).$$

If $\phi$ is injective, then $\phi(\text{Tors}(M_1)) = \text{Tors}(M_2) \cap \text{Im}(\phi)$. If $\phi$ is surjective with $\ker(\phi) \subset \text{Tors}(M_1)$, then $\phi(\text{Tors}(M_1)) = \text{Tors}(M_2)$.

(ii) If $M$ is a $D$-module, then $M/\text{Tors}(M)$ is $D$-module free of torsion.

(iii) If $\{M_i\}_{i \in I}$ is a family of $D$-modules, then

$$\text{Tors}(\bigoplus_{i \in I} M_i) = \bigoplus_{i \in I} \text{Tors}(M_i).$$

The following examples show that one should not confuse the concepts of a free module and free of torsion.
Examples 6.4.3.

1. If $M$ is a free module over an integral domain, then $\text{Tors}(M) = 0$ (exercise) and $M$ is free of torsion.

2. The $\mathbb{Z}$-module $\mathbb{Q}$ is free of torsion, but $\mathbb{Q}$ is not a free $\mathbb{Z}$-module.

3. The modules $\mathbb{Z}_n$ are torsion $\mathbb{Z}$-modules.

4. If $V$ is a finite dimension vector space over $K$ and $T : V \to V$ is a linear transformation, then $V$ is a torsion $K[x]$-module (exercise).

Let $M$ be a module over an integral domain $D$ and denote by $K = \text{Frac}(D)$ the field of fractions of $D$. Since $K$ is an extension of $D$, we can extend the ring of scalars of $M$ to $K$, resulting in a vector space $M_K$ over $K$. This vector space reflects the properties of $M$ up to its torsion.

Proposition 6.4.4. Let $M$ be a $D$-module, $K = \text{Frac}(D)$, and $\phi : M \to M_K$ the $D$-linear transformation $v \to 1 \otimes v$. Then:

(i) Every element of $M_K$ is of the form $\frac{1}{d} \phi(v)$, where $0 \neq d \in D$ and $v \in M$.

(ii) The kernel of $\phi$ is the torsion submodule $\text{Tors}(M)$.

Proof. In order to prove (i), observe that the module $M_K$ is generated by all elements $k \otimes v$, with $k \in \text{Frac}(D)$ and $v \in M$. Hence, if $w \in M_K$, then:

$$w = \sum_{i=1}^{n} k_i \otimes v_i = \sum_{i=1}^{n} \frac{a_i}{b_i} \otimes v_i.$$ 

Denoting by $d$ the product of the $b_i$’s, there exist $c_i \in D$ such that $\frac{a_i}{b_i} = \frac{c_i}{d}$, hence:

$$w = \frac{1}{d} \otimes \left( \sum_{i=1}^{n} c_i v_i \right) = \frac{1}{d} \phi(v).$$

The proof of (ii) is left for the exercises. \qed

We call the vector space $M_K$ over $K = \text{Frac}(D)$ the vector space associated with the $D$-module $M$. The previous proposition suggests the following definition:

Definition 6.4.5. If $M$ is a $D$-module and $S \subset M$, we call the rank of $S$ the dimension of the linear subspace $M_K$ generated by $\phi(S)$. In particular, the rank of $M$ is the dimension $\dim M_K$. 
Recall that $M$ is a finite type if it is finitely generated. The proposition above shows that:

**Corollary 6.4.6.** Every $D$-module $M$ of finite type has finite rank.

**Proof.** If $S$ is a finite set generating $M$, then $\phi(S)$ is finite and contains a basis of $M_K$, hence $\dim M_K < \infty$. In particular, $M$ has finite rank. \hfill $\Box$

Notice that the rank of a $D$-module is an invariant: if $M_1 \cong M_2$, then $M_1$ and $M_2$ have the same rank. The converse in general fails, so the rank is certainly not a complete invariant.

**Examples 6.4.7.**

1. Since $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Q} = \mathbb{Q}$ the rank of $\mathbb{Q}$, as a $\mathbb{Z}$-module, is 1. Since $\mathbb{Q}$ is not finitely generated over $\mathbb{Z}$, the converse to the corollary does not hold.

2. Since $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}_n = \{0\}$, the rank of $\mathbb{Z}_n$ is zero. More generally, if $M$ is a torsion module, then its rank is zero.

**Corollary 6.4.8.** Let $M$ be a $D$-module. Then $\{e_i\}_{i \in I} \subseteq M$ is a linearly independent family over $D$ if and only if $\{1 \otimes e_i\}_{i \in I} \subseteq M_K$ is a linearly independent family over $K$.

**Proof.** If $\{e_i\}_{i \in I} \subseteq M$ is a linearly independent family, the submodule $N = \bigoplus_{i \in I} De_i$ is free of torsion, hence the restriction of $\phi : M \to M_K$ to $N$ is injective. \hfill $\Box$

It follows from this corollary that if $M$ is a free $D$-module then its rank equals its dimension. On the other hand, if $M$ is a free $D$-module then a submodule $N \subseteq M$ may fail to be free. In fact, we have:

**Proposition 6.4.9.** If $D$ is an integral domain such that for any free $D$-module $M$ the submodules $N \subseteq M$ are all free, then $D$ is a PID.

**Proof.** Since $M = D$ is a free $D$-module, by assumption its submodules, i.e., the ideals $I \subseteq D$, must be free. Let $I \subseteq D$ be an ideal. A basis of $I$ contains only one element, since any two elements $a, b \in I$ are linearly dependent:

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

If $\{d\}$ is a basis of $I$, then $I = \langle d \rangle$, so $I$ is a principal ideal. \hfill $\Box$

Actually, the principal ideal domains are characterized by the property expressed in this proposition, since we have the following result:
Theorem 6.4.10. If $D$ is a PID and $M$ is a free $D$-module, then any submodule $N \subset M$ is free and $\dim N \leq \dim M$.

Proof. Let $\{e_i\}_{i \in I}$ be a basis for $M$ over $D$ and $\{0\} \neq N \subset M$ a submodule. If $J \subset I$, consider an ordered pair $(N_J, B_J)$, where

$$N_J = N \cap \left( \bigoplus_{j \in J} D e_j \right),$$

and $B_J = \{f_j\}_{j \in J'}$ is a basis of $N_J$, with $J' \subset J$. We denote by $P$ the set formed by all such ordered pairs. In $P$ we have the partial order relation given by:

$$(N_J, B_J) \leq (N_{J'}, B_{J'}) \iff J \subset J' \text{ and } B_J \subset B_{J'}.$$

We claim that we can apply Zorn’s Lemma to $(P, \leq)$:

(i) $P$ is non-empty: Since $N \neq \{0\}$, there exists $J_0 = \{j_1, \ldots, j_n\} \subset I$ such that $N \cap \bigoplus_{i=1}^{n-1} D e_{j_i} = \{0\}$ and $N \cap \bigoplus_{i=1}^{n} D e_{j_i} \neq \{0\}$. The set

$$\{a \in D : ae_{j_n} + \sum_{i=1}^{n-1} b_i e_{j_i} \in N\}$$

is an ideal in $D$, hence takes the form $\langle d_0 \rangle$. It follows that there exists $f_0 = d_0 e_{j_n} + \sum_{i=1}^{n-1} b_i e_{j_i} \in N$. If $v = ae_{j_n} + \sum_{i=1}^{n-1} b_i e_{j_i} \in N$, we must have $a = kd_0$ and

$$v - kf_0 = \sum_{i=1}^{n-1} (b_i - kb_0) e_{j_i} \in N \cap \bigoplus_{i=1}^{n-1} D e_{j_i} = \{0\}.$$

We conclude that $B = \{f_0\}$ is a basis of $N_{\{j_0\}}$, so $P$ is non-empty.

(ii) Every ascending chain $\{(N_{J_0}, B_{J_0})\}_{\alpha \in A}$ in $(P, \leq)$ has an upper bound: It is enough to consider the pair $(\bigcup_{\alpha \in A} N_{J_\alpha}, \bigcup_{\alpha \in A} B_{J_\alpha})$. Zorn’s Lemma applied to $(P, \leq)$ gives a maximal element $(N_J, B_J)$. To finish the proof it is enough to show that $J = I$, for this implies that $N_j = N$, so that $B_J$ is a basis for $N$.

Assume that $I - J \neq \emptyset$. Then there exists $l \in I - J$ and $a \in D$ such that

$$ae_l + v_a \in N \text{ for some } v_a \in \bigoplus_{j \in J} D e_j.$$
The elements \( a \in D \) which satisfy \((6.4.1)\) form an ideal, which must be principal: \( a \in \langle d_0 \rangle \). We claim that \( \mathcal{B}_{j'} \cup \{ f_0 \} \), where \( f_0 = d_0 e_l + v_{d_0} \), is a basis for \( N_{j' \cup \{l\}} \). Letting \( \mathcal{B}_{j'} = \{ f_j \}_{j \in j'} \) we have:

(a) \( \mathcal{B}_{j'} \cup \{ f_0 \} \) is a generating set: In fact, every element of \( v \in N_{j' \cup \{l\}} \) is of the form \((6.4.1)\), hence:

\[
v = ae_l + va = a'd_0e_l + va = a'f_0 - a'v_{d_0} + va, \quad a' \in D.
\]

It follows that \(-a'v_{d_0} + va \in N \cap (\bigoplus_{j \in j'} De_j) = N_j\), where

\[
v = a'f_0 + \sum_{j \in j'} a_j f_j,
\]

and so \( \mathcal{B}_{j'} \cup \{ f_0 \} \) is generating.

(b) \( \mathcal{B}_{j'} \cup \{ f_0 \} \) is linearly independent: given a linear combination

\[
\sum_{j \in j'} a_j f_j + af_0 = \sum_{j \in j'} a_j \left( \sum_{k \in j} c_{jk} e_k \right) + ad_0 e_l + a \sum_{k \in j} b_k e_k = \sum_{k \in j} \left( \sum_{j' \in j'} a_j c_{jk} + ab_k \right) e_j + ad_0 e_l.
\]

we see that if this linear combination is zero, then \( ad_0 = 0 \), hence \( a = 0 \). Since the elements \( \{ f_j \} \) are linearly independent, also \( a_j = 0 \) and the elements \( \mathcal{B}_{j'} \cup \{ f_0 \} \) are linearly independent.

Therefore the pair \( (N_{j' \cup \{l\}}, \mathcal{B}_{j' \cup \{l\}}) \) contradicts the maximality of \( (N_j, \mathcal{B}_{j'}) \). We conclude that \( I = J \), as claimed.

One consequence of the previous theorem is that for modules of finite type over a PID, the concepts “free” and “free of torsion” actually coincide:

**Proposition 6.4.11.** Let \( M \) be a module of finite type over a PID. If \( \text{Tors}(M) = 0 \), then \( M \) is free.

**Proof.** Let \( S \) be a finite generating set. In \( S \) choose a set \( \mathcal{B} = \{ v_1, \ldots, v_n \} \) which is maximal linearly independent in \( S \). If \( v \in S \), there exist \( a_v, a_1, \ldots, a_n \in D \) (non-unique) such that

\[
a_v v = a_1 v_1 + \cdots + a_n v_n \quad (a_v \neq 0).
\]
6.4. MODULES OVER INTEGRAL DOMAINS

Set \( a \equiv \prod_{v \in S} a_v \). Since \( M \) is torsion free, the map \( \phi : M \to M, \, w \mapsto a w \), is an injective homomorphism. We claim that \( \phi(M) \subset \bigoplus_{i=1}^n D v_i \). Since \( M = \langle S \rangle \) it is enough to observe that if \( w \in S \), then

\[
a w = \left( \prod_{v \in S - \{w\}} a_v \right) a_w w
= \left( \prod_{v \in S - \{w\}} a_v \right) (a_1 v_1 + \cdots + a_n v_n) \in \bigoplus_{i=1}^n D v_i.
\]

Hence \( M \) is isomorphic to a submodule of a free module free. By Theorem 6.4.10 \( M \) is free. \( \square \)

Exercises.

1. Show that Proposition 6.4.2 holds

2. Show that, if \( M \) is a free \( D \)-module over an integral domain \( D \), then \( M \) is torsion free. Give an example of a free module \( N \) over a ring \( A \) such that \( \text{Tors}(N) \neq 0 \).

3. Let \( V \) be a finite dimensional vector space \( K \) and \( T : V \to V \) a linear transformation. Show that \( V \) is a torsion \( K[x] \)-module.

4. Give an example of a free module over a integral domain which admits submodules which are not free.

5. Let \( D \) be an integral domain and \( K = \text{Frac}(D) \), seen as a \( D \)-module. In \( D - \{0\} \) consider the partial order defined by:

\[
d_1 \leq d_2 \iff K_{d_1} \subset K_{d_2},
\]

where \( K_d \subset K \) is a \( D \)-submodule \( \{ \frac{a}{d} : a \in D \} \).

(a) Show that \( K = \varinjlim K_d \).

(b) If \( M \) is a \( D \)-module and \( M_K \) is the associated vector space show that \( M_K = \varinjlim (K_d \otimes M) \).

(c) Conclude that \( 1 \otimes v \in M_K \) is zero vector if and only if \( v \in \text{Tors}(M) \).

6. Let \( D \) be a PID. If \( M \) is a free \( D \)-module, \( N \subset M \) is a submodule, and \( \text{Tors}(M/N) = M/N \), show that \( \dim N = \dim M \).
6.5 Finitely Generated Modules over a PID

We will now study modules which are finitely generated over PID, eventually obtaining its classification. As particular instances of this classification one obtains the classification of finitely generated abelian groups and canonical forms for linear transformations between finite dimensional vector spaces.

Recall that in the previous section we have shown that for a PID $D$ every submodule of a free $D$-module is free, and that “free” and “free of torsion” coincide for $D$-modules of finite type. This leads to the first step in the classification of modules of finite type over a PID:

**Theorem 6.5.1.** Let $M$ be a module of finite type over a PID. There exists a free submodule $L \subset M$ such that $M = \text{Tors}(M) \oplus L$ and $\dim L = \text{rank } M$.

**Proof.** The module $M/\text{Tors}(M)$ is free of torsion and of finite type, hence it is free. This means we can choose elements $e_1, \ldots, e_n \in M$, linearly independent, such that

$$M/\text{Tors}(M) = \bigoplus_{i=1}^n D\pi(e_i),$$

where $\pi : M \to M/\text{Tors}(M)$ denotes the quotient map. Let $L = \bigoplus_{i=1}^n De_i \subset M$. Then:

(a) $L \cap \text{Tors}(M) = \{0\}$: If $v \in L \cap \text{Tors}(M)$ there exist scalars $d, d_1, \ldots, d_n \in D$, with $d \neq 0$, such that

$$dv = 0, \quad v = \sum_{i=1}^n d_i e_i.$$

Hence $(dd_1)e_1 + \cdots + (dd_n)e_n = 0$ and we conclude that $dd_1 = \cdots = dd_n = 0$. By the cancellation law, $d_1 = \cdots = d_n = 0$, and so $v = 0$.

(b) $M = L + \text{Tors}(M)$: If $v \in M$ define $d_1, \ldots, d_n \in D$ by

$$\pi(v) = \sum_{i=1}^n d_i \pi(e_i).$$

Then $v = v_T + v_L$, where $v_L = \sum_{i=1}^n d_i e_i \in L$ and $v_T = v - v_L \in N(\pi) = \text{Tors}(M)$. 
By (a) and (b), we have $M = \text{Tors}(M) \oplus L$. If $K = \text{Frac}(D)$ and $\phi : M \to M_K$ is the canonical injection, the restriction of $\phi$ to $L$ is injective. Since $\phi(L)$ generates $M_K$, we conclude that the rank of $M$ equals the dimension of $L$. 

In the above decomposition the free factor $L$ is not unique. This is clear from the proof, where a different choice of a basis $\{e_1, \ldots, e_n\}$ may lead to a different factor. However, its dimension coincides with the rank of $M$ and hence it is an invariant.

The rank of $M$ classifies, up to isomorphism, the free part of $M$. To complete the classification, one needs to classify the finitely generated torsion modules. This will be done in the next paragraphs.

### 6.5.1 Diagonalization of matrices with entries in a PID

Recall that $M_n(D)$ denotes the ring of $n \times n$ matrices with entries in a PID $D$. The following result will be very useful to find special bases in a free module:

**Proposition 6.5.2.** Let $A \in M_n(D)$. There exist invertible matrices $P, Q \in M_n(D)$ such that:

$$Q^{-1}AP = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{pmatrix},$$

where $d_1 \mid d_2 \mid \cdots \mid d_n$. The $d_i$’s are unique up to multiplication by units.

Note that this result does not say that a matrix can be diagonalized by a change of basis: in general, the matrices $P$ and $Q$ are not related in any way.

The normal form for a matrix given by Proposition 6.5.2 can be obtained by performing elementary operations in the rows and columns of the matrix, as in the usual Gauss elimination. Let $E_{ij}$ denote the matrix whose entries are all zero with the exception of the entry $(i, j)$ which equals $1 \in D$. Right and left multiplication by the following invertible matrices leads to the usual elementary operations:

- Exchange of columns (rows): $P_{ij} = I - E_{ii} - E_{jj} + E_{ij} + E_{ji};$
- Multiplication of columns (rows) by units: $D_i(u) = I + (u - 1)E_{ii}$ ($u \in D$ a unit);
• Addition of a multiple of a column (row) to another column (row):

\[ T_{ij}(a) = I + aE_{ij} \ (a \in D). \]

Also, we define the length \( \delta(d) \) of an element \( 0 \neq d \in D \) to be the number of prime factors in a prime factorization of \( d \).

Given any \( n \times n \) matrix \( A = (a_{ij}) \) we would like to show that \( A \) is equivalent\(^\text{12}\) to a diagonal matrix. If \( A = 0 \), there is nothing to prove. If not, then there is a non-zero entry of \( A \) of minimal length, and using elementary operations we can transport it to the \((1,1)\) entry. Let \( a_{1k} \) be some entry such that \( a_{11} \nmid a_{1k} \). By exchange of columns 2 and \( k \), we can assume that this entry is \( a_{12} \). If \( d = \gcd(a_{11}, a_{12}) \), then there exist elements \( p, q \in D \) such that \( pa_{11} + qa_{12} = d \). If we let \( r = a_{12}d^{-1} \) and \( s = a_{11}d^{-1} \), then we see that the matrices

\[
P = \begin{pmatrix}
p & r \\
q & -s \\
1 & \ddots \\
& \ddots & 1
\end{pmatrix}, \quad P^{-1} = \begin{pmatrix}
s & r \\
q & -p \\
1 & \ddots \\
& \ddots & 1
\end{pmatrix},
\]

are inverse to each other. Right multiplication of \( A = (a_{ij}) \) by the matrix \( P \) gives an equivalent matrix whose first row is \((d, 0, a_{13}, \ldots, a_{1n})\) and \( \delta(d) < \delta(a_{11}) \). Similarly, if \( a_{11} \nmid a_{k1} \), we can obtain a new equivalent matrix where the new \((1,1)\) entry \( d \) has length \( \delta(d) < \delta(a_{11}) \), so the minimal \( \delta \) has decreased again. Since \( \delta \) takes values in \( \mathbb{N} \), we can repeat this process a finite number of times leading to an equivalent matrix where \( a_{11} \mid a_{1k} \) and \( a_{11} \mid a_{k1} \), for all \( k = 2, \ldots, n \). Performing elementary operations we obtain finally an equivalent matrix of the form:

\[
\begin{pmatrix}
d_1 & 0 & \ldots & 0 \\
0 & \hat{a}_{22} & \ldots & \hat{a}_{2n} \\
& \vdots & \ddots & \vdots \\
0 & \hat{a}_{n2} & \ldots & \hat{a}_{nn}
\end{pmatrix}.
\]

Obviously we can repeat this process again with the second row and column, etc., so we see that the original matrix is equivalent to a diagonal matrix:

\[
\begin{pmatrix}
d_1 & 0 & \ldots & 0 \\
& \ddots \\
0 & & d_n
\end{pmatrix}.
\]

\(^{12}\)In the following we will say that two matrices \( A \) and \( B \) are equivalent if there exist invertible matrices \( P \) and \( Q \) such that \( B = PAQ \).
If \( d_1 \not| d_2 \), then we can add the second row to the first row and repeat the all process again. Since \( \delta \) always decreases, we will eventually obtain a diagonal matrix where \( d_1 | d_2 \). It should now be clear that we can obtain an equivalent diagonal matrix where \( d_1 | d_2 | \cdots | d_n \).

The elements \( d_1, \ldots, d_n \) in the normal form given by Proposition 6.5.2 are called invariant factors of the matrix \( A \). The uniqueness of the invariant factors follows from the following result which also gives a more efficient method to compute them than the elimination that we used above to prove existence. We leave the proof as an exercise.

**Lemma 6.5.3.** Let \( A \in M_n(D) \) and assume that \( A \) is equivalent to a diagonal matrix

\[
\begin{pmatrix}
d_1 & 0 \\
\vdots & \ddots \\
0 & d_n
\end{pmatrix},
\]

with \( d_1 | d_2 | \cdots | d_n \). If the rank of \( A \) is \( r \), then \( d_i = 0 \), for \( i > r \), and \( d_i = \Delta_i \) for \( i \leq r \), where \( \Delta_0 = 1 \) and \( \Delta_i \) is the greatest common divisor of the minors of \( A \) of size \( i \).

The formulas in this lemma lead immediately to the following corollary:

**Corollary 6.5.4.** The invariant factors are unique up to multiplication by units. Two matrices are equivalent if and only if they possess the same invariant factors.

**Example 6.5.5.**

Let \( D = \mathbb{C}[x] \) and consider the matrix

\[
A = \begin{pmatrix}
x - 2 & 0 & 0 \\
-1 & x & -1 \\
-2 & 4 & x - 4
\end{pmatrix}.
\]

Computing the minors given in Lemma 6.5.3, we obtain

\[
\Delta_1 = 1, \\
\Delta_2 = x - 2, \\
\Delta_3 = (x - 2)^3.
\]

Hence: \( d_1 = 1 \), \( d_2 = (x - 2) \) and \( d_3 = (x - 2)^2 \). In fact, the method of
elimination used above gives the following invertible matrices:

\[
\begin{pmatrix}
0 & -1 & 0 \\
-1 & -x + 2 & 0 \\
1 & x - 4 & 1
\end{pmatrix}
\begin{pmatrix}
x - 2 & 0 & 0 \\
-1 & x & -1 \\
-2 & 4 & x - 4
\end{pmatrix}
\begin{pmatrix}
1 & -1 & 0 \\
0 & 0 & 1 \\
0 & 1 & x
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & x - 2 & 0 \\
0 & 0 & (x - 2)^2
\end{pmatrix}.
\]

### 6.5.2 Decomposition into invariant factors

Let \( M \) be a module over a PID \( D \). For each \( v \in M \), one defined the order ideal of \( v \) by

\[ \text{ann } v = \{ d \in D : dv = 0 \}. \]

This is a principal ideal: \( \text{ann } v = \langle a \rangle \), and the element \( a \in D \) is called the order of \( v \). The order is defined up to multiplication by units. Obviously, the cyclic submodule \( \langle v \rangle \) is isomorphic to \( D/\text{ann } v \).

**Example 6.5.6.**

Let \( G \) be an abelian group, viewed as a \( \mathbb{Z} \)-module. If \( g \in G \), then the cyclic subgroup \( \langle g \rangle \) generated by \( g \) is isomorphic to \( \mathbb{Z}/\text{ann } g \). The order of \( g \), as defined above, coincides with the usual notion of order up to a sign (the units in this case are \( \pm 1 \)).

We can now state and prove our first version of the classification of modules of finite type over a PID:

**Theorem 6.5.7** (Invariant factors decomposition). Let \( D \) be a PID and let \( M \) be a module of finite type over \( D \). Then

\[ M = \langle v_1 \rangle \oplus \cdots \oplus \langle v_k \rangle, \]

where \( \text{ann } v_1 \supset \text{ann } v_2 \supset \cdots \supset \text{ann } v_k \). Letting \( \text{ann } v_i = \langle d_i \rangle \), there is an isomorphism

\[ M \simeq D/\langle d_1 \rangle \oplus \cdots \oplus D/\langle d_k \rangle \]

where \( d_1 | d_2 | \cdots | d_k \). The ideals \( \langle d_i \rangle \) are uniquely determined by \( M \).

**Proof.** If \( \text{rank } M = r \) then

\[ M \simeq \text{Tors}(M) \oplus D \oplus \cdots \oplus D, \]

\( r \text{ terms} \).
hence it is enough to prove the result for a torsion module $M$.

Let $\{w_1, \ldots, w_n\}$ be a finite set of generators of $M$. Denote by $L$ the free module generated by the $w_i$’s. In $L$ there is a basis $\{\hat{w}_1, \ldots, \hat{w}_n\}$ such that $\pi(\hat{w}_i) = w_i$, where $\pi : L \to M$ is the projection. Let $R$ be the kernel of $\pi$, so that $M \cong L/R$. Since $R$ is a free submodule of $L$ and since $M$ is torsion, we must have $\dim R = \dim L = n$ (see Exercise 6.4.6).

Let $\{\hat{v}_1, \ldots, \hat{v}_n\}$ be a basis of $R$, so there exist scalars $a_{ij} \in D$ satisfying the relations

$$\hat{v}_i = \sum_j a_{ji} \hat{w}_j, \quad i = 1, \ldots, n.$$ 

Changing basis in $L$ and $R$,

$$\hat{w}'_i = \sum_j q_{ji} \hat{w}_j, \quad \hat{v}'_i = \sum_j p_{ji} \hat{v}_j,$$

we obtain new relations

$$\hat{v}'_i = \sum_j b_{ji} \hat{w}'_j, \quad i = 1, \ldots, n,$$

and it is immediate to verify that the matrices $A = (a_{ij}), B = (b_{ij}), P = (p_{ij})$ and $Q = (q_{ij})$ are related by

$$B = Q^{-1} AP.$$ 

As we saw above, we can choose invertible matrices $P$ and $Q$ (or, equivalently, we can choose bases of $L$ and $R$) such that $B = \text{diag}(d_1, \ldots, d_n)$ with $d_1 | d_2 | \cdots | d_n$. In this case:

$$\hat{v}'_i = d_i \hat{w}'_i, \quad i = 1, \ldots, n.$$ 

If $w'_i = \pi(\hat{w}'_i)$, we claim that

$$M = \langle w'_1 \rangle \oplus \cdots \oplus \langle w'_n \rangle.$$ 

Since $\text{ann} w'_i = \langle d_i \rangle$, this proves the existence part of the proposition.

Obviously, $M = \sum_i \langle w'_i \rangle$, since the $w'_i$ form a generating set of $L$ and $\pi : L \to M$ is surjective. Hence, to prove the claim, it is enough to show that $\langle w'_k \rangle \cap \sum_{i \neq k} \langle w'_i \rangle = \{0\}$. Let $w$ be an element in this intersection. Then there exist $a_i \in D$ such that

$$w = a_k w'_k = \sum_{i \neq k} a_i w'_i.$$
It follows that in \( L \) we have:

\[
a_k w'_k - \sum_{i \neq k} a_i w'_i \in R
\]

so we conclude that there exist \( b_i \in D \) such that \( a_i = b_id_i \), \( i = 1, \ldots, n \). But then \( w = a_k w'_k = \pi(a_k w'_k) = \pi(b_k d_k w'_k) = \pi(b_k v'_k) = 0. \)

The proof of uniqueness will be given later.

Note that in the decomposition above we can assume that the \( v_k \neq 0 \), i.e., that the \( d_i \) are non-invertible. The elements \( d_1, \ldots, d_k \) in the decomposition above are then defined up to multiplication by a unit, and are called the invariant factors of the module \( M \).

**Corollary 6.5.8.** Two modules of finite type over a PID are isomorphic if and only if they have the same invariant factors.

### 6.5.3 Decomposition into primary factors

There is an alternative way of encoding the classification of modules over a PID, which is based on the prime factorization of elements of \( D \), and which we now explain.

Recall that a PID is a UFD so that any \( 0 \neq a \in D \) can be expressed in the form

\[
a = u \cdot p_1 \cdots p_n,
\]

where \( u \in D \) is a unit and the \( p_i \in D \) are primes. This factorization is unique, up to the order of the factors and multiplication by units. Recall that \( a, b \in D \) are associate, i.e., differ by multiplication by a unit, we write \( a \sim b \).

**Lemma 6.5.9.** Let \( D \) be a PID and assume that \( \gcd(a, b) = 1 \). Then:

\[
D/\langle ab \rangle \simeq D/\langle a \rangle \oplus D/\langle b \rangle.
\]

We leave the proof as an exercise. Using this lemma, we can show:

**Theorem 6.5.10 (Primary factors decomposition).** Let \( D \) be a PID and let \( M \) be a module of finite type over \( D \). Then:

\[
M = L \oplus \langle w_1 \rangle \oplus \cdots \oplus \langle w_n \rangle \simeq L \oplus D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_n^{m_n} \rangle,
\]

where \( L \) is a free submodule with \( \dim L = \operatorname{rank} M \), \( \operatorname{ann} w_i = \langle p_i^{m_i} \rangle \), and the elements \( p_1, \ldots, p_n \in D \) are all primes. The ideals \( \langle p_1^{m_1} \rangle, \ldots, \langle p_n^{m_n} \rangle \) are uniquely determined, up to its order, by \( M \).
Proof. Consider the decomposition of \( M \) into invariant factors:

\[
M = \langle \mathbf{v}_1 \rangle \oplus \cdots \oplus \langle \mathbf{v}_k \rangle.
\]

If \( \text{ann} \mathbf{v}_i = \langle d_i \rangle \), then \( d_1 \mid d_2 \mid \cdots \mid d_k \) and \( d_{k-r+1} = \cdots = d_k = 0 \), where \( r = \text{rank} M \). Hence, we have

\[
\langle \mathbf{v}_{k-r+1} \rangle \oplus \cdots \oplus \langle \mathbf{v}_k \rangle = L,
\]

where \( L \) is free of dimension \( r \). On the other hand, if \( p_1^{m_1}, \ldots, p_n^{m_n} \) are the prime powers in the prime factorizations of \( d_1, \ldots, d_{k-r} \), Lemma 6.5.9 shows that

\[
\langle \mathbf{v}_1 \rangle \oplus \cdots \oplus \langle \mathbf{v}_k \rangle \simeq D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_n^{m_n} \rangle \oplus L.
\]

The proof of uniqueness will be given later.

The elements \( p_1^{m_1}, \ldots, p_n^{m_n} \) associated with the module \( M \) are defined up to multiplication by units. They are called the \textsc{elementary divisors} of \( M \). The elementary divisors together with the rank form a complete set of invariants of \( M \):

\textbf{Corollary 6.5.11.} Two modules of finite type over a PID are isomorphic if and only if they have the same list of elementary divisors and the same rank.

The proof of the theorem shows that the decomposition of \( M \) into invariant factors determines a decomposition of \( M \) into primary factors.

Conversely, let

\[
M \simeq L \oplus D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_n^{m_n} \rangle
\]

be a decomposition de \( M \) into primary factors. Let \( p_1, \ldots, p_s \) be the distinct (i.e., non-associate) primes appearing in this decomposition. We arrange the prime powers in the decomposition as follows:

\[
\begin{array}{cccc}
p_1^{n_{11}} & p_2^{n_{12}} & \cdots & p_s^{n_{1s}} \\
p_1^{n_{21}} & p_2^{n_{22}} & \cdots & p_s^{n_{2s}} \\
\vdots & \vdots & \ddots & \vdots \\
p_1^{n_{t1}} & p_2^{n_{t2}} & \cdots & p_s^{n_{ts}}
\end{array}
\]

(6.5.1)

where \( n_{1i} \leq n_{2i} \leq \cdots \leq n_{ti}, i = 1, \ldots, s \). Note that one may eventually need to add some factors \( 1 = p_i^0 \). We let \( d_j \) be the product of all the prime
power in row \( j \), i.e., \( d_j = p_1^{n_{j1}} \cdot p_2^{n_{j2}} \cdots p_s^{n_{js}} \). Then \( d_1 \mid d_2 \mid \cdots \mid d_t \), and since the prime powers in each \( d_j \) correspond to distinct primes they are relatively prime, so the preceding lemma gives an isomorphism

\[
\text{Tors}(M) \simeq D/\langle d_1 \rangle \oplus \cdots \oplus D/\langle d_t \rangle.
\]

If the dimension of the free part \( L \) is \( r \), then we add to the list of \( d_j \)'s the elements \( d_{t+1} = \cdots = d_{t+r} = 0 \), and this leads to the decomposition of \( M \) into invariant factors.

Given the list of elementary divisors \( \{p_i^{n_{ji}}\} \), the invariant factors \( d_k \) are determined by the process above. Conversely, given the list of invariant factors \( \{d_k\} \), the elementary divisors \( \{p_i^{n_{ji}}\} \) are the prime powers in the prime factorizations of the \( d_k \)'s. Hence, the uniqueness of the ideals \( \langle d_1 \rangle, \ldots, \langle d_k \rangle \) follows from the uniqueness of the ideals \( \langle p_1^{m_1} \rangle, \ldots, \langle p_n^{m_n} \rangle \), which we will study next.

### 6.5.4 Primary components and uniqueness of decompositions

If \( M \) is a \( D \)-module and \( p \in D \) is a prime, the \( p \)-primary component of \( M \) is the submodule

\[
M(p) = \{ v \in M : p^k v = 0, \text{ for some } k \in \mathbb{N} \}.
\]

It is an easy exercise to check that if \( \text{Tors}(M) = M \), then

\[
M = \bigoplus_{\text{prime } p} M(p).
\]

Since \( M \) is of finite type, only a finite number of terms in this direct sum is non-zero.

Given a primary decomposition:

\[
M = L \oplus D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_s^{m_s} \rangle
\]

note that the primary components of \( M \) are:

\[
M(p) = \bigoplus_{\{p_i : p_i \sim p\}} D/\langle p_i^{m_i} \rangle.
\]

Hence, if we are given two primary decompositions

\[
M = L \oplus D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_s^{m_s} \rangle
\]

\[
= L \oplus D/\langle \tilde{p}_1^{m_1} \rangle \oplus \cdots \oplus D/\langle \tilde{p}_t^{m_t} \rangle,
\]

then the uniqueness of the ideals \( \langle p_1^{m_1} \rangle, \ldots, \langle p_n^{m_n} \rangle \) follows from the uniqueness of the ideals \( \langle \tilde{p}_1^{m_1} \rangle, \ldots, \langle \tilde{p}_t^{m_t} \rangle \).
we have that:

\[ M(p) = \bigoplus_{\{p_i: p_i \sim p\}} D/(p_i^{m_i}) = \bigoplus_{\{\tilde{p}_i: \tilde{p}_i \sim p\}} D/\langle \tilde{p}_i^{n_i} \rangle. \]

Hence, to prove the uniqueness of the primary decomposition of \( M \), we can assume that \( M = M(p) \) and that the primes appearing in the two decompositions coincide:

\[ M(p) = D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_s^{m_s} \rangle = D/\langle p_1^{n_1} \rangle \oplus \cdots \oplus D/\langle p_t^{n_t} \rangle \]

Then we can order the decompositions so that \( m_1 \leq m_2 \leq \cdots \leq m_s \) and \( n_1 \leq n_2 \leq \cdots \leq n_t \). If \( v_s \in M \) is such that \( \text{ann} v_s = \langle p_s^{m_s} \rangle \), then the second decomposition gives \( p^n v_s = 0 \), so \( n_t \geq m_s \). Similarly, we find that \( m_s \geq n_t \), so we must have \( m_s = n_s \). The quotient module \( M(p)/\langle v_s \rangle \) admits the decompositions

\[ M(p)/\langle v_s \rangle \cong D/\langle p_1^{m_1} \rangle \oplus \cdots \oplus D/\langle p_s^{m_s-1} \rangle \]
\[ \cong D/\langle p_1^{n_1} \rangle \oplus \cdots \oplus D/\langle p_t^{n_t-1} \rangle \]

By exhaustion, we conclude that \( m_i = n_i \) and \( s = t \). This proves the uniqueness of the decompositions into primary factors and invariant factors.

**Exercises.**

1. An ideal \( I \) in a commutative ring \( A \) is called a **primary ideal** if \( ab \in I \) implies that either \( a \in I \) or \( b^n \in I \), for some \( n \in \mathbb{N} \). Show that if \( D \) is a UFD and \( p \in D \) is a prime, then \( \langle p^n \rangle \) is a primary ideal. This justifies the name “primary decomposition” used in Theorem 6.5.10.

2. Deduce the formulas for the invariant factors given in Lemma 6.5.3.

3. Prove Lemma 6.5.9.

4. Determine diagonal matrices equivalent to the matrices

   (a) \[
   \begin{pmatrix}
   36 & 12 \\
   16 & 18
   \end{pmatrix}
   \]
   over \( \mathbb{Z} \);

   (b) \[
   \begin{pmatrix}
   x - 1 & -2 & -1 \\
   0 & x & 1 \\
   0 & -2 & x - 3
   \end{pmatrix}
   \]
   over \( \mathbb{R}[x] \).
5. Show that if $p \in \mathbb{N}$ is a prime, the following two matrices in $M_n(\mathbb{Z}_p)$ are equivalent:

\[
\begin{pmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 0
\end{pmatrix}, \quad \begin{pmatrix}
1 & 1 & 0 & \cdots & 0 \\
0 & 1 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 1 \\
0 & 0 & \cdots & 1
\end{pmatrix}.
\]

6. Show that $M = \bigoplus_{p \text{ prime}} M(p)$ if and only if $\text{Tors} \, M = M$.

7. If $M = D/\langle p_1 p_2^2 p_3 \rangle \oplus D/\langle p_1 p_3 p_2^3 p_4 \rangle \oplus D/\langle p_1^3 p_2 p_4^5 \rangle$ is a module over a PID $D$, where $p_1, \ldots, p_4$ are non-associate primes, determine the decompositions of $M$ into invariant factors and primary factors.

8. Let $M_1$ and $M_2$ modules of finite type over a PID $D$.

(a) Show that if $M_1$ and $M_2$ are cyclic, then $M_1 \otimes M_2$ is cyclic.

(b) Determine the decompositions of $M_1 \otimes M_2$ into invariant and primary factors in terms of the corresponding decompositions of $M_1$ and $M_2$.

9. Let $M_1$ and $M_2$ be cyclic modules over a PID $D$, of order $a$ and $b$, respectively. Show that if $\gcd(a, b) \neq 1$, then the invariant factors of $M_1 \oplus M_2$ are $\gcd(a, b)$ and $\text{lcm}(a, b)$.

6.6 Classification of Abelian Groups and Canonical Forms for Matrices

We will now show how the structure theory that we developed for modules of finite type over a PID can be used to classify finitely generated abelian groups and to establish canonical forms for linear transformations between vector spaces. This correspond, of course, to consider the cases where $D = \mathbb{Z}$ and $D = K[x]$, with $K$ a algebraically closed field. When $D = \mathbb{Z}$, the units are $\pm 1$, so every non-zero ideal has a non-negative integer as unique generator. If $D = K[x]$ the units are the constant, non-zero, polynomials, so every non-zero ideal has a monic polynomial as a unique generator. Hence, we have unique representatives for the invariant factors and the elementary divisors, and the classifications becomes particularly nice.
6.6. CLASSIFICATIONS

6.6.1 Classification of Finitely Generated Abelian Groups

Let \( G \) be a group. We say that \( G \) is of finite type if there exist elements \( g_1, \ldots, g_m \in G \) such that \[ \forall g \in G, \exists n_1, \ldots, n_m \in \mathbb{Z} : \quad g = g_1^{n_1} \cdots g_m^{n_m}. \]

If \( G \) is an abelian group, then \( G \) is of finite type if and only if \( G \) is a \( \mathbb{Z} \)-module of finite type. Since \( \mathbb{Z} \) is a PID, the classifications theorems in the previous sections immediately yield:

**Theorem 6.6.1** (Classification of abelian groups of finite type). Let \( G \) be an abelian group of finite type. Then \[ G \cong \mathbb{Z}^{d_1} \oplus \cdots \oplus \mathbb{Z}^{d_n}, \]
where \( d_1, \ldots, d_n \) are non-negative integers uniquely determined by the condition \( d_1 \mid d_2 \mid \cdots \mid d_n \). If \( p_1^{n_1}, \ldots, p_s^{n_s} \) are the prime powers in the prime factorization of the non-zero \( d_i \)'s, then \[ G \cong \mathbb{Z}^{p_1^{n_1}} \oplus \cdots \oplus \mathbb{Z}^{p_s^{n_s}} \oplus \mathbb{Z}^r, \]
where \( r \) is the number of \( d_i \)'s equal to zero, i.e., the rank of \( G \).

The integers \( d_i \) (respectively, \( p_i^{n_i} \)) are called the invariant factors (respectively, elementary divisors) of the abelian group \( G \). Together with the rank, they both form two distinct complete set of invariants of an abelian group. Notice that one can find one list of invariants, given the other list.

**Examples 6.6.2.**

1. If \( n \in \mathbb{N} \) admits a prime factorization \( n = p_1^{n_1} \cdots p_s^{n_s} \), then \[ \mathbb{Z}_n \cong \mathbb{Z}^{p_1^{n_1}} \oplus \cdots \oplus \mathbb{Z}^{p_s^{n_s}}. \]
   In this case, there exists only one invariant factor \( n \) and the elementary divisors are the powers \( p_1^{n_1}, \ldots, p_s^{n_s} \).

2. Let \( G = \mathbb{Z}_6 \oplus \mathbb{Z}_{15} \oplus \mathbb{Z}_{18} \). The primary decomposition de \( G \) is \[ G \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_5 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3^2 \]
so the list of elementary divisors is \( \{2, 2, 3, 3^2, 5\} \). We can obtain the list of invariant factors from Table 6.5.1, which in this case reads:

\[
\begin{array}{ccc}
2^0 & 3 & 5^0 \\
2 & 3 & 5^0 \\
2 & 3^2 & 5.
\end{array}
\]
The invariant factors are the products of the prime powers in each line of this table: \( d_1 = 3, d_2 = 6, d_3 = 90 \). Hence, the decomposition of \( G \) into invariant factors is:

\[
G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_6 \oplus \mathbb{Z}_{90}
\]

### 6.6.2 Jordan Canonical Form

Let \( K \) be an algebraically closed field. Let \( V \) be a finite-dimensional vector space over \( K \), and \( T : V \to V \) a linear transformation. The \( K[x] \)-module structure on \( V \) defined by \( T \) is defined by: \( p(x) = a_n x^n + \cdots + a_0 \in K[x] \) and \( v \in V \), then

\[
p(x) \cdot v = a_n T^n(v) + \cdots + a_0 v.
\]

We can use this \( K[x] \)-module structure to obtain information about \( T \): for example, \( \tilde{V} \subset V \) is a \( K[x] \)-submodule if and only if \( \tilde{V} \) linear subspace of \( V \), invariant under \( T \) (see Exercise 6.1.1).

Since \( K[x] \) is a PID and \( V \) is of finite type, one of the classification theorems for modules over PID leads to the following:

**Theorem 6.6.3 (Jordan Canonical Form).** Let \( T : V \to V \) be a linear transformation of a finite-dimensional vector space over an algebraically closed field \( K \). There exists a basis \( \{ e_1, \ldots, e_n \} \) for \( V \) where the matrix representing the linear transformation \( T \) is

\[
J = \begin{pmatrix}
J_1 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & J_m
\end{pmatrix},
\]

where \( J_i \) is a \((n_i \times n_i)\) matrix of the form

\[
\begin{pmatrix}
\lambda_i & 1 \\
0 & \ddots \\
0 & \cdots & 1 \\
\end{pmatrix}.
\]

**Proof.** Since \( K \) is algebraically closed a polynomial \( p(x) \in K[x] \) is prime if and only if \( p(x) = x - \lambda \). Hence, the decomposition of \( V \) as a \( K[x] \)-module into primary factors takes the form:

\[
V \cong V_1 \oplus \cdots \oplus V_m,
\]
where $V_i \simeq \langle v_i \rangle$ are $T$-invariant subspaces, and $\text{ann}(v_i) = \langle (x - \lambda_i)^{n_i} \rangle$. The vectors
\[ \{ (x - \lambda_i)^{n_i-1}v_i, \ldots, (x - \lambda_i)v_i, v_i \} \]
form a basis for $V_i$ over $K$ (exercise), and the matrix of $T$ relative to this basis is precisely $J_i$.

To compute the the Jordan canonical form one needs only to know the elementary divisors (or the invariant factors) of the $K[x]$-module $V$. These can be determined as follows, as shown by inspection of the proof of Theorem 6.5.7. Let \( \{ f_1, \ldots, f_n \} \) be a basis of $V$ over $K$, and let $A = (a_{ij})$ be the matrix representing $T$ in this basis. The set \( \{ f_1, \ldots, f_n \} \) generates $V$ as a $K[x]$-module. Forming the free module $L$ generated by these elements, we obtain the surjective homomorphism $\pi : L \to V$ which has kernel $R$. The elements
\[ e_i := xf_i - \sum_j a_{ji}f_j \]
form a basis for $R$ (as a $K[x]$-module) and the invariant factors of $V$ can be obtained by applying Proposition 6.5.2 to the matrix
\[
\begin{pmatrix}
  x - a_{11} & -a_{12} & \cdots & -a_{1n} \\
  -a_{21} & x - a_{22} & \cdots & -a_{2n} \\
  \vdots & \ddots & \ddots & \vdots \\
  -a_{n1} & -a_{n2} & \cdots & x - a_{nn}
\end{pmatrix}.
\]
In other words, we have an equivalent diagonal matrix
\[
\begin{pmatrix}
  1 & & & \\
  & \ddots & & 0 \\
  & & 1 & d_1(x) \\
  & & 0 & \ddots \\
  & & & d_s(x)
\end{pmatrix},
\]
where $d_1(x) | \cdots | d_s(x)$ are the invariant factors.

**Example 6.6.4.**

Let $T : \mathbb{C}^3 \to \mathbb{C}^3$ be the linear transformation defined relative to the canonical basis by the matrix
\[
A = \begin{pmatrix}
  2 & 0 & 0 \\
  1 & 0 & 1 \\
  2 & -4 & 4
\end{pmatrix}.
\]
As we saw in Example 6.5.5, we have
\[
\begin{pmatrix}
0 & -1 & 0 \\
-1 & -x+2 & 0 \\
1 & x-4 & 1
\end{pmatrix}
\begin{pmatrix}
x-2 & 0 & 0 \\
-1 & x & -1 \\
-2 & 4 & x-4
\end{pmatrix}
\begin{pmatrix}
1 & -1 & 0 \\
0 & 0 & 1 \\
0 & 1 & x
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 \\
0 & x-2 & 0 \\
0 & 0 & (x-2)^2
\end{pmatrix},
\]
and so the elementary divisors are \((x-2)\) and \((x-2)^2\). We conclude that the Jordan canonical form of \(T\) is
\[
J = \begin{pmatrix}
2 & 0 & 0 \\
0 & 2 & 1 \\
0 & 0 & 2
\end{pmatrix}.
\]

The Jordan canonical form is a consequence of the decomposition of the \(K[x]\)-module \(V\) into primary factors. The decomposition of \(V\) into invariant factor leads to another canonical form known as the rational canonical form (see Exercise 6, below).

**Exercises.**

1. Determine all abelian groups of order 120.

2. If \(K\) is an algebraically closed field of characteristic zero, show that \(U = \{r : r \text{ is a root of } x^n - 1 = 0\}\) is an abelian group isomorphic to \(\mathbb{Z}_n\).

3. Let \(T : V \to V\) be a linear transformation of a finite-dimensional vector space over a field \(K\) and assume that \(V \simeq \langle v \rangle\) (as \(K[x]\)-module), where \(\text{ann}(v) = \langle (x-\lambda)^m \rangle\). Show that the elements
\[
\{ (x-\lambda)^{m-1}v, \ldots, (x-\lambda)v, v \}
\]
form a basis of \(V\) over \(K\).

4. Determine the Jordan canonical form of the matrices:

(a) \(A = \begin{pmatrix}
-1 & 1 & -2 \\
0 & -1 & 4 \\
0 & 0 & 1
\end{pmatrix}\).

\[^{13}\text{Actually, the rational canonical form, contrary to the Jordan canonical form, does not require } K \text{ to be algebraically closed.}\]
6.6. CLASSIFICATIONS

(b) \[ B = \begin{pmatrix} 0 & 0 & 0 & -8 \\ 1 & 0 & 0 & 16 \\ 0 & 1 & 0 & -14 \\ 0 & 0 & 1 & 6 \end{pmatrix}. \]

5. Let \( T : V \to V \) be a linear transformation of a finite-dimensional vector space over a field \( K \), and let \( d_1(x) \mid \cdots \mid d_s(x) \) be the invariant factors of \( K[x]\)-module \( V \). One calls \( m(x) = d_s(x) \) the minimal polynomial and \( p(x) = d_1(x) \cdots d_s(x) \) the characteristic polynomial of the linear transformation \( T \).

(a) Show that \( m(x) \neq 0 \), \( m(T) = 0 \) and that if \( q(x) \) is a polynomial such that \( q(T) = 0 \) then \( m(x)|q(x) \);

(b) Show that \( p(x) \neq 0 \), \( p(T) = 0 \) and that \( p(x) = \det(xI - T) \).

6. (RATIONAL CANONICAL FORM) Let \( T : V \to V \) be a linear transformation of a finite-dimensional vector space over a field \( K \). Using the decomposition of \( V \) into invariant factors, as a \( K[x]\)-module, show that there exists a basis \( \{e_1, \ldots, e_n\} \) for \( V \) over \( K \) where the matrix representing \( T \) is

\[
R = \begin{pmatrix} R_1 & 0 & \cdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R_m \end{pmatrix},
\]

where \( R_i \) is a \((n_i \times n_i)\) matrix of the form

\[
\begin{pmatrix} 0 & 0 & -a_0 \\ 1 & 0 & \vdots \\ \vdots & \ddots & \vdots \\ 0 & 0 & -a_{n_i-2} \\ 1 & \cdots & -a_{n_i-1} \end{pmatrix}.
\]

One calls \( R \) the rational canonical form of the linear transformation \( T \).

7. Recall that a scalar linear differential ordinary equation (o.d.e)

\[
\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_1 \frac{dy}{dt} + a_0 y = 0
\]

is equivalent to a system of first order linear o.d.e’s:

\[
\frac{dx}{dt} = xA,
\]
where \( \mathbf{x} = (y, y', y'', \ldots, y^{(n-1)}) \) and \( A \) is the companion matrix

\[
\begin{pmatrix}
0 & 0 & -a_0 \\
1 & 0 & \vdots \\
\vdots & \ddots & \vdots \\
0 & 0 & -a_{n-2} \\
1 & & -a_{n-1}
\end{pmatrix}
\]

Using the rational canonical form, show the following converse: every system of first order linear o.d.e’s is equivalent to a system of decoupled scalar linear o.d.e.’s.

6.7 Categories and Functors

Many of the algebraic structures that we have been studying, albeit distinct, often admit similar properties. We have seen that groups, rings or modules share common constructions and that the methods of proof are the same. We shall now see that one can formalize these similarities in a very precise way. This relies on the following fundamental concept:

**Definition 6.7.1.** A category \( \mathcal{C} \) consists of:

(i) A class of objects.

(ii) For each pair of objects \( (X, Y) \), a set \( \text{Hom}(X, Y) \), whose elements are called morphisms.

(iii) A map \( \text{Hom}(X, Y) \times \text{Hom}(Y, Z) \rightarrow \text{Hom}(X, Z) \), called composition of morphisms.

The image of the pair \( (\phi, \psi) \) under composition of morphisms will be denoted by \( \psi \circ \phi \), and the following properties must hold:

(C1) **Associativity:** if \( \phi \in \text{Hom}(X, Y) \), \( \psi \in \text{Hom}(Y, Z) \) and \( \tau \in \text{Hom}(Z, W) \), then \( \tau \circ (\psi \circ \phi) = (\tau \circ \psi) \circ \phi \).

(C2) **Identity:** for any object \( X \), there exists a morphism \( 1_X \in \text{Hom}(X, X) \) which satisfies

\[1_X \circ \phi = \phi \quad \text{and} \quad \psi \circ 1_X = \psi,\]

for any \( \phi \in \text{Hom}(W, X) \) and \( \psi \in \text{Hom}(X, Y) \), where \( Y \) and \( W \) are arbitrary objects.
Examples 6.7.2.

1. The category \( S \) of sets is the category where the objects are the sets, the morphisms \( \text{Hom}(X,Y) \) are the maps \( \phi : X \to Y \), and the composition of morphisms is the usual composition of maps.

2. The category \( G \) of groups is the category where the objects are the groups, the morphisms \( \text{Hom}(G,H) \) are the group homomorphisms \( \phi : G \to H \), and the composition of morphisms is the composition of homomorphisms.

3. The category \( M_A \) of (left/right) modules over a ring \( A \) is the category where the objects are the (left/right) \( A \)-modules, the morphisms \( \text{Hom}(M,N) \) are the \( A \)-linear transformations \( \phi : M \to N \), and the composition of morphisms is the composition of linear transformations.

4. The category \( T \) of topological spaces is the category where the objects are the topological spaces, the morphisms \( \text{Hom}(X,Y) \) are the continuous maps, and the composition of morphisms is the usual composition of continuous maps.

The first example above shows that, in general, the objects of a category form a classe, rather than a set: one cannot define a “set of all sets” without getting into paradoxes (e.g., is the set of all sets a set?). This distinction, whose complete justification requires a deep study of Set Theory\(^{14}\), means that to classes one does not apply the usual operations over sets, such as passing to subsets. A category where all the objects are elements of a set is called a SMALL CATEGORY.

It maybe convenient of think of a morphism \( \phi \in \text{Hom}(X,Y) \) as an arrow from \( X \) to \( Y \). However, it is important to note that, in spite of its name, a morphism \( \phi \in \text{Hom}(X,Y) \) does not need to be a map from \( X \) to \( Y \), as shown in the following simple example:

Example 6.7.3.

Fix a group \( G \). Let \( C \) be the category with only one object \( \{*\} \) and where the morphisms \( \text{Hom}(*,*) \) are the elements of \( G \). The composition of two morphisms is the multiplication in the group \( G \). If \( G \) is a non-trivial group, the morphisms are not maps between the objects.

A CONCRETE CATEGORY is a category \( C \) where every morphism \( \phi \in \text{Hom}(X,Y) \) is actually a map \( X \to Y \), where the identity morphism \( 1_X \in \)
Hom(\(X, Y\)) is the identity map \(X \to X\), and where composition of morphisms is the usual composition of maps. Most of the categories that we have found up to now are concrete categories. In any case, we will always represent a morphism \(\phi \in \text{Hom}(X, Y)\) symbolically as \(\phi : X \to Y\), keeping in mind that \(\phi\) is not necessarily a map from \(X\) to \(Y\).

Let us give some elementary properties of categories.

**Proposition 6.7.4.** In a category \(C\), for each object \(X\), the identity morphism \(1_X\) is unique.

*Proof.* In fact, if \(1_X\) and \(1'_X\) are two identities in \(X\), then by the identity property (C2) applied to both \(1_X\) and \(1'_X\), we find

\[
1_X \circ 1'_X = 1'_X \quad \text{and} \quad 1_X \circ 1'_X = 1_X.
\]

Hence, \(1_X = 1'_X\). □

In a category \(C\), given a morphism \(f : X \to Y\), we say that \(g : Y \to X\) is a **left inverse** of \(f\) if

\[g \circ f = 1_X.\]

Similarly, one defines a **right inverse** of \(f\).

**Proposition 6.7.5.** If \(f : X \to Y\) had both a left inverse \(g\) and a right inverse \(g'\), then \(g = g'\).

*Proof.* From the definition of left/right inverse we obtain:

\[
(g \circ f) \circ g' = 1_X \circ g' = g',
\]

\[
g \circ (f \circ g') = g \circ 1_Y = g.
\]

Therefore, by the associativity property (C1), we find that \(g = g'\). □

When \(f : X \to Y\) has both a right and a left inverse \(g\), one calls \(g\) the inverse of \(f\) and writes \(g = f^{-1}\). In this case we say that \(f\) is an **isomorphism** in the category.

**Examples 6.7.6.**

1. In Examples 6.7.2 the isomorphisms are the bijections (in the category of sets), the group isomorphisms (in the category of groups), the linear isomorphisms (in the category of modules over a ring) and the homeomorphisms (in the category of topological spaces).

2. In Example 6.7.3 every morphism is an isomorphism. A small category where every morphism is invertible is called a **groupoid**.
6.7. CATEGORIES AND FUNCTORS

Next we introduce a notion which allows us to relate two categories.

**Definition 6.7.7.** A **covariant functor** $F$ from a category $\mathcal{C}$ to a category $\mathcal{D}$ is a map that associates to an object $X$ of $\mathcal{C}$ an object $F(X)$ of $\mathcal{D}$, and to each morphism $\phi : X \to Y$, a morphism $F(\phi) : F(X) \to F(Y)$, such that the following properties hold:

(i) $F$ preserves identities: $F(1_X) = 1_{F(X)}$, for every object $X$ of $\mathcal{C}$;

(ii) $F$ preserves compositions: $F(\phi \circ \psi) = F(\phi) \circ F(\psi)$, for all morphisms $\phi$ and $\psi$ that are composable.

Similarly, a **contravariant functor** $F$ from a category $\mathcal{C}$ to a category $\mathcal{D}$ is a map that associates to each object $X$ of $\mathcal{C}$ an object $F(X)$ of $\mathcal{D}$, and to each morphism $\phi : X \to Y$, a morphism $F(\phi) : F(Y) \to F(X)$, and:

(i) $F$ preserves identities: $F(1_X) = 1_{F(X)}$ for every object $X$ of $\mathcal{C}$;

(ii') $F$ inverts compositions: $F(\phi \circ \psi) = F(\psi) \circ F(\phi)$ for all morphisms $\phi$ and $\psi$ that are composable.

Let us see some examples of functors:

**Examples 6.7.8.**

1. The map that associates to each group its underlying set and to each group homomorphism the underlying map of sets, is a covariant functor from the category of groups to the category of sets. More generally, for any concrete category $\mathcal{C}$, there is a **forgetful functor** $F : \mathcal{C} \to \mathcal{S}$ from $\mathcal{C}$ to the category of sets $\mathcal{S}$ which "forgets" the structure: to each object $X$ of $\mathcal{C}$, the functor $F$ associates the underlying set of $X$, and to each morphism $\phi \in \text{Hom}(X,Y)$ the functor $F$ associates the underlying map $X \to Y$.

2. The map that associates to each module $M$ over a commutative ring $A$ its dual $M^*$, and to each $A$-linear map $\phi : M \to N$ the transpose $\phi^* : N^* \to M^*$, is a contravariant functor from the category of left $A$-modules to the category of right $A$-modules.

3. The map that associates to each topological space $X$ the $k$-th homology group $H_k(X,\mathbb{Z})$ (respectively, the $k$-th cohomology group $H^k(X,\mathbb{Z})$) and to each continuous map $\phi : X \to Y$ the group homomorphism $\phi_* : H_k(X,\mathbb{Z}) \to H_k(Y,\mathbb{Z})$ (respectively, $\phi^* : H^k(Y,\mathbb{Z}) \to H^k(X,\mathbb{Z})$), is a covariant (respectively, contravariant) functor from the category of topological spaces to the category of abelian groups.
Many constructions that we have seen before are special instances of general constructions in Category Theory. Consider, for example, the notion of direct product: if \( \{X_i\}_{i \in I} \) is a family of objects in some category \( \mathcal{C} \) the product of the objects \( \{X_i\}_{i \in I} \) is a pair \((Z, \{\pi_i\}_{i \in I})\), where \( Z \) is an object in \( \mathcal{C} \) and each \( \pi_i : Z \to X_i \) is a morphism in \( \mathcal{C} \), satisfying the following universal property:

- For any object \( Y \) and any collection of morphisms \( \{\phi_i : Y \to X_i\}_{i \in I} \), there exists a unique morphism \( \phi : Y \to Z \) such that, for each \( i \in I \), the following diagram commutes:

\[
\begin{array}{ccc}
Y & \xrightarrow{\phi} & Z \\
\downarrow{\phi_i} & & \downarrow{\pi_i} \\
X_i & \end{array}
\]

It is easy to see that if products exist, then they are unique up to an isomorphism in \( \mathcal{C} \). In this case, we denote the product by \( \prod_{i \in I} X_i \). Note that the product may or may not exist, depending on the category. For a given category it is necessary to show the existence, and for that one needs some concrete model (usually quite obvious). We have followed exactly this procedure in the case of the category of groups (direct product of groups) and in the case of the category of modules over a ring (direct product of modules).

Another advantage of Category Theory is to make it possible to give a precise meaning to certain expressions one often uses when discussing some mathematical formalism. For example, one may say that a certain map is “induced” or that a certain construction is “natural”. In order to illustrate this usage of Category Theory, let us see how one can make the notion of “natural” precise. We recall that this expression is often is used as a synonymous of “not depending on choices”.

**Definition 6.7.9.** A **natural transformation** \( T \) between two functors \( F : \mathcal{C} \to \mathcal{D} \) and \( G : \mathcal{C} \to \mathcal{D} \) is a map which associates to each object \( X \) in a category \( \mathcal{C} \) a morphism \( T_X : F(X) \to G(X) \) in the category \( \mathcal{D} \), so that for any morphism \( \phi : X \to Y \) of \( \mathcal{C} \) the following diagram commutes:

\[
\begin{array}{ccc}
F(X) & \xrightarrow{T_X} & G(X) \\
\downarrow{F(\phi)} & & \downarrow{G(\phi)} \\
F(Y) & \xrightarrow{T_Y} & G(Y) \\
\end{array}
\]
If $T_X$ is an isomorphism for every object $X$, we say that $T$ is a **natural equivalence**.

As a simple illustration, we show in the next example how one can formalize precisely the statement that for any finite-dimensional vector space $V$ there exists a natural isomorphism between the double dual $(V^*)^*$ and $V$.

**Example 6.7.10.**

Let $C$ be the category of (left) modules over a ring $A$. For each $A$-module $M$, denote by $M^{**}$ the double dual:

$$M^{**} \equiv (M^*)^*,$$

and for each $A$-linear transformation $\phi : M \to N$ denote by $\phi^{**} : M^{**} \to N^{**}$ the double transpose:

$$\phi^{**} \equiv (\phi^*)^*.$$

This defines a covariant functor from $C$ to $C$.

Now, for each $A$-module $M$, we denote by $T_M : M \to M^{**}$ the $A$-linear transformation defined as follows: for each $v \in M$, $T_M(v) : M^* \to A$ the linear transformation

$$\xi \mapsto \xi(v).$$

One checks easily that $M \mapsto T_M$ is a natural transformation between the double dual functor and the identity functor.

If instead of the category of $A$-modules we consider instead the category of finite-dimensional vector spaces over a field $K$, then $V \mapsto T_V$ defines a natural equivalence natural between the double dual functor and the identity functor.

We will not make heavy use of Category Theory (but in some situations it we will be quite useful as a language!). Hence, we will not develop further Category Theory, but you should keep in mind that Category Theory is an important subject, that plays a major role in many branches of Mathematics, including Algebraic Topology, Algebraic Geometry, etc.

**Exercises.**

1. Let $\{X_i\}_{i \in I}$ be a family of objects in a category $C$. Define **coproduct** $\prod_{i \in I} X_i$ of the objects $\{X_i\}_{i \in I}$ by diverting the direction of the arrows in the definition of product, and verify that the direct sum of abelian groups and modules, and the free product of groups, are coproducts in the appropriate categories.
2. A set with a marked point is a pair \((X, x)\) where \(X\) is a set and \(x \in X\). A morphism \((X, x) \to (Y, y)\) between sets with marked points is a map \(f : X \to Y\) such that \(f(x) = y\).

(a) Show that sets with a marked point and the morphisms between marked sets form a category.

(b) Show that products exist in this category and describe them explicitly.

(c) Show that coproducts exist in this category and describe them explicitly.

3. Let \(L\) be an object in a category \(\mathcal{C}\), \(S\) a non-empty set, and \(\iota : X \to L\) a map. One says that \(L\) is a free object in the set \(S\) if for each object \(X\) in \(\mathcal{C}\) and map \(\phi : S \to X\) there exists a unique morphism \(\tilde{\phi} : L \to X\) such that the following diagram commutes:

\[
\begin{array}{ccc}
S & \xrightarrow{\iota} & L \\
\downarrow^{\phi} & & \downarrow^{\tilde{\phi}} \\
X.
\end{array}
\]

Show that, in any category \(\mathcal{C}\), if \(L\) is a free object in the set \(S\), \(L'\) is a free object in the set \(S'\) and \(|S| = |S'|\), then \(L'\) is isomorphic to \(L\).

4. An object \(I\) in a category \(\mathcal{C}\) is called universal or initial if for each object \(X\) in \(\mathcal{C}\) there exists a unique morphism \(\phi : I \to X\):

(a) Show that any two initial objects in a category \(\mathcal{C}\) are isomorphic.

(b) Determine the initial objects in the category of groups.

(c) Show that the coproduct can be consider as an initial object in an appropriate category.

5. Define, similarly to the previous exercise, a co-universal or terminal object in a category \(\mathcal{C}\).

(a) Show that any two terminal objects in a category \(\mathcal{C}\) are isomorphic.

(b) Determine the terminal objects in the category of groups.

(c) Show that the product can be consider as a terminal object in an appropriate category.
Chapter 7

Galois Theory

The solution of a quadratic equation was known to the mathematicians in Babylon, which knew how to “complete the square”, and was popularized in the western world during the Renascence in Latin translations of the Arab book by al-Khowarizmi *Al-jabr wa’l muqabalah* (mentioned in Chapter 7). In 1545, it was published the monograph *Ars Magna* by Geronimo Cardano (1501–1576), also known by the name Cardan, which included formulas to solve third and fourth degree polynomial equations. Cardan attributed these formulas, respectively, to Niccolo Tartaglia (1500–1565) and Ludovico Ferrari (1522–1565). The discovery of these formulas, and the fight for the priority in its discovery, has a very curious and amusing history, which can be found in the books mentioned above.

The formula for the solution of the cubic equation $x^3 + px = q$, known today as Cardan’s formula, is

$$x = \sqrt[3]{\sqrt{(p/3)^3 + (q/2)^2} + q/2} - \sqrt[3]{\sqrt{(p/3)^3 + (q/2)^2} - q/2}.$$

The general cubic equation $y^3 + by^2 + cy + d = 0$ can be reduced to this case by performing the substitution $y = x - b/3$. One can have a good idea of the degree of difficulty in solving these equations if one verifies by substitution that Cardan’s formula does indeed furnish a solution.

A quartic equation can also be reduced to the solution of a cubic. First one can assume, eventually after some translation, that the quartic is of the form $x^4 + px^2 + qx + r = 0$. Completing the square, one obtains

$$(x^2 + p)^2 = px^2 - qx - r + p^2.$$ 

The trick consists then in observing that, for any \( y \), one has:
\[
(x^2 + p + y)^2 = px^2 - qx - r + p^2 + 2y(x^2 + p) + y^2
= (p + 2y)x^2 - qx + (p^2 - r + 2py + y^2).
\]
This last equation is quadratic in \( x \), and we can choose \( y \) so that we obtain a perfect square. This is achieved precisely when the discriminant vanishes:
\[
q^2 - 4(p + 2y)(p^2 - r + 2py + y^2) = 0
\]
Now the last equation is the cubic in \( y \):
\[
-8y^3 - 20py^2 + (-16p^2 + 8r)y + (q^2 - 4p^3 + 4pr) = 0,
\]
which can be solved by using Cardan’s formula. For this value of \( y \), the right hand side of the above auxiliary equation is then a perfect square, so taking square roots we obtain a quadratic equation which can be easily solved.

The solutions to the cubic and the quartic contained in the *Ars Magna* provided a strong stimulus in the search of general formulas for the solutions of algebraic equations of higher order. These efforts were unsuccessful for over 300 years, and one had to wait for the beginning of the 19th Century when Abel and Ruffini reach the opposite conclusion: for an equation of the 5th degree there is no general formula that expresses the solutions in terms of radicals of the coefficients of the equation.

Inspired by Abel’s proof of the impossibility of solving the quintic equation, Galois initiated a systematic study of algebraic equations of arbitrary degree and showed that it was not only impossible to give a general formula for equations of any degree higher than 5, but gave also a criterion to decide if a particular equation could be solved and, in the affirmative case, a method to solve it. Galois investigations, in spite of its premature death, were fundamental in shaping the field of Algebra as we know it nowadays, and had much more consequences that the complete solution to the original problem of solving algebraic equation through radicals.

In order to illustrate Galois ideas, consider the quartic equation with rational coefficients:
\[
p(x) = x^4 + x^3 + x^2 + x + 1 = 0.
\]
The roots of this equation are \( r_k = e^{2\pi i k/5} \) (why?). Consider all the possible polynomial equations, with rational coefficients, which are satisfied by these roots, such as:
\[
r_1 + r_2 + r_3 + r_4 - 1 = 0, \quad r_1 r_4 = 1,
\]
\[
(r_1 + r_4)^2 + (r_1 + r_4) - 1 = 0, \quad (r_1)^5 - 1 = 0, \quad \ldots
\]
The key observation is the following: the set of all permutations of the roots which transform equations of this type into equation of this type form a group, now called the \textit{Galois group} of the equation. In the present case the group is: \( G = \{I, (1243), (14)(23), (1342)\} \). Galois understood that the structure of this group is the key to solve the equation \(2\).

Consider, for example, the subgroup \( H = \{I, (14)(23)\} \subset G \). One can check that among the polynomial expressions above the ones that are \textit{fixed} by \( H \) are precisely the polynomials in \( w_1 = r_1 + r_4 \) and \( w_2 = r_2 + r_3 \). Now \( w_1 \) and \( w_2 \) are the roots of the quadratic equation

\[ x^2 + x - 1 = 0. \]

Hence, assuming that we did not know the solutions to the original quartic equation, we could try to solve it by solving first this quadratic equation, obtaining:

\[ r_1 + r_4 = \frac{-1 + \sqrt{5}}{2}, \quad r_2 + r_3 = \frac{-1 - \sqrt{5}}{2}, \]

and then solving the quadratic equation:

\[(x - r_1)(x - r_4) = x^2 - (r_1 + r_4)x + r_1r_4 = 0,\]

Note that the last equation is an equation whose coefficients are polynomials in \( w_1 \) and \( w_2 \), since \( r_1r_4 = 1 \).

Note that the Galois Group of a polynomial equation \( p(x) \in \mathbb{Q}[x] \) can be characterized as the \textit{group of symmetries} of the equation: it consists of the transformations that take solutions (roots) to solutions, preserving the algebraic structure of the roots. In the modern formulation of Galois Theory one constructs the field \( \mathbb{Q}(r_1, \ldots, r_n) \) generated by the roots of \( p(x) \) and the Galois Group consists of the automorphisms of this field \(3\). One can say that Galois Theory consists in replacing questions concerning the structure of this field into questions about the structure of the Galois group.

### 7.1 Field Extensions

Recall that a field \( L \) is said to be an \textit{extension} of the field \( K \), if \( K \) is a subfield of \( L \). We will now initiate a systematic study of field extensions. We start by recalling some results that were deduced in Chapter \(3\).

\footnote{This discovery is even more remarkable, since Galois had first to invent the concept of a group, which at the time was not known!}

\footnote{The notion of field was only formalized by Dedekind in 1879, more that 50 years after the tragic death of Galois.}
Let $K$ be a field and $L$ an extension of $K$. Recall that we call $L$ a finite extension (respectively infinite extension) if $L$, seen as a vector space over $K$, has finite dimension (respectively, infinite dimension). The dimension of $L$ over $K$ is denoted by $[L : K]$. If $S \subset L$ is some subset, we denote by $K(S)$ the smallest subfield of $L$ which contains $K$ and $S$. Obviously, $K(S)$ is an extension of $K$ generated by $S$. If $S = \{u_1, \ldots, u_n\}$, we write $K(u_1, \ldots, u_n)$ instead of $K(\{u_1, \ldots, u_n\})$.

Let us consider in detail the case where $S$ consists of a single element, i.e., $K(u)$ where $u \in L$. If $x$ is an indeterminate, there is a ring homomorphism $\phi : K[x] \to L$ which associates to a polynomial $g(x)$ its value at $u$: $g(x) \mapsto g(u)$. The kernel $\ker(\phi)$ is an ideal in $K[x]$, which must be principal. There are two possibilities:

(i) $u$ is transcendental over $K$. This corresponds to the case $\ker(\phi) = \{0\}$, so $\phi$ is a monomorphism whose image is $K[u]$ and which extends uniquely to the field of fractions $K(x) = \text{Frac}(K[x])$. It follows that $K(u) \simeq K(x)$, and the elements in $K(u)$ take the form $g(u)/f(u)$ where $g(x), f(x) \in K[x]$ and $f(x) \neq 0$. The extension $K(u)$ has infinite dimension over $K$;

(ii) $u$ is algebraic over $K$. In this case $\ker(\phi) = \langle p(x) \rangle$, where $p(x)$ is an irreducible polynomial, and $K(u) = K[u] \simeq K[x]/\langle p(x) \rangle$. If we require $p(x)$ to be monic, then $p(x)$ is unique and is called the minimal polynomial of $u$. The extension $K(u)$ has dimension $[K(u) : K] = \deg p(x)$. We call this dimension the algebraic degree of $u$ over $K$.

**Definition 7.1.1.** An extension $L$ of $K$ is called a simple extension if there exists $u \in L$ such that $L = K(u)$. In this case, one says that $u$ is a primitive element of $L$.

A extension $L$ of $K$ is called algebraic (respectively transcendental) if all the elements of $L$ are algebraic (respectively, there exists a transcendental element in $L$) over $K$. A simple extension $L$ is algebraic or transcendental exactly when its primitive elements are algebraic or transcendental. A finite extension $L$ over $K$ is always algebraic (see Exercise 3.1.8), but there are algebraic extensions of infinite dimension over $K$. A transcendental extension is always infinite.

**Examples 7.1.2.**

1. Consider $L = \mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{R}$ and $K = \mathbb{Q}$. By Eisenstein Criterion, the polynomial $p(x) = x^n - 2$ ($n \geq 2$) is irreducible over $\mathbb{Q}$. The number $u = \sqrt[3]{2}$ is a root of $p(x)$ in $\mathbb{R}$. Hence, $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = n$ and $\sqrt[3]{2}$ is algebraic of degree $n$ over $\mathbb{Q}$. 

2. Consider \( L = \mathbb{C} \) and \( K = \mathbb{R} \). The polynomial \( x^2 + 1 \) is clearly irreducible over \( \mathbb{R} \). The number \( i \in \mathbb{C} \) is a root of \( x^2 + 1 \) in \( \mathbb{C} \), hence \( i \) is algebraic of degree 2 over \( \mathbb{R} \). In fact, \( \mathbb{C} = \mathbb{R}[x]/(x^2 + 1) = \mathbb{R}(i) \), so \( \mathbb{C} \) is a simple extension of \( \mathbb{R} \), and \( i \) is a primitive element.

3. Consider \( L = \mathbb{C} \) and \( K = \mathbb{Q} \). Then \( L \) is a transcendental extensional of \( K \) and \([\mathbb{C} : \mathbb{Q}] = \infty\).

We leave the proof of the following proposition as an exercise:

**Proposition 7.1.3.** Let \( M \supset L \supset K \) be successive extensions of a field \( K \). Then \([M : K]\) is finite if and only if both \([M : L]\) and \([L : K]\) are finite and in this case:

\[ [M : K] = [M : L] \cdot [L : K]. \]

**Example 7.1.4.**

We saw above that \([\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2\). The polynomial \( q(x) = x^2 - 3 \) is irreducible over \( \mathbb{Q}(\sqrt{2}) \): a root \( a + b\sqrt{2} \) of \( q(x) \) in \( \mathbb{Q}(\sqrt{2}) \) must satisfy

\[ (a + b\sqrt{2})^2 = 3, \quad a, b \in \mathbb{Q}. \]

Hence, \((a^2 + 2b^2) + 2ab\sqrt{2} = 3\), so that \( a^2 + 2b^2 = 3 \) and \( ab = 0 \). If \( b = 0 \), then \( a^2 = 3 \) which is impossible since \( \sqrt{3} \notin \mathbb{Q} \). If \( a = 0 \) we have \((2b)^2 = 6\) which is also impossible since \( \sqrt{6} \notin \mathbb{Q} \). We conclude that \( \mathbb{Q}(\sqrt{2})(\sqrt{3}) \cong \mathbb{Q}(\sqrt{2})[x]/(x^2 - 3) \) so that \([\mathbb{Q}(\sqrt{2})(\sqrt{3}), \mathbb{Q}(\sqrt{2})] = 2\). By Proposition 7.1.3, we conclude that

\[ [\mathbb{Q}(\sqrt{2}, \sqrt{3}), \mathbb{Q}] = [\mathbb{Q}(\sqrt{2})(\sqrt{3}), \mathbb{Q}(\sqrt{2})] \cdot [\mathbb{Q}(\sqrt{2}), \mathbb{Q}] = 2 \cdot 2 = 4. \]

In the examples so far, we have consider only subfields of the field of complex numbers. In this case, there is no problem in finding extensions where a given polynomial has roots due to a fundamental property of \( \mathbb{C} \): every polynomial (of degree \( \geq 1 \)) with coefficients in \( \mathbb{C} \) has at least one root, i.e., \( \mathbb{C} \) is algebraically closed. We recall:

**Proposition 7.1.5.** The following statements are equivalent:

(i) \( L \) is an algebraically closed field.

(ii) Every polynomial \( p(x) = a_nx^n + \cdots + a_1x + a_0 \in L[x] \) factors into a product of linear factors: \( p(x) = a_n \prod_{i=1}^{n} (x - r_i) \).

(iii) Every irreducible polynomial in \( L[x] \) has degree 1.

(iv) There are no proper algebraic extensions of \( L \).
There are many interesting examples of fields which are not subfields of \( \mathbb{C} \). For example, the number fields \( \mathbb{Z}_p \), relevant in Number Theory, or the field of fractions \( \mathbb{C}(z) \), fundamental in Algebraic Geometry. For such fields it is not completely obvious that given a polynomial there exists an extension where the polynomial factors into linear polynomials. We will deal with these issues in the next sections.

**Exercises.**

1. Show that the subset of \( \mathbb{C} \) formed by all algebraic elements over \( \mathbb{Q} \) is an algebraic extension of \( \mathbb{Q} \) of infinite dimension over \( \mathbb{Q} \) (such elements of \( \mathbb{C} \) are usually called *algebraic numbers*).

2. Give a proof of Proposition 7.1.3.

3. Show that:
   
   \begin{enumerate}
     \item every algebraically closed field \( K \) is infinite;
     \item if \( K \) is an infinite field, show that any algebraic extension of \( K \) has the same cardinality as \( K \);
   \end{enumerate}

4. Show that there are elements in \( \mathbb{C} \) which are not algebraic over \( \mathbb{Q} \).

5. Let \( L \subset \mathbb{C} \) be the smallest extension of \( \mathbb{Q} \) which contains the roots of the polynomial \( x^3 - 2 \). Decide if \( L \) is simple or not and in the affirmative case give an example of a primitive element.

6. Prove Proposition 7.1.5.

7. Let \( K \) be a field and \( x \) an indeterminate. In \( K(x) \) consider the element \( u = x^2 \). Show that \( K(x) \) is a simple extension of \( K(u) \). What is the dimension of \( [K(x) : K(u)] \)?

### 7.2 Compass-and-Straightedge Constructions

The mathematicians in Ancient Greece used to express in a geometric manner many mathematical concepts and constructions. In general, a construction was consider to be valid only if it could be obtained by using only a compass and a straightedge (an unmarked ruler). In spite of their spectacular achievements, there were a few simple constructions that they could not find a way to perform using only compass and straightedge.
Among the most famous constructions were the Greeks failed to find a compass-and-straightedge construction, were the following:

(i) trisecting an angle;
(ii) duplicating a cube;
(iii) constructing a regular heptagon;
(iv) squaring the circle, i.e., constructing a square with area equal to a given circle.

The problem of existence of such compass-and-straightedge constructions can be transformed into a problem in Field Theory. Then Galois Theory can be applied to solve the problem. We will formulate here a first version of the theory which allows us already to show that (i)-(iv) do not have a compass-and-straightedge construction.

Starting with two points in the plane let us attempt to determine all possible compass-and-straightedge constructions based at these points. Without loss of generality, we can assume that the two points are the numbers 0 and 1 in the complex plane $\mathbb{C}$. We define inductively subsets $X_m \subset \mathbb{C}$, $m = 1, 2, \ldots$, as follows:

- $X_1 \equiv \{0, 1\}$, and
- $X_{m+1}$ is the union of $X_m$ with
  
  (C1) the points of intersection of straight lines connecting points of $X_m$;
  
  (C2) the points of intersection of straight lines connecting points of $X_m$ with circumferences with center in points of $X_m$ and radius segments with end points in $X_m$;
  
  (C3) the points of intersection of pairs of circumferences with center in points of $X_m$ and radius segments with end points in $X_m$;

A point $z \in \mathbb{C}$ is said to be constructible with compass and straightedge if it belongs to the union:

$$\mathcal{C} \equiv \bigcup_m X_m.$$ 

We call $\mathcal{C}$ the set of constructible points.

Let us give some alternative descriptions of the set $\mathcal{C}$ of constructible points. These will be complemented later with more complete descriptions, after we study Galois Theory.
Theorem 7.2.1. \(\mathcal{C}\) is the smallest subfield of \(\mathbb{C}\) which contains \(\mathbb{Q}\), and is closed for the operations of complex conjugation \((z \mapsto \bar{z})\) and taking square roots \((z \mapsto z^{\frac{1}{2}})\).

Proof. The proof is split into 3 parts:

(a) \(\mathcal{C}\) is a subfield of \(\mathbb{C}\): Assume that \(z_1, z_2 \in \mathcal{C}\). Then \(z_1 - z_2\) and \(z_1/z_2 (z_2 \neq 0)\) belong to \(\mathcal{C}\), since the usual operations of addition and multiplications of complex can be expressed geometrically exclusively using (C1), (C2) and (C3) (exercise).

(b) \(\mathcal{C}\) is closed for the operations of complex conjugation and taking square roots It is obvious that \(\mathcal{C}\) is closed for the operation of complex conjugation. On the other hand, if \(z \in \mathcal{C}\) and \(r = |z|\), we can obtain \(\sqrt{r}\) using (C1), (C2) and (C3), through the geometric construction sketched in the figure. Since bisecting an angle can also be performed using these operations, we see that \(z^{\frac{1}{2}}\) can be constructed using (C1), (C2) and (C3).

![Figure 7.2.1: Finding the square root.](image)

(c) If \(\mathbb{Q} \subset \mathcal{C}' \subset \mathbb{C}\) is any subfield closed for the operations of complex conjugation and taking square roots, then \(\mathcal{C} \subset \mathcal{C}'\): We need to show that the points of intersection of

(i) straight lines connecting points of \(\mathcal{C}'\),

(ii) straight lines connecting points of \(\mathcal{C}'\) with circumferences with center in \(\mathcal{C}'\) and radius segments with end-points in \(\mathcal{C}'\),

(iii) pairs of circumferences with center in \(\mathcal{C}'\) and radius segments with end-points in \(\mathcal{C}'\),
7.2. COMPASS-AND-STRAIGHTEDGE CONSTRUCTIONS

belong to \( \mathcal{C}' \). Observe first that if \( z = x + iy \in \mathcal{C}' \), then \( x \) and \( y \) belong to \( \mathcal{C}' \), since \( \mathcal{C}' \) is closed under conjugation. It follows that if \( ax + by + c = 0 \) is the equation of a straight line connecting points of \( \mathcal{C}' \), then \( a, b, c \in \mathcal{C}' \). Similarly, if \( x^2 + y^2 + dx + ey + f = 0 \) is the equation of a circumference with center in \( \mathcal{C}' \) and radius a segment with end-points in \( \mathcal{C}' \), then \( d, e, f \in \mathcal{C}' \).

Finally, observe that the points of intersection of two objects of this type, have coordinates certain fractions which, at most, contain square roots in the coefficients \( a, b, c, d, e, f \), hence, they belong to \( \mathcal{C}' \).

Next, we observe that one can translate the property of being constructible to an issue about the structure of fields.

**Theorem 7.2.2.** A complex number \( z \) belongs to \( \mathcal{C} \) if and only if it belongs to a subfield \( \mathbb{Q}(u_1, u_2, \ldots, u_r) \subset \mathbb{C} \), where \( u_1^2 \in \mathbb{Q} \) and \( u_{m+1}^2 \in \mathbb{Q}(u_1, \ldots, u_m) \) for \( m = 1, 2, \ldots, r-1 \).

**Proof.** Since \( \mathcal{C} \) is a field which contains \( \mathbb{Q} \) and is closed under taking square roots, it follows that fields of the form \( \mathbb{Q}(u_1, u_2, \ldots, u_r) \) with \( u_1^2 \in \mathbb{Q} \) and \( u_{m+1}^2 \in \mathbb{Q}(u_1, \ldots, u_m) \) for \( m < r \), are subfields of \( \mathcal{C} \). To complete the proof it is enough to check that the complex numbers that belong to some field \( \mathbb{Q}(u_1, u_2, \ldots, u_r) \) with \( u_1^2 \in \mathbb{Q} \) and \( u_{m+1}^2 \in \mathbb{Q}(u_1, \ldots, u_m) \) for \( m < r \), form a subfield of \( \mathcal{C} \), closed under conjugation (\( z \mapsto \bar{z} \)) and taking square roots (\( z \mapsto z^{\frac{1}{2}} \)). For this, note that if \( z \in \mathbb{Q}(u_1, u_2, \ldots, u_r) \) and \( z' \in \mathbb{Q}(u'_1, u'_2, \ldots, u'_s) \), then \( z - z' \) and \( z/z'(z' \neq 0) \) belong to \( \mathbb{Q}(u_1, u_2, \ldots, u_r, u'_1, u'_2, \ldots, u'_s) \). This field is obviously closed under conjugation and taking square roots is obvious.

As a corollary, we obtain the following criterion to exclude a complex number from \( \mathcal{C} \):

**Corollary 7.2.3.** If a number \( z \in \mathbb{C} \) is constructible, then it is algebraic over \( \mathbb{Q} \) of degree a power of 2.

**Proof.** If \( K(u) \) is an extension of \( K \) where \( u^2 \in K \), then either \( K(u) = K \) or \( [K(u) : K] = 2 \). Hence, the fields \( \mathbb{Q}(u_1, u_2, \ldots, u_r) \) with \( u_1^2 \in \mathbb{Q} \) and \( u_{m+1}^2 \in \mathbb{Q}(u_1, \ldots, u_m) \), if \( m < r \), satisfy \( [\mathbb{Q}(u_1, \ldots, u_r) : \mathbb{Q}] = 2^s \), for some integer \( s \leq r \). The corollary now follows from the theorem, if one observes that \( z \in \mathcal{C} \), then \( \mathbb{Q}(z) \subset \mathbb{Q}(u_1, \ldots, u_r) \), hence, \( [\mathbb{Q}(z) : \mathbb{Q}] = 2^t \).

We will see later, in our study of Galois Theory, that the converse of this corollary does not hold. However, we can use the corollary to show that constructions (i)–(iv) are not possible if one uses only a compass and straightedge.

Examples 7.2.4.

1. **Trisecting an angle.** One can construct a 60° degree angle with compass and straightedge. We claim that one cannot trisect an 60° angle. If this was possible, then the number \( \cos 20° + i \sin 20° \) would be constructible and, in particular, \( \cos 20° \) would be constructible. The trigonometric identity

\[
\cos 3\theta = 4\cos^3 \theta - 3\cos \theta,
\]

shows that \( \cos 20° \) is a root of the polynomial \( 4x^3 - 3x - \frac{1}{2} = 0 \). This polynomial is irreducible over \( \mathbb{Q} \) (exercise), hence the degree of \( \cos 20° \) over \( \mathbb{Q} \) is 3. By the corollary, \( \cos 20° \) is not constructible.

2. **Duplicating a cube.** We need to show that \( \sqrt[3]{2} \) is not constructible. For that, it is enough to observe that \( [\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3 \), since \( \sqrt[3]{2} \) is a root of the irreducible polynomial \( x^3 - 2 \).

3. **Constructing a regular do heptagon.** If this construction exists, then the number \( z = \cos \frac{2\pi}{7} + i \sin \frac{2\pi}{7} \) is constructible. This number is a root of the polynomial \( x^7 - 1 = (x - 1)(x^6 + x^5 + \cdots + 1) \). Since \( x^6 + x^5 + \cdots + 1 \) is irreducible over \( \mathbb{Q} \), we conclude that \( [\mathbb{Q}(z) : \mathbb{Q}] = 6 \). Since 6 is not a power of 2, the corollary shows that \( z \) is not constructible.

4. **Squaring the circle.** Starting from a circle of radius 1, we see that an affirmative answer implies that the number \( \pi \) is constructible. As we know, \( \pi \) is transcendental over \( \mathbb{Q} \), hence is not constructible.

There exist algebraic numbers of degree a power of 2 which are not constructible. We will see later that Galois Theory yields a more efficient criterion to determine if a given algebraic number is constructible. For now, we remark that the structure of certain extensions of \( \mathbb{Q} \) was the relevant factor to decide about this property.

**Exercises.**

1. Interpret the usual addition and multiplication operations on complex numbers geometrically, showing that it involves exclusively (C1), (C2) and (C3) and, hence, maybe constructed with a compass and straightedge.

---

\^4We will not discuss in this book the transcendence of \( \pi \). The existence of transcendental numbers was established for the first time by Liouville in 1844. Liouville observed that algebraic numbers cannot be the limit of rational sequences which converge “very fast”. For example, the number \( \sum_{n=1}^{\infty} 10^{-n!} \) cannot be algebraic. Hermite, in 1873, showed that the base of the natural logarithms “e” is transcendental. Finally, Lindemann, in 1882, used methods similar to Hermite to prove that \( \pi \) is transcendental. In 1874, Cantor gave an argument without using limits (see Exercises 3 and 4 in Section 7.1) showing that there exist transcendental numbers.
7.3. SPLITTING FIELDS

2. Show that the number $\sqrt{2} + i\sqrt{2}$ is constructible.

3. Show that the polynomial $q(x) = 4x^3 - 3x - \frac{1}{2}$ is irreducible over $\mathbb{Q}$.

4. Show that $\arccos \frac{1}{125}$ can be trisected.

5. Let $p$ be a prime. Show that the polynomial $x^{p-1} + x^{p-2} + \cdots + 1$ is irreducible over $\mathbb{Q}$.
   (Hint: Replace $x$ by $x + 1$ in the expression $x^{p-1} + x^{p-2} + \cdots + 1 = (x^p - 1)/(x - 1)$ and apply Eisenstein’s Criterion.)

6. Let $p$ be a prime.
   (a) Show that if a regular polygon with $p$ sides can be constructed with a compass and straightedge then $p = 2^s + 1$.
   (b) Show that if $p = 2^s + 1$ is prime then $p$ is a Fermat number: $p = 2^{2^t} + 1$.
   (c) Conclude that if a regular polygon with $p$ sides can be constructed with a compass and straightedge then $p$ is a Fermat number.

7. Decide if a regular polygon of 9 sides has a compass-and-straightedge construction.

7.3 Splitting Fields

As usual, we assume that $K$ is a field. Given a polynomial $p(x) \in K[x]$ is there some extension $L$ of $K$ where $p(x)$ splits into a product of linear factors? Since $K$ is not a priori a subfield of an algebraically closed field this is not completely obvious. Such an extension, if it exists, is necessarily obtained through some “abstract” construction. In fact, such an extension always exists and its construction is similar to the construction of the complex field $\mathbb{C}$, as an extension of $\mathbb{R}$, in the form of the “abstract” quotient $\mathbb{R}[x]/(x^2 + 1)$.

**Theorem 7.3.1.** Let $p(x) \in K[x]$ be polynomial of degree $n \geq 1$. There exists an extension $L \supset K$ where $p(x)$ splits into a product of linear terms. Moreover, one can take this extension to be of the form

$$L = K(r_1, \ldots, r_n),$$

where $r_1, \ldots, r_n$ are the roots of $p(x)$ in $L$. 

Proof. Without loss of generality, we can assume that \( p(x) \) is a monic polynomial of degree \( n \). Denoting by \( l \) the number of irreducible factors of \( p(x) \), the proof proceeds by induction on \( n - l \).

If \( n - l = 0 \), then \( p(x) \) is already a product of linear terms and the roots of \( p(x) \) belong to \( K \). Hence, the result holds in this case.

Assume now that \( n - l > 0 \). Then there exists a factor \( q(x) \) of \( p(x) \) which is irreducible of degree > 1. The field \( M = K[x]/\langle q(x) \rangle \) is an extension of \( K \), containing the root \( r = x + \langle q(x) \rangle \) of \( q(x) \), and coincides with \( K(r) \). In \( M \), we have the factorization \( q(x) = (x - r)\tilde{q}(x) \), so the polynomial \( p(x) \) splits into a product of \( \tilde{l} \) irreducible factors with \( \tilde{l} > l \). Since \( n - \tilde{l} < n - l \) the induction hypothesis gives an extension \( L \) of \( M \), where \( p(x) \) splits into a product of linear terms, and such that:

\[
L = M(r_1, \ldots, r_n),
\]

where \( r_1, \ldots, r_n \) are the roots of \( p(x) \) in \( L \). Since the root \( r \) is among the roots \( r_1, \ldots, r_n \), we see that

\[
L = M(r_1, \ldots, r_n) = K(r)(r_1, \ldots, r_n) = K(r, r_1, \ldots, r_n) = K(r_1, \ldots, r_n).
\]

This result leads to the following definition.

**Definition 7.3.2.** Let \( p(x) \in K[x] \). An extension \( L \supseteq K \) is called a **splitting extension** of \( p(x) \) if:

(i) \( p(x) \) splits in \( L \) into a product of terms of degree 1;

(ii) \( L = K(r_1, \ldots, r_n) \) where \( r_1, \ldots, r_n \) are the roots of \( p(x) \) in \( L \).

Similarly, if \( \{p_i(x)\}_{i \in I} \subset K[x] \) is a family of polynomials, an extension \( L \supseteq K \) is called a **splitting extension** of the family \( \{p_i(x)\}_{i \in I} \) if:

(i) each \( p_i(x) \) splits in \( L \) into a product of terms of degree 1;

(ii) \( L \) is generated by the roots of the polynomials \( p_i(x) \).

Often, we will use the term **splitting field** instead of **splitting extension**, if the base field is clear from the context. We will see in the next section that two splitting extensions of a polynomial (or a family) are isomorphic.

The proof of Theorem 7.3.1 gives a practical algorithm to construct a splitting extension of a polynomial, as we illustrate in the next examples.
Examples 7.3.3.

1. Consider a polynomial \( p(x) = x^2 + bx + c \) with coefficients in a field \( K \). If \( p(x) \) has a root in \( K \), then \( K \) is already a splitting field of \( p(x) \). Assume then that \( p(x) \) is irreducible. Then \( K[x]/\langle p(x) \rangle \) is an extension of \( K \), \( r_1 = x - \langle p(x) \rangle \) is a root of \( p(x) \), and \( K[x]/\langle p(x) \rangle = K(r_1) \). In \( K(r_1) \) we have \( x^2 + bx + c = (x - r_1)(x - r_2) \), hence \( K(r_1) = K(r_1, r_2) \) is a splitting field of \( x^2 + bx + c \) and \( [K(r_1, r_2) : K] = 2 \).

2. We have seen in a previous example that \( \mathbb{Q}(\sqrt{2}, \sqrt{3}) \subset \mathbb{C} \) is a splitting field of \( (x^2 - 2)(x^2 - 3) \) and that \( [\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] = 4 \).

3. The polynomial \( p(x) = x^p - 1 \) (\( p \) a prime) has the root \( r_0 = 1 \) in \( \mathbb{Q} \), so that \( x^p - 1 = (x - 1)(x^{p-1} + x^{p-2} + \cdots + 1) \). The polynomial \( x^{p-1} + x^{p-2} + \cdots + 1 \) is irreducible in \( \mathbb{Q} \) (exercise). Let \( r \in \mathbb{Q}[x]/\langle x^{p-1} + x^{p-2} + \cdots + 1 \rangle \) be the element \( x + (x^{p-1} + x^{p-2} + \cdots + 1) \). In \( \mathbb{Q}(r) \) we have \( (r^{k})^p = r^{kp} = 1 \), and that \( r, r^2, \ldots, r^{p-1} \) are all distinct, so \( x^p - 1 = \prod_{i=1}^{p-1} (x - r^i) \). We conclude that \( \mathbb{Q}(r) \) is a splitting field of \( x^p - 1 \) and that \( [\mathbb{Q}(r) : \mathbb{Q}] = p - 1 \). If \( z \in \mathbb{C} \) is a complex root of \( x^p - 1 \) distinct from 1, one checks easily that \( \mathbb{Q}(z) \) is isomorphic to \( \mathbb{Q}(r) \).

4. The polynomial \( p(x) = x^3 + x^2 + 1 \) is irreducible over \( \mathbb{Z}_2[x] \). In fact, \( p(x) \) does not have any roots in \( \mathbb{Z}_2 \), since \( p(0) = 0 + 0 + 1 = 1 \) and \( p(1) = 1 + 1 + 1 = 1 \) (mod 2). The field \( \mathbb{Z}_2[x]/\langle x^3 + x^2 + 1 \rangle \) is an extension of \( \mathbb{Z}_2 \) where \( p(x) \) has the root \( r = x + (x^3 + x^2 + 1) \). Hence, in this field we can factor \( p(x) = (x - r)(x^2 + bx + c) \). One checks by comparison that \( b = 1 + r \) and \( c = r^2 + r \). Using the relations \( r^3 = r^2 + 1 \) and \( r^4 = r^2 + r + 1 \), one checks immediately that \( r^2 \) is a root of \( x^2 + (r + 1)x + (r^2 + r) \). Hence, \( p(x) \) splits into linear factors in \( \mathbb{Z}_2(r)[x] \). This shows that \( \mathbb{Z}_2(r) \) is a splitting field of \( x^3 + x^2 + 1 \) over \( \mathbb{Z}_2 \), and that \([\mathbb{Z}_2(r) : \mathbb{Z}_2] = 3\).

5. The polynomial \( p(x) = x^3 - 2 \) is irreducible over \( \mathbb{Q} \). We form the extension \( L = \mathbb{Q}[x]/\langle x^3 - 2 \rangle \) and let \( r_1 = x + (x^3 - 2) \). Then \( L = \mathbb{Q}(r_1) \) and \( L \) the polynomial \( x^3 - 2 \) factors as \( (x - r_1)(x^2 + r_1 x + r_1^2) \). The polynomial \( x^2 + r_1 x + r_1^2 \) is irreducible over \( \mathbb{Q}(r_1) \) (exercise), so we can form a new extension \( M = \mathbb{Q}(r_1)[x]/\langle x^2 + r_1 x + r_1^2 \rangle \). Setting \( r_2 = x + (x^2 + r_1 x + r_1^2) \in M \), we find that \( M = \mathbb{Q}(r_1, r_2) \) and in \( \mathbb{Q}(r_1, r_2)[x] \) we have the factorization \( x^3 - 2 = (x - r_1)(x - r_2)(x - r_3) \). We conclude that \( \mathbb{Q}(r_1, r_2) = \mathbb{Q}(r_1, r_2, r_3) \) is a splitting field of \( x^3 - 2 \). Also, by Proposition 7.1.3, we have \([\mathbb{Q}(r_1, r_2, r_3) : \mathbb{Q}] = 3 \cdot 2 = 6\).

As we have mentioned before, the splitting field \( \mathbb{Q}(r_1, \ldots, r_n) \) of a polynomial \( p(x) = a_n x^n + \cdots + a_1 x + a_0 \in \mathbb{Q}[x] \) can always be realized as a subfield of \( \mathbb{C} \), since this field is an algebraically closed extension of \( \mathbb{Q} \). Actually, we can now show that any field admits an algebraically closed extension.
Theorem 7.3.4. Let $K$ be a field. There exists an algebraically closed extension of $K$.

Proof. We split the proof into 4 steps:

(a) There exists an extension $L_1$ of $K$ where every polynomial in $K[x]$ of degree $\geq 1$ has a root: For each $p(x) \in K[x]$ of degree $\geq 1$ choose an indeterminate $x_p$. Denote by $S$ the set of all such indeterminates. We claim that in the ring $K[S]$ the polynomials $p(x_p)$ generate a proper ideal. Assume that

$$g_1 p_1(x_{p_1}) + \cdots + g_m p_m(x_{p_m}) = 1,$$

for some $g_i = g_i(x_{p_1}, \ldots, x_{p_m}) \in K[S]$. There exists an extension $M$ of $K$ where the polynomials $p_1, \ldots, p_m$ have roots $\alpha_1, \ldots, \alpha_m$, so replacing $x_{p_i} \mapsto \alpha_i$ in (7.3.1) we obtain $0 = 1$, a contradiction. This proves the claim, and so there exists a maximal ideal $I$ in $K[S]$ which contains the polynomials $p(x_p)$. The field $L_1 \equiv K[S]/I$ is the sought extension.

(b) By(a), using induction, we construct a chain of extensions $K \subset L_1 \subset L_2 \subset \ldots \subset L_n \subset \ldots$, where for each $k$, any polynomial of $L_k[x]$ of degree $\geq 1$ has a root in $L_{k+1}$.

(c) Let $L \equiv \bigcup_i L_i$. Then $L$ is a field: if $a, b \in L$, there exists a $k$ such that $a, b \in L_k$, and one can define $a + b$ and $ab$ as the sum and product in $L_k$. This definition is independent of $k$, since the $L_i$ are successive extensions.

(d) The field $L$ is a algebraically closed extension: if $p(x) \in L[x]$ has degree $\geq 1$, then the coefficients of $p(x)$ belong to some $L_k$. Hence, $p(x) \in L_k[x]$, so $p(x)$ must have a root in $L_{k+1}$, hence in $L$.

\[\square\]

Corollary 7.3.5. Let $K$ be a field. There exists an algebraic extension $\tilde{L} \supset K$ which is algebraically closed.

Proof. Let $L \supset K$ be an algebraically closed extension and denote by $\tilde{L}$ the set of all algebraic elements of $L$ over $K$. It is easy to check that $\tilde{L}$ is a subfield of $L$ which has the desired properties. \[\square\]

\(^5\)One should not confuse $p(x)$ with $p(x_p)$!
7.3. SPLITTING FIELDS

An algebraic extension of $K$ which is algebraically closed is called an algebraic closure of $K$. We will see in the next section that any two such extensions are isomorphic. We will denote by the symbol $\overline{K}$ any algebraic closure of $K$.

The existence of a splitting field of a polynomial (Theorem 7.3.1) can be easily shown if one appeals to the algebraic closure. However, note that the proof of the existence of an algebraic closure uses precisely Theorem 7.3.1. Still, we can now easily prove the existence of a splitting field of an arbitrary family of polynomials.

**Corollary 7.3.6.** Let $\{p_i(x)\}_{i \in I} \subset K[x]$ be a family of polynomials. There exists a splitting extension $L \supset K$ for the family $\{p_i(x)\}_{i \in I}$.

**Proof.** It is enough to consider in the algebraic closure $\overline{K}$ the subfield $L$ generated by all the roots of the polynomials $\{p_i(x)\}_{i \in I}$.

Let us consider now the converse problem: when is an extension $L$ of $K$ a splitting extension of some family of polynomials with coefficients in $K$?

**Definition 7.3.7.** An algebraic extension $L$ of $K$ is called a normal extension of $K$ if every irreducible polynomial in $K[x]$ which has a root in $L$ splits into a product of linear terms in $L[x]$.

**Examples 7.3.8.**

1. $\mathbb{Q}(\sqrt{2})$ is a normal extension of $\mathbb{Q}$. To check this let $p(x) \in \mathbb{Q}[x]$ be an irreducible polynomial with a root $r = a + b\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. Then $p(x)$ is a multiple of the minimal polynomial $q(x)$ of $r$. Since

   
   $$q(x) = (x - a)^2 - 2b^2 \in \mathbb{Q}[x]$$

   
   and $p(x)$ is irreducible over $\mathbb{Q}$, we must have $p(x) = \lambda q(x)$, for some $\lambda \in \mathbb{Q}$. Finally, in $\mathbb{Q}(\sqrt{2})$ we have $p(x) = c(x - a - b\sqrt{2})(x - a + b\sqrt{2})$.

2. $\mathbb{Q}(\sqrt{2})$ is not a normal extension of $\mathbb{Q}$. In fact, the polynomial $x^4 - 2$ is irreducible over $\mathbb{Q}$, and has the root $\sqrt{2}$ in $\mathbb{Q}(\sqrt{2})$. However, this polynomial does not split into linear terms in $\mathbb{Q}(\sqrt{2})$, since this field does not contain the imaginary roots of $x^4 - 2$.

The last example suggest that there must exist some relation between splitting extensions and normal extensions. The next result clarifies completely this issue:
Theorem 7.3.9. Let $L \subset \overline{K}$ be an algebraic extension of $K$. The following statements are equivalent:

(i) $L$ is a normal extension of $K$.

(ii) $L$ contains a splitting field of the minimal polynomial of any $u \in L$.

(iii) $L$ is a splitting extension of a family $\{p_i(x)\}_{i \in I} \subset K[x]$.

(iv) Every monomorphism $\phi : L \to \overline{K}$ such that $\phi|_K = id$ is an automorphism of $L$: $\phi(L) = L$.

We defer the proof of this theorem for the next section. An immediate corollary is the following:

Corollary 7.3.10. An extension $L$ of $K$ is a splitting extension of a polynomial $p(x) \in K[x]$ if and only if $L$ is a normal, finite-dimensional extension of $K$.

We have provide above a simple example of a non normal extension. If an algebraic extension is not normal, one can try to find a larger normal extension which is not “too big”.

Proposition 7.3.11. Let $L$ be an algebraic extension of $K$. There exists an extension $\tilde{L} \supset L$ such that:

(i) $\tilde{L}$ is a normal extension of $K$.

(ii) for any normal extension $M$ of $K$ such that $L \subset M \subset \tilde{L}$, one has $M = \tilde{L}$.

(iii) $[\tilde{L} : K] < \infty$ if and only if $[L : K] < \infty$.

Proof. Let $S = \{u_i : i \in I\}$ be a basis for $L$ over $K$ and for each $i \in I$ denote by $p_i(x) \in L[x]$ the minimal polynomial of $u_i$. Let $\tilde{L}$ be a splitting extension for the family $\{p_i(x)\}_{i \in I}$. By Theorem 7.3.9, $\tilde{L}$ is a normal extension of $K$, so (i) holds. Also by Theorem 7.3.9 any normal extension of $K$ must contain a splitting extension of the minimal polynomial of any of its elements so we conclude that $\tilde{L}$ is the smallest normal extension of $K$ that contains $L$, and (ii) also holds. Finally, to see that (iii) holds, note that if $[L : K] < \infty$, then $I$ is finite and we can replace the family $\{p_i(x)\}_{i \in I}$ by the polynomial $p(x) = \prod_{i \in I} p_i(x)$. It follows that $[\tilde{L} : K] < \infty$. Conversely, if $[\tilde{L} : K] < \infty$ then, by Proposition 7.3.9, $[L : K] < \infty$. 

$\square$
7.3. SPLITTING FIELDS

From the uniqueness of the the splitting extension it follows that the extension \( \tilde{L} \) given in the previous proposition is unique up to an isomorphism. Such an extension is called the normal closure of \( L \) and is denoted \( \overline{L} \).

Example 7.3.12.

We saw in a example above that \( L = \mathbb{Q}(\sqrt{2}) \) is not a normal extension of \( \mathbb{Q} \). It is easy to check that its normal closure is the extension \( \overline{L} = \mathbb{Q}(\sqrt{2}, i) \).

Exercises.

1. Let \( L = \mathbb{Q}[x]/(x^3 - 2) \) and \( r = x + (x^3 - 2) \in L \). Show that the polynomial \( x^2 + rx + r^2 \) is irreducible over \( L \).

2. Prove the following extension of Theorem 7.3.1, without using the algebraic closure: If \( p_1(x), \ldots, p_m(x) \in K[x] \), then there exists a splitting extension of the family of polynomials \( p_1(x), \ldots, p_m(x) \).

3. For the following polynomials over \( \mathbb{Q} \) find splitting extensions and determine their dimensions:
   (a) \( x^2 + x + 1 \);
   (b) \( (x^3 - 2)(x^2 - 1) \);
   (c) \( x^6 + x^3 + 1 \);
   (d) \( x^5 - 7 \).

4. Consider the polynomial \( x^3 - 3 \). Find a splitting extension and its dimension over each of the following ground fields:
   (a) \( \mathbb{Q} \);
   (b) \( \mathbb{Z}_3 \);
   (c) \( \mathbb{Z}_5 \).

5. If \( p(x) \in K[x] \) is a polynomial of degree \( n \), and \( L \) is a splitting extension of \( p(x) \), show that \( [L : K] \mid n! \).

6. Determine a splitting extension of the polynomial \( x^{p^n} - 1 \) over the field \( \mathbb{Z}_p \).

7. Show that the extension \( \tilde{L} \) constructed in the proof of Corollary 7.3.5 is algebraically closed.

8. If \( L \) is an extension of \( K \) and \( [L : K] = 2 \), show that \( L \) must be a normal extension of \( K \).
9. If \( M \supset L \supset K \) are successive extensions, with \( M \) normal over \( L \) and \( L \) normal over \( K \), is it true that \( M \) must be normal over \( K \)?

10. Show that if \( M \supset L \supset K \) are successive extensions and \( M \) is normal over \( K \), then \( M \) is normal over \( L \).

### 7.4 Homomorphisms of Extensions

#### Definition 7.4.1.
Let \( L_1 \supset K_1 \) and \( L_2 \supset K_2 \) be extensions. A field homomorphism \( \phi : L_1 \rightarrow L_2 \) such that \( \phi(K_1) \subset K_2 \) is called a **homomorphism of extensions**. When \( K_1 = K_2 = K \) and \( \phi|_K = \text{id} \) we say that \( \phi \) is a **homomorphism over** \( K \) or a **\( K \)-homomorphism**.

Notice that a \( K \)-homomorphism \( \phi : L_1 \rightarrow L_2 \) is the same thing as a field homomorphism which is also a linear transformation from \( L_1 \) to \( L_2 \), as vector spaces over \( K \).

We are interested in following questions: Given an isomorphism of fields \( \phi : K_1 \rightarrow K_2 \), when can one extend \( \phi \) to an isomorphism of extensions \( \Phi : L_1 \rightarrow L_2 \)? How many such extensions are there? The answers will allow us to prove uniqueness of algebraic and normal closures, as well as of splitting fields.

#### Proposition 7.4.2.
Let \( \phi : K_1 \rightarrow K_2 \) be an isomorphism of fields, \( L_1 \supset K_1 \) and \( L_2 \supset K_2 \) extensions, and \( r \in L_1 \) algebraic over \( K_1 \) with minimal polynomial \( p(x) \). The isomorphism \( \phi \) can be extended to a monomorphism \( \Phi : K_1(r) \rightarrow L_2 \) if and only if the polynomial \( p^\phi(x) \) has a root in \( L_2 \). The number of such extensions equals the number of distinct roots of \( p^\phi(x) \) in \( L_2 \).

**Proof.** Let \( s \) be a root of \( p^\phi(x) \) in \( L_2 \). It is easy to check that there exist a unique field homomorphism \( \Phi : K_1(r) \rightarrow L_2 \) such that \( \Phi|_{K_1} = \phi \) and \( \Phi(r) = s \).

On the other hand, if \( \Phi \) is any extension of \( \phi \), then

\[
p^\phi(\Phi(r)) = \Phi(p(r)) = \Phi(0) = 0,
\]

so \( p^\phi(x) \) has a root in \( L_2 \).

Here is a first answer to the questions posed above:

---

6Recall that a homomorphism of fields is necessarily injective, hence in these definitions one can replace “homomorphism” by “monomorphism”.
**Theorem 7.4.3.** Let \( \phi : K_1 \to K_2 \) be an isomorphism of fields and \( p(x) \in K_1[x] \). If \( L_1 \) and \( L_2 \) are splitting extensions of \( p(x) \) and \( p^\phi(x) \), respectively, then there exists an isomorphism \( \Phi : L_1 \to L_2 \) which extends \( \phi \). The number of such extension is \( \leq [L_1 : K_1] \), and equals \( [L_1 : K_1] \) when \( p^\phi(x) \) has distinct roots in \( L_2 \).

**Proof.** We use induction on \( [L_1 : K_1] \).

If \( [L_1 : K_1] = 1 \), then \( p(x) = a_n \prod_{i=1}^n (x - r_i) \), where \( r_i \in K_1 = L_1 \). But then \( p^\phi(x) = \phi(a_n) \prod_{i=1}^n (x - \phi(r_i)) \), with \( \phi(a_n), \phi(r_i) \in K_2 \). Since the roots of a polynomial generate its splitting field we conclude that \( L_2 = K_2 \), and that there exists exactly \( [L_1 : K_1] = 1 \) extension.

Assume now that \( [L_1 : K_1] > 1 \). Then \( p(x) \) has an irreducible factor \( q(x) \in K_1[x] \) of degree \( \geq 1 \). Let \( r \) be a root of \( q(x) \) in \( L_1 \). By Proposition 7.4.2 the isomorphism \( \phi : K_1 \to K_2 \) can be extended to a monomorphism \( \phi : K_1 \cdot r \to L_2 \) and there exist as many such extensions as distinct roots of \( q^\phi(x) \) in \( L_2 \). We can also view \( L_1 \) and \( L_2 \) as splitting extensions of \( p(x) \) and \( p^\phi(x) \) over \( K_1(r) \) and \( \tilde{\phi}(K_1(r)) \), respectively. Since

\[
[L_1 : K_1(r)] = [L_1 : K_1]/[K_1(r) : K_1] = [L_1 : K_1]/\deg q(x) < [L_1 : K_1],
\]

by the induction hypothesis, there exists an extension of \( \tilde{\phi} \) to an isomorphism \( \Phi : L_1 \to L_2 \), and the number of such extensions is \( \leq [L_1 : K_1(r)] \), and equal \( [L_1 : K_1(r)] \) if \( p^\phi(x) \) has distinct roots in \( L_2 \). It follows that \( \Phi \) is an extension of \( \phi \), and the number of extensions \( \Phi \) of this type is

\[
[L_1 : K_1(r)] \cdot \deg q(x) = [L_1 : K_1(r)] \cdot [K_1(r) : K_1] = [L_1 : K_1],
\]

if \( p^\phi(x) \) has distinct roots in \( L_2 \).

Finally, observe that we can obtain any extension of \( \phi \) in this way. In fact, given any extension \( \Phi : L_1 \to L_2 \) its restriction to \( K_1(r) \) gives a monomorphism \( K_1(r) \to L_2 \), which by Proposition 7.4.2 is an extension of \( \phi \) of the type \( \tilde{\phi} \), as above. \( \square \)

Setting \( K_1 = K_2 = K \) and \( \phi = \text{id} \), in this theorem, yields:

**Corollary 7.4.4.** Any two splitting fields of \( p(x) \in K[x] \) are isomorphic.

**Example 7.4.5.**

In an example in the previous section, we have constructed an (abstract) splitting field \( L_1 = \mathbb{Q}(r_1, r_2, r_3) \) of \( x^3 - 2 \in \mathbb{Q}[x] \), with \( [L_1 : \mathbb{Q}] = 6 \). Another splitting field \( L_2 \) can be obtained by taking the subfield of \( \mathbb{C} \) generated by \( \mathbb{Q} \) and the roots of \( x^3 - 2 \) in \( \mathbb{C} : \sqrt[3]{2}, \sqrt[3]{2}/2(-1 + i\sqrt{3}) \). By the theorem, all the 6 bijections between \( \{r_1, r_2, r_3\} \) and \( \{\sqrt[3]{2}, \sqrt[3]{2}/2(-1 + i\sqrt{3})\} \) can be realized by a \( \mathbb{Q} \)-isomorphism \( L_1 \to L_2 \).
Let us consider now the case of arbitrary algebraic extensions:

**Theorem 7.4.6.** Let $\phi : K_1 \to K_2$ be an isomorphism of fields, $L_1 \supseteq K_1$ an algebraic extension and $L_2 \supseteq K_2$ an algebraically closed extension. There exists a monomorphism $\Phi : L_1 \to L_2$ which extends $\phi$. If $L_1$ is algebraically closed and $L_2$ is algebraic over $K_2$, then $\Phi$ is an isomorphism $L_1 \cong L_2$.

**Proof.** Consider the set $\mathcal{P}$ formed by ordered pairs $(N, \tau)$ where $N \subseteq L_1$ is an extension of $K_1$, and $\tau : M \to L_2$ is a monomorphism which extends $\phi$. In $\mathcal{P}$ one defines a partial order relation by declaring $(N_1, \tau_1) \leq (N_2, \tau_2)$ if and only if $N_1 \subseteq N_2$ and $\tau_2|N_1 = \tau_1$. The set $\mathcal{P}$ is non-empty, since it contains the pair $(K_1, \phi)$, and any chain $\{(N_i, \tau_i)\}_{i \in I}$ of elements of $\mathcal{P}$ is bounded above by the pair $(N, \tau)$, where $N \equiv \bigcup_i N_i$ and $\tau|N_i \equiv \tau_i$. By Zorn’s Lemma, $\mathcal{P}$ contains a maximal element $(M, \Phi)$.

We claim that $M = L_1$, so that $\Phi$ gives the desired extension. In fact, assume that there exists $r \in L_1 - M$. We can then form the extension $M(r)$ and, by Proposition 7.4.2, there exists an extension $\tilde{\Phi} : M(r) \to L_2$. The pair $(M(r), \tilde{\Phi})$ contradicts the maximality of $(M, \Phi)$, so we must have $L_1 = M$, as claimed.

Finally, if $L_1$ is algebraically closed, then $\Phi(L_1)$ is also algebraically closed. If $L_2$ is algebraic over $K_2$, then we must have $L_2 \subseteq \Phi(L_1)$, hence $\Phi$ is onto.

The previous result leads to the uniqueness of the algebraic closure and of the splitting extension of any family of polynomials. For example, if one takes $K_1 = K_2 = K$ and $\phi = \text{id}$ in the theorem, we obtain immediately:

**Corollary 7.4.7.** Let $K$ be a field, $L$ and $\tilde{L}$ algebraically closed extensions, algebraic over $K$. There exists a $K$-isomorphism $\Phi : L \to \tilde{L}$.

Since a splitting extension of a family of polynomials can be taken inside an algebraic closure, and it is generated by the all roots of the polynomials in the family, we also obtain:

**Corollary 7.4.8.** Let $\{p_i(x)\}_{i \in I} \subseteq K[x]$ be a family of polynomials. Any two splitting extensions of the family $\{p_i(x)\}_{i \in I}$ are isomorphic.

Finally, the uniqueness of the normal closure of an extension $L$ of $K$ follows from Theorem 7.3.9 whose proof we are now able to furnish.

**Proof of Theorem 7.3.9.** We will prove the implications $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. 

7.4. HOMOMORPHISMS OF EXTENSIONS

(i) ⇒ (ii): Let \( r \in L \) and denote by \( p(x) \in K[x] \) the minimal polynomial of \( r \). Since \( L \) is normal and \( p(r) = 0 \), \( p(x) \) splits into a product \( \prod_{i=1}^{n}(x-r_i) \) in \( L[x] \). The subfield \( K(r_1, \ldots, r_n) \subset L \) is a splitting field of \( p(x) \).

(ii) ⇒ (iii): \( L \) is a splitting extension of the family \( \{p_r(x)\}_{r \in L} \), where \( p_r(x) \) is the minimal polynomial of the element \( r \in L \).

(iii) ⇒ (iv): Let \( r \in L \subset \mathbb{K} \) be a root of a polynomial \( p_i(x) \). Then \( \phi(r) \) is also a root of \( p_i(x) \), since \( \phi|_{\mathbb{K}} = \text{id} \). Since the roots of the family \( \{p_i(x)\}_{i \in I} \) generate the field \( L \), we conclude that \( \phi(L) = L \).

(iv) ⇒ (i): Let \( p(x) \in K[x] \) be an irreducible polynomial with a root \( r \in L \). If \( \tilde{r} \in \mathbb{K}^a \) is any other root of \( p(x) \), the map \( \phi : K(r) \to K(\tilde{r}) \subset \mathbb{K}^a \) which takes \( r \mapsto \tilde{r} \) and is the identity in \( K \), is a \( K \)-isomorphism, which by Theorem 7.4.6 can be extended to \( L \). Then \( \tilde{r} = \phi(r) \in L \), hence all the roots of \( p(x) \) belong to \( L \), so \( p(x) \) factors into a product of linear terms in \( L[x] \). Hence, \( L \) is a normal extension of \( K \).

Exercises.

1. Let \( \phi : K_1 \to K_2 \) be an isomorphism, \( L_1 \supset K_1 \) and \( L_2 \supset K_2 \) extensions, and \( r \in L_1 \) algebraic over \( K_1 \) with minimal polynomial \( p(x) \). If \( s \) is a root of \( p_\phi(x) \) in \( L_2 \), show that there exists a unique homomorphism of fields \( \Phi : K_1(r) \to L_2 \) such that \( \Phi|_{K_1} = \phi \) and \( \Phi(r) = s \).

2. Consider the extensions \( \mathbb{Q}(i), \mathbb{Q}(\sqrt{2}), \mathbb{Q}(i, \sqrt{2}) \subset \mathbb{C} \) of \( \mathbb{Q} \).

(a) Show that \( \mathbb{Q}(i) \) and \( \mathbb{Q}(\sqrt{2}) \) are isomorphic \( \mathbb{Q} \)-vector spaces, but are not isomorphic as fields.

(b) Find all the \( \mathbb{Q} \)-automorphisms of \( \mathbb{Q}(i, \sqrt{2}) \).

3. Let \( r \) be a root of \( x^6 + x^3 + 1 \). Determine all the \( \mathbb{Q} \)-homomorphisms \( \Phi : \mathbb{Q}(r) \to \mathbb{C} \).

(Hint: \( x^6 + x^3 + 1 \) is a factor of \( x^9 - 1 \).)

4. Determine all the \( \mathbb{Z}_2 \)-automorphisms of the splitting extension of \( x^3 + x^2 + 1 \in \mathbb{Z}_2[x] \).

5. Show that an algebraic extension \( L \) of \( K \) is a normal extension if and only if every irreducible polynomial of \( K[x] \) factors in \( L[x] \) into a product of irreducible factors where all the factors have the same degree.
7.5 Separability

As it will be clear in the sequel, the $K$-automorphisms of an extension $L \supset K$ play a fundamental role in Galois Theory. As we have seen in the previous section, the precise counting of the number of $K$-automorphisms depend on the number of distinct roots of certain polynomials. We will now study when a given polynomial has multiple roots and we will establish a criterion for its existence.

Any polynomial $p(x) \in K[x]$ splits into a product of linear terms in $K^a[x]$. Hence, if $r \in K^a$ is a root of $p(x)$, we can define its multiplicity to be the largest integer $k$ such that $(x-r)^k \mid p(x)$ in $K^a[x]$. The root is called simple if $k = 1$, and multiple if $k > 1$.

The criterion for the existence of multiple roots will require a notion of derivative of a polynomial. When the polynomial can be viewed as real or complex valued function, this is just the usual notion of derivative, but in general this will be only a formal derivative.

**Definition 7.5.1.** Let $K$ be a field. The formal derivative operator of $K[x]$ is the only map $D: K[x] \to K[x]$ which satisfies the following properties:

(i) **Linearity:** $D(p(x) + q(x)) = D(p(x)) + D(q(x))$ and $D(a \cdot p(x)) = a \cdot D(p(x))$.

(ii) **Leibniz Rule:** $D(p(x)q(x)) = D(p(x))q(x) + p(x)D(q(x))$.

(iii) **Normalization:** $D(x) = 1$.

One checks easily that there is exactly one operator satisfying these properties. In fact, if $p(x) = a_nx^n + \cdots + a_1x + a_0 \in K[x]$, then by applying (i), (ii) and (iii), we find:

$$D(p(x)) = na_nx^{n-1} + \cdots + 2a_2x + a_1.$$  \hspace{1cm} (7.5.1)

We will often write $p'(x)$ instead of $D(p(x))$.

**Theorem 7.5.2.** Let $p(x) \in K[x]$ be a monic polynomial of degree $\geq 1$. The roots of $p(x)$ in $K^a$ are simple if and only if $\gcd(p(x), p'(x)) = 1$.

**Proof.** If $r \in K^a$ is a multiple root of $p(x)$, then $p(x) = (x-r)^k q(x)$, where $k > 1$ and $q(x) \in K^a[x]$. Differentiating:

$$p'(x) = (x-r)^k q'(x) + k(x-r)^{k-1} q(x),$$
so we see that \((x - r)^{k-1} \mid p'(x)\), and \((x - r)\) is a common factor of \(p(x)\) and \(p'(x)\). We conclude that if \(\gcd(p(x), p'(x)) = 1\), then \(p(x)\) has no multiple roots.

Assume now that \(p(x)\) has no multiple roots. Then in \(\mathbb{K}^a[x]\) we have \(p(x) = (x-r_1)(x-r_2)\cdots(x-r_n)\), where the \(r_i\) are all distinct. Differentiating we obtain:

\[
p'(x) = \sum_{i=1}^{n} (x-r_1)\cdots(x-r_{i-1})(x-r_{i+1})\cdots(x-r_n),
\]

so that \((x - r_i) \nmid p'(x)\), for all \(i = 1, \ldots, n\). Hence, \(\gcd(p(x), p'(x)) = 1\).

Notice that to apply this criterion we do not need to know the algebraic closure \(\mathbb{K}^a\) or even a splitting field of \(p(x)\): the computation of \(\gcd(p(x), q(x))\) is independent of the extension where one takes the coefficients of the polynomials \(p(x)\) and \(q(x)\).

**Definition 7.5.3.** One calls \(p(x) \in \mathbb{K}[x]\) a **separable polynomial** if each of its irreducible factors has no multiple roots. A **perfect field** is a field \(K\) where every polynomial in \(\mathbb{K}[x]\) is separable.

The criterions above leads immediately to:

**Corollary 7.5.4.** Every field of characteristic zero is perfect.

**Proof.** Let \(p(x) \in \mathbb{K}[x]\) be a monic irreducible polynomial with multiple roots. Then \(\gcd(p(x), p'(x)) \neq 1\), so we must have \(p(x) \mid p'(x)\). Since \(\deg p'(x) < \deg p(x)\), it follows that \(p'(x) = 0\). Since the characteristic of \(K\) is assumed to be zero, formula (7.5.1) for the derivative shows that \(p(x) \in \mathbb{K}\), a contradiction.

The proof fails in characteristic \(p\), since then the condition \(q'(x) = 0\) only implies that \(q(x) = a_0 + a_1x^p + a_2x^{2p} + \ldots\). One can then find examples of non-separable polynomials in characteristic \(p \neq 0\) (see the exercises). When \(p \neq 0\) the study of separability is based on the following simple lemma:

**Lemma 7.5.5.** If \(K\) has characteristic \(p \neq 0\) and \(a \in \mathbb{K}\), then the polynomial \(x^p - a\) is either irreducible or takes the form \((x - b)^p\), \(b \in \mathbb{K}\).

**Proof.** Suppose that \(x^p - a = g(x)h(x)\), where \(g(x)\) and \(h(x)\) are monic polynomials. Let \(b \in \mathbb{K}^a\) be a root of \(x^p - a\). Then

\[
(x - b)^p = \sum_{k=0}^{p} \binom{p}{k} x^k (-b)^{p-k} = x^p - b^p = x^p - a,
\]
since for $0 < k < p$ the coefficients $\binom{p}{k}$ are all divisible by $p$. Hence, $g(x) = (x - b)^p$ and we must have $b^p \in K$. Since $\gcd(k, p) = 1$, there exist $r, s \in \mathbb{Z}$ such that $rk + sp = 1$, and it follows that $b = b^{rk+ksp} = (b^k)^r + (b^p)^s \in K$. This shows that $x^p - a = (x - b)^p$, with $b \in K$.

\textbf{Theorem 7.5.6.} A field $K$ of characteristic $p \neq 0$ is perfect if and only if $K = K^p$, where $K^p$ denotes the subfield of $K$ formed by all the powers $a^p$, with $a \in K$.

\textit{Proof.} Let $a \in K - K^p$. By Lemma 7.5.5, $x^p - a$ is irreducible and since $(x^p - a)' = px^{p-1} = 0$, this polynomial is not separable. Hence if $K \neq K^p$ then $K$ is not perfect.

Conversely, assume that $K$ is not perfect, so there exists $q(x) \in K[x]$ irreducible and non-separable. Since $q'(x) = 0$, we have $q(x) = a_0 + a_1x^p + a_2x^{2p} + \ldots$. Note that at least one $a_j \notin K^p$, for if $a_j = b_j^p$ with $b_j \in K$ for all $j$, then $q(x) = (b_0 + b_1x + b_2x^2 + \ldots)^p$, which contradicts the assumption that $q(x)$ is irreducible. Hence, $K \neq K^p$. \hfill $\square$

If $K$ is a field of characteristic $p \neq 0$, the monomorphism $\phi : K \to K$, $a \mapsto a^p$, is called the \textsc{Frobenius Endomorphism}. The previous theorem states that $K$ is perfect if and only if the Frobenius endomorphism is an automorphism.

\textbf{Corollary 7.5.7.} If $K$ is finite of characteristic $p \neq 0$, then $K$ is a perfect.

\textit{Proof.} If $K$ is finite, the Frobenius endomorphism is surjective. \hfill $\square$

In most examples that we will present in our discussion of Galois Theory the ground field will be a perfect field. When this is not the case, in order for Galois Theory to work, we need the following extra assumption:

\textbf{Definition 7.5.8.} Let $L$ be an extension of $K$.

(i) $u \in L$ is called \textsc{separable element} over $K$ if $u$ is algebraic over $K$, and the minimal polynomial of $u$ is separable.

(ii) $L$ is called a \textsc{separable extension} over $K$ if all elements of $L$ are separable over $K$.

Given an algebraic extension $L \supset K$ one defines its \textsc{separable degree}, denoted $[L : K]_s$, to be the cardinality of the set of $K$-homomorphisms from $L$ to the algebraic closure $\overline{K}$:

$$[L : K]_s := \#\{\phi : L \to \overline{K} : \phi \text{ is a } K\text{-homomorphism}\}.$$
Proposition 7.5.9 (Properties of the separable degree). Let $K$ be a field.

(i) If $M \supset L \supset K$ are algebraic extensions then

$$[M : K]_s = [M : L]_s \cdot [L : K]_s.$$ 

(ii) If $L$ is an extension of finite dimension over $K$, then $[L : K]_s \leq [L : K]$, with equality precisely when $L$ is separable over $K$.

(iii) If $L = K(u_1, \ldots, u_m)$, then $L$ is separable over $K$ if and only if $u_1, \ldots, u_m$ are separable over $K$.

Proof.

(i) If $\phi : M \to \overline{K}^a$ is a $K$-homomorphism, then $\phi$ is an extension of the $K$-homomorphism $\phi|_L : L \to \overline{K}^a$. Hence, it is enough to show that for a $K$-homomorphism $\psi : L \to \overline{K}^a$ its extensions to a $K$-homomorphism $\phi : M \to \overline{K}^a$ are in bijection with the $L$-homomorphisms $\tau : M \to \overline{L}^a$. In fact, we can extend $\psi$ to an isomorphism $\tilde{\psi} : \overline{L}^a \to \overline{K}^a$, hence the correspondence that to a $L$-homomorphism $\tau : M \to \overline{L}^a$ associates the extension $\phi = \tilde{\psi} \circ \tau$ is a bijection: the inverse associates to an extension $\phi : M \to \overline{K}^a$ the $L$-homomorphism $\tau = \tilde{\psi}^{-1} \circ \phi$.

(ii) By (i), it is enough to show that $[K(u) : K]_s \leq [K(u) : K]$ for any algebraic element $u$ over $K$, with equality if and only if $K(u)$ is separable over $K$. Since $[K(u) : K]_s$ represents the number of $K$-homomorphisms $\phi : K(u) \to \overline{K}^a$, it equals the number of distinct roots of the minimal polynomial of $u$ over $K$. Hence, $[K(u) : K]_s \leq [K(u) : K]$, with equality if and only if $u$ is separable over $K$. But $u$ is separable over $K$ if and only if $K(u)$ is a separable extension over $K$: if $u' \in K(u)$ is not separable, then

$$[K(u) : K]_s = [K(u) : K(u')]_s [K(u') : K]_s$$

$$< [K(u) : K(u')] [K(u') : K] = [K(u) : K].$$

(iii) If the $u_i$ are separable over $K$, then $u_i$ is separable over $K(u_1, \ldots, u_{i-1})$. From (i) and (ii), we conclude that

$$[K(u_1, \ldots, u_m) : K]_s = [K(u_1, \ldots, u_m) : K],$$

so $K(u_1, \ldots, u_m)$ is separable over $K$. \hfill \qed

If $K$ is perfect, then obviously all algebraic extensions of $K$ are separable. In particular, if $K$ has characteristic 0 or $K$ has characteristic $p \neq 0$ and $K = K^p$, then every algebraic extension of $K$ is separable.
Exercises.

1. Show that if $K$ has characteristic 0 and $p(x) \in K[x]$ is monic, then $q(x) = p(x)[\gcd(p(x), p'(x))^{-1}$ is a polynomial whose roots are all simple, and that these are precisely the roots of $p(x)$.

2. Show that if $K$ is a field with characteristic $\neq 2$, then any quadratic polynomial $x^2 + ax + b \in K[x]$ is separable.

3. Consider the field $\mathbb{Z}_p(x)$ of fractions $q(x)/r(x)$, where $q(x)$ and $r(x) \neq 0$ are polynomials over $\mathbb{Z}_p$.
   (a) Show that $\mathbb{Z}_p(x)$ has characteristic $p$.
   (b) Show that the element $x \in \mathbb{Z}_p(x)$ is not a power of degree $p$, i.e., there is no $b(x) \in \mathbb{Z}_p(x)$ such that $x = b(x)^p$.
   (c) Given an example of a polynomial over $\mathbb{Z}_p(x)$ which is not separable.

4. Let $K$ be a field of characteristic $p$, and $q(x) \in K[x]$ an irreducible polynomial. Show that all the roots of $q(x)$ have the same multiplicity $p^n$, for some integer $n$.

5. Let $L$ be an extension of $K$, with $[L : K] < \infty$. Show that $[L : K]_s \mid [L : K]$.
   (Note: One calls $[L : K]_s = [L : K]/[L : K]_s$ the inseparable degree of $L$ over $K$.)

6. Let $M \supset L \supset K$ be successive extensions. Show that:
   (i) If $M$ is separable over $K$, then $M$ is separable over $L$ and $L$ is separable over $K$;
   (ii) If $M$ is separable over $L$ and $L$ is separable over $K$, then $M$ is separable over $K$.

7.6 The Galois Group

As we have already mentioned, Galois’s great insight was to replace questions about field extensions by questions in group theory. We are now ready to introduce the relevant groups.

If $L$ is an extension of $K$, obviously the $K$-automorphisms of $L$ form a group under composition: if $\phi_1$ and $\phi_2$ are $K$-automorphisms of $L$, then $\phi_1 \circ \phi_2$ is also a $K$-automorphism.

**Definition 7.6.1.** The group of $K$-automorphisms of an extension $L$ of $K$ is called the Galois group of $L$ over $K$ and is denoted $\text{Aut}_K(L)$. 
The next examples show that this group can be of a very different nature, depending on the extension.

Examples 7.6.2.

1. Let \( L = \mathbb{Q}(\sqrt{2}) \). The element \( \sqrt{2} \) has minimal polynomial \( x^2 - 2 \). Any \( \mathbb{Q} \)-automorphism \( \phi : L \to L \) transforms roots of this polynomial into roots. Hence, we have two automorphisms, namely the identity and
\[
\phi(a + b\sqrt{2}) = a - b\sqrt{2}.
\]
We conclude that the Galois group \( \text{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt{2})) \) is isomorphic to \( \mathbb{Z}_2 \).

2. Let \( L = \mathbb{Q}(\sqrt{2}, \sqrt{3}) \). As in the previous example we see that the \( \mathbb{Q} \)-automorphisms of \( L \) are completely determined by its action on the set \( \{\sqrt{2}, \sqrt{3}\} \). There are 4 possibilities: the identity and
\[
\phi_1(\sqrt{2}) = -\sqrt{2}, \quad \phi_1(\sqrt{3}) = \sqrt{3};
\]
\[
\phi_2(\sqrt{2}) = \sqrt{2}, \quad \phi_2(\sqrt{3}) = -\sqrt{3};
\]
\[
\phi_3(\sqrt{2}) = -\sqrt{2}, \quad \phi_3(\sqrt{3}) = -\sqrt{3}.
\]
In this case the Galois group is isomorphic to \( \mathbb{Z}_2 \oplus \mathbb{Z}_2 \).

3. Let \( K \) be a field of characteristic \( p \) such that \( K \neq K^p \). If \( a \notin K^p \), the polynomial \( q(x) = x^p - a \) is irreducible. Let \( L \) be a splitting extension of \( q(x) \). In \( L \) we have \( q(x) = (x-r)^p \), so \( L = K(r) \). If \( \phi : L \to L \) is a \( K \)-automorphism, then \( \phi(r) = r \) and we conclude that \( \phi = \text{id} \). The Galois group \( \text{Aut}_K(L) \) is trivial.

4. Let \( L = K(x) \) be the field of fractions \( p(x)/q(x) \) with \( p(x), q(x) \in K[x] \) and \( q(x) \neq 0 \). It is easy to see that the primitive elements of \( L \) are all of the form
\[
t = \frac{ax + b}{cx + d}, \quad a, b, c, d \in K, \quad ad - bc \neq 0.
\]
Any \( K \)-automorphism of \( L \) transforms primitive elements into primitive elements. Hence, a \( K \)-automorphism \( \phi : L \to L \), transforms the element \( p(x)/q(x) \in L \) in an element \( p(t)/q(t) \in L \). We conclude that the Galois group is isomorphic to \( \text{PGL}_2(K) := \text{GL}_2(K)/Z \), where \( Z \) is the center of \( \text{GL}_2(K) \) consisting of \( 2 \times 2 \) matrices of the form
\[
Z = \{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} : a \neq 0 \}.
\]
If \( K \) is infinite, this group is infinite.

Since we know now that all the splitting fields of a polynomial \( p(x) \) are isomorphic, we can set:
**Definition 7.6.3.** The **Galois group of** \( p(x) \in K[x] \) (or of \( p(x) = 0 \)) is the Galois group of any splitting field of \( p(x) \) over \( K \).

It is natural to identify the Galois group of an equation \( p(x) = 0 \) with the subgroup of group of permutations of the roots of \( p(x) \). If \( L \) is a splitting field of \( p(x) \), and \( S = \{ r_1, \ldots, r_n \} \) is the set of distinct roots of \( p(x) \), then \( L = K(r_1, \ldots, r_n) \). If \( \phi \in \text{Aut}_K(L) \) is an element of the Galois group of \( p(x) \), then \( \phi \) transforms roots of \( p(x) \) into roots, so it induces a permutation of \( S \). On the other hand, if we know the effect of \( \phi \) on the roots of \( p(x) \), then we know how \( \phi \) transforms any element of \( L = K(r_1, \ldots, r_n) \). Hence, the map \( \phi \mapsto \phi|_S \) is a monomorphism \( \text{Aut}_K(L) \to S_n \). This shows that we can identify \( \text{Aut}_K(L) \) with a subgroup of the group of permutations the roots.

In general, as it was already clear in the examples above, \( \text{Aut}_K(L) \subseteq S_n \), even when \( p(x) \) is irreducible.

**Example 7.6.4.**

Let \( L \subset \mathbb{C} \) be a splitting extension of polynomial \( p(x) = x^6 - 2 \in \mathbb{Q}[x] \).

\[ r_2 \]
\[ r_1 \]
\[ r_3 \]
\[ r_4 \]
\[ r_5 \]
\[ r_6 \]

Figure 7.6.1: Roots of \( x^6 - 2 = 0 \).

This polynomial is irreducible with roots \( r_k = \sqrt[3]{2} e^{\frac{2\pi k}{6}} \), \( k = 1, \ldots, 6 \). Notice that, for example

\[ r_3 + r_6 = 0. \]

\(^7\)It was precisely in this way that Galois thought of this group, much before the abstract notion of group was discovered!
7.6. THE GALOIS GROUP

Since \( r_3 + r_1 \neq 0 \), there is no automorphism in the Galois group corresponding to the transposition (16). From the figure below, note also that

\[(r_1 + r_5)^6 = r_6^6 = 2,\]

hence there are no automorphisms of the Galois group corresponding to the permutations (13)(56) and (16)(35). One can exclude many other elements of \( S_6 \) in this way. In fact, we will see shortly that \(|G| = 12\).

The computation of the Galois group of an equation \( p(x) = 0 \) or of an extension maybe a very hard task. However, we have already developed the tools to find its order.

**Theorem 7.6.5.** Let \( L \) be an extension of finite dimension over \( K \), and \( G = \text{Aut}_K(L) \) its Galois group. Then \(|G| \leq [L : K]\). If \( L \) a normal, separable, extension of \( K \) then \(|G| = [L : K]\).

**Proof.** We can assume that \( L \subset \bar{K}^a \). If \( \phi \in G \), then we obtain a \( K \)-homomorphism \( \phi : L \to \bar{K}^a \). The number of such homomorphisms is \([L : K]_s \leq [L : K]\). Hence \(|G| \leq [L : K]\).

If \( L \) is normal, then every \( K \)-homomorphism \( \psi : L \to \bar{K}^a \) is in fact an automorphism of \( L \), by Theorem 7.3.9. If \( L \) is separable, then \([L : K]_s = [L : K]\), by Proposition 7.5.9. Hence, if \( L \) is normal and separable over \( K \), then \(|G| = [L : K]\). \(\Box\)

For a separable polynomial we conclude that:

**Corollary 7.6.6.** If \( p(x) \) is a separable polynomial over \( K \) with Galois group \( G \), and \( L \) is a splitting extension of \( p(x) \), then \(|G| = [L : K]\).

**Examples 7.6.7.**

1. The extension \( \mathbb{Q}(i, \sqrt{2}) \) is a splitting extension of the polynomial \( x^4 - 2 \in \mathbb{Q}[x] \). Since \([\mathbb{Q}(i, \sqrt{2}) : \mathbb{Q}(i)] = 4 \) and \([\mathbb{Q}(i) : \mathbb{Q}] = 2\), we have \([\mathbb{Q}(i, \sqrt{2}) : \mathbb{Q}] = 8\), and the Galois group of \( \mathbb{Q}(i, \sqrt{2}) \) has order 8. We have \( \mathbb{Q} \)-automorphisms of \( \mathbb{Q}(i, \sqrt{2}) \) defined by (check this!)

\[
\sigma(i) = -i, \quad \tau(i) = i, \quad \sigma(\sqrt{2}) = \sqrt{2}, \quad \tau(\sqrt{2}) = i\sqrt{2}.
\]

These automorphisms have orders 2 and 4, respectively, and satisfy the relation \( \tau\sigma = \sigma\tau^3 \). It is then easy to see that

\[
\text{Aut}_\mathbb{Q} \mathbb{Q}(i, \sqrt{2}) = \{1, \tau, \tau^2, \tau^3, \sigma, \sigma\tau, \sigma\tau^2, \sigma\tau^3\} \equiv G.
\]
This group is isomorphic to the dihedral group $D_4$. In terms of the roots $r_k = e^{\frac{2\pi i}{3}}$, these automorphisms correspond to the permutations
\[ \sigma = (13), \quad \tau = (1234). \]

2. The Galois group $G$ of the polynomial $p(x) = x^6 - 2$ over $\mathbb{Q}$ has order 12, since $[\mathbb{Q}(\sqrt[3]{2}, e^{\frac{2\pi i}{3}}) : \mathbb{Q}] = 12$ and $\mathbb{Q}(\sqrt[3]{2}, e^{\frac{2\pi i}{3}})$ is a splitting extension of $p(x)$. We leave it as an exercise to check that $G \cong D_6$ and to determine its representation as a permutation group of the roots.

The previous examples suggest that one should be able to compute the Galois group of a polynomial $x^n - a$ over a field $K$ of characteristic 0. This depends on the ability to compute the Galois group of $x^n - 1$.

**Definition 7.6.8.** Let $K$ be a field. A splitting extension of the polynomial $x^n - 1$ is called the **cyclotomic field of order $n$ over $K$**.

The Galois group of a cyclotomic field is much simple in the case of characteristic zero.

**Proposition 7.6.9.** If $K$ has characteristic 0, the Galois group of the cyclotomic field of order $n$ is isomorphic to a subgroup of $\mathbb{Z}_n^*$.

**Proof.** Let $L$ be a splitting extension of $x^n - 1$ over $K$. Since $K$ has characteristic zero, and $(x^n - 1)' = nx^{n-1} \neq 0$, the roots of $x^n - 1$ are all simple roots. We conclude that the set of roots $U = \{ r \in L : r^n - 1 = 0 \}$ is isomorphic to $\mathbb{Z}_n$ (see Exercise 6.6.2.2). On the other hand, if $\phi \in \text{Aut}_K(L) \equiv G$, then $\phi|_U$ is an automorphism of $U$ and this restriction determines completely $\phi$. Hence $G$ is isomorphic to a subgroup of $\text{Aut}(U) \cong \text{Aut}(\mathbb{Z}_n) \cong \mathbb{Z}_n^*$, the group of units of the ring $\mathbb{Z}_n$. \[ \square \]

The previous result shows that in characteristic 0 the Galois group of a cyclotomic field is abelian. In general, there is nothing else we can say about the Galois group: for example, if $K$ contains the roots of $x^n - 1 = 0$, then obviously the cyclotomic field of order $n$ coincides with $K$, so its Galois group is trivial. However, when $K = \mathbb{Q}$, the proof above shows that:

**Corollary 7.6.10.** The Galois group of the cyclotomic field of order $n$ over $\mathbb{Q}$ is $\mathbb{Z}_n^*$. In particular, its order is given by the Euler function:
\[ |\text{Aut}_\mathbb{Q}(\mathbb{Q}(e^{\frac{2\pi i}{n}}))| = \varphi(n) = n \prod_{i=1}^{k} \left( 1 - \frac{1}{p_i} \right), \]
where $p_1, \ldots, p_n$ are the distinct prime factors of $n$. 
7.6. THE GALOIS GROUP

Proof. Let \( \zeta = e^{\frac{2\pi i}{n}} \), so \( U = \{ \zeta, \zeta^2, \ldots, \zeta^n \} \simeq \mathbb{Z}_n \) is set of \( n \)-th roots of the unit. In this case the map \( \text{Aut}_\mathbb{Q}(\zeta) \to \text{Aut}(\mathbb{Z}_n), \phi \mapsto \phi|_U \) is an isomorphism, so the result follows.

Once one has information about the cyclotomic fields, one can find the Galois group of an equation of the form \( x^n - a = 0 \). For example, one has:

**Proposition 7.6.11.** If \( K \) has characteristic 0 and contains the roots of \( x^n - 1 = 0 \), then the Galois group of \( x^n - a \) over \( K \) is cyclic of order a divisor of \( n \).

**Proof.** Let \( L \) be a splitting extension of \( x^n - a \) over \( K \), and denote again by \( U = \{ z \in L : z^n - 1 = 0 \} \) the set of roots of \( x^n - 1 \). If \( r \in L \) is a root of \( x^n - a \), then \( \{ zr : z \in U \} \) is the set of \( n \) roots of \( x^n - a \) in \( L \). Hence, we have \( L = K(r) \). If \( \phi_1, \phi_2 \in \text{Aut}_K(L) \equiv G \), then \( \phi_1(r) = z_1r, \phi_2(r) = z_2r \), for some \( z_1, z_2 \in U \), and \( \phi_1 \circ \phi_2(r) = z_1z_2r \). Hence, the map \( \phi \mapsto z \) is a monomorphism form \( G \) into the cyclic group \( U \simeq \mathbb{Z}_n \) and we conclude that \( G \) is isomorphic to a subgroup of \( \mathbb{Z}_n \).

**Exercises.**

1. Determine the Galois group of the extension \( \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}) \) over \( \mathbb{Q} \).

2. Let \( L = K(x) \) be the field of fractions of \( K[x] \). Show that the primitive elements of \( L \) take the form

\[
t = \frac{ax + b}{cx + d}, \quad a, b, c, d \in K, \quad ad - bc \neq 0
\]

(HINT: If \( t = p(x)/q(x) \), where \( \gcd(p(x), q(x)) = 1 \), define the degree of \( t \) to be the maximum of the degrees of \( p(x) \) and \( q(x) \). Show that \( p(w) - yq(w) \) is irreducible in \( K[w, y] \), hence also in \( K(y)[w] \), and that \( x \) is algebraic over \( K(t) \) with minimal polynomial a multiple of \( p(w) - tq(w) \). Conclude that \( [K(x) : K(t)] = 1 \) (or that \( K(x) = K(t) \)) if and only if the degree of \( t \) is 1.)

3. Determine the Galois group of \( x^3 - 2 \) over \( \mathbb{Q} \) and over \( \mathbb{Z}_2 \).

4. Find the Galois group of the polynomial \( x^n - 2 \in \mathbb{Q}[x] \) and represent it as a group of permutations of the roots, for \( n = 4, 5, 6, 7 \).

5. Determine the Galois group of the polynomial \( p(x) = 2x^3 + 3x^2 + 6x + 6 \in \mathbb{Q}[x] \).

6. Show that the Galois group of \( p(x) = x^3 + x^2 - 2x - 1 \in \mathbb{Q}[x] \) is isomorphic to the alternating group \( A_4 \).

(HINT: If \( r \) is a root of \( p(x) \), then \( r^2 - 2 \) is also a root.)
7.7 The Galois Correspondence

Finally we are now able to explain how Galois Theory allows one to replace a problem concerning (extensions of fields of) polynomials by a simpler problem in Group Theory. The correspondence, which goes back to Galois, allows one to associate to an intermediate extension a subgroup of the Galois group, as we now explain.

Let $L \supseteq K$ be an extension of $K$ and $H \subset \text{Aut}_K(L)$ a subgroup of the Galois group. We can think of $H$ as a group of transformations of $L$, in which case the fixed point set of $H$ is an intermediate field $L \supseteq \text{Fix}(H) \supseteq K$. Conversely, given an intermediate field $L \supseteq \tilde{K} \supseteq K$ then $\text{Aut}_{\tilde{K}}(L)$ is a subgroup of the Galois group $\text{Aut}_K(L)$.

The following properties are immediate for the definitions of these correspondences:

**Proposition 7.7.1** (Properties of the Galois Correspondence). Let $L$ be an extension of $K$ with Galois group $G = \text{Aut}_K(L)$. Let $\tilde{K}, \tilde{K}_1$ and $\tilde{K}_2$ intermediate extensions, and $H, H_1, H_2 \subset G$ subgroups.

(i) If $H_1 \supseteq H_2$, then $\text{Fix}(H_1) \subset \text{Fix}(H_2)$.

(ii) If $\tilde{K}_1 \supseteq \tilde{K}_2$, then $\text{Aut}_{\tilde{K}_1}(L) \subset \text{Aut}_{\tilde{K}_2}(L)$.

(iii) $\text{Fix}(\text{Aut}_{\tilde{K}}(L)) \supseteq \tilde{K}$.

(iv) $\text{Aut}_{\text{Fix}(H)}(L) \supseteq H$.

**Example 7.7.2.**

Consider the extension $L = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ of $K = \mathbb{Q}$. As we saw before, the Galois group of this extension contain 4 elements:

$$\text{Aut}_K(L) = \{\text{id}, \phi_1, \phi_2, \phi_3\} \equiv G.$$  

Besides the trivial subgroup $H_0 = \{\text{id}\}$, this group has the subgroups $H_1 = \{\text{id}, \phi_1\}$, $H_2 = \{\text{id}, \phi_2\}$ and $H_3 = \{\text{id}, \phi_3\}$. This lattice of subgroups can be represented by the diagram:

$$\begin{array}{ccc}
G & & \\
\downarrow & & \downarrow \\
H_1 & & H_2 \\
\downarrow & & \downarrow \\
H_0 & & H_3
\end{array}$$

---

8A lattice is a partially ordered set in which every subset of two elements has a supremum and a infimum. The set of subgroups of a group $G$, ordered by inclusion, is a lattice. The set of intermediate extensions $K \subset \tilde{K} \subset L$ of a fixed extension $L$ of $K$, ordered by inclusion, is a also a lattice.
The field fixed by the Galois group \( G = \text{Aut}_\mathbb{Q}(L) \) is the ground field: \( \text{Fix}(G) = \mathbb{Q} \), while obviously \( \text{Fix}(H_0) = \mathbb{Q}(\sqrt{2}, \sqrt{3}) \). On the other hand, one checks easily that
\[
\text{Fix}(H_1) = \mathbb{Q}(\sqrt{3}), \quad \text{Fix}(H_2) = \mathbb{Q}(\sqrt{2}), \quad \text{Fix}(H_3) = \mathbb{Q}(\sqrt{6}).
\]

Hence, the lattice of intermediate extensions is given by the diagram:

\[
\begin{array}{ccc}
\mathbb{Q} & \mathbb{Q}(\sqrt{3}) & \mathbb{Q}(\sqrt{2}) \\
& \mathbb{Q}(\sqrt{2}) & \mathbb{Q}(\sqrt{6}) \\
& & \mathbb{Q}(\sqrt{2}, \sqrt{3})
\end{array}
\]

As in the previous example, extensions with a rich group of automorphisms so that the correspondences above between intermediate extensions and subgroups are inverse to each other (so in properties (iii) and (iv), inclusions can be replaced by equalities) are specially important and deserve a special name:

**Definition 7.7.3.** A finite dimensional extension \( L \supset K \) is called a **Galois extension** of \( K \) if \( \text{Fix}(\text{Aut}_K(L)) = K \).

We have the following alternative characterizations of a Galois extension.

**Proposition 7.7.4.** The following statements are equivalent:

(i) \( L \) is a Galois extension of \( K \);

(ii) \( L \) is a splitting extension of a separable polynomial over \( K \);

(iii) \( L \) is a finite dimensional extension, normal and separable over \( K \).

If any of these equivalent conditions hold, then:
\[
|\text{Aut}_K(L)| = [L : K].
\]

**Proof.** Note that (7.7.1) follows from Theorem 7.6.5. So it remains to prove the implications (i) \( \Rightarrow \) (ii) \( \Rightarrow \) (iii) \( \Rightarrow \) (i).

(i) \( \Rightarrow \) (ii): Since \( [L : K] < \infty \), there exist \( r_1, \ldots, r_n \in L \), algebraic over \( K \), such that \( L = K(r_1, \ldots, r_n) \). Let \( p_i(x) \) be the minimal polynomial of each \( r_i \) in \( K[x] \), and \( O_i = \{ \phi(r_i) : \phi \in \text{Aut}_K(L) \} \) the orbit of \( r_i \) under the
action of the Galois group. This set is finite and consists of roots of \( p_i(x) \).

The polynomial

\[
q_i(x) = \prod_{r \in \mathcal{O}_i} (x - r) \in L[x]
\]

divides \( p_i(x) \) and is separable, since all its roots are distinct. If \( \phi \in \text{Aut}_K(L) \), then

\[
q_i^\phi(x) = \prod_{r \in \mathcal{O}_i} (x - \phi(r)) = \prod_{r \in \mathcal{O}_i} (x - r) = q_i(x),
\]

so the coefficients of \( q_i(x) \) belong to \( \text{Fix}(\text{Aut}_K(L)) = K \). We conclude that \( q_i(x) \in K[x] \), so we must have \( p_i(x) = q_i(x) \). It follows that \( p_i(x) \) is separable and \( L \) is a splitting extension of \( p(x) \) over \( K \).

(ii) \( \Rightarrow \) (iii): If \( L \) is a splitting extension of a polynomial \( p(x) \) over \( K \), then, by Corollary 7.3.10, \( L \) is a normal extension of finite dimension over \( K \). If \( p(x) \) is a separable polynomial, then \( L = K(r_1, \ldots, r_m) \), where the \( r_1, \ldots, r_m \) are separable. By Proposition 7.5.9, \( L \) is a separable extension.

(iii) \( \Rightarrow \) (i): Let \( \tilde{K} = \text{Fix}(\text{Aut}_K(L)) \). From the general properties of the Galois correspondence, \( \tilde{K} \) is an intermediate field \( L \supseteq \tilde{K} \supseteq K \). Let \( r_1 \in \tilde{K} - K \) and denote by \( p(x) \) the minimal polynomial of \( r_1 \) over \( K \). Since \( L \) is normal and separable, \( p(x) \) splits in \( L[x] \) into a product of distinct linear factors: \( p(x) = \prod_{i=1}^m (x - r_i) \), with \( m > 1 \) and the \( r_i \) all distinct. If \( \phi : K(r_1) \to \bar{K} \) is the \( K \)-monomorphism mapping \( r_1 \) to \( r_2 \), by Theorem 7.4.6 we can extend \( \phi \) to a monomorphism \( \Phi : L \to \bar{K} \). By Theorem 7.3.9 \( \Phi \) is actually a \( K \)-automorphism of \( L \). It follows that \( \Phi \) is an element of \( \text{Aut}_K(L) \) which does not fix \( r_1 \). This contradicts \( r_1 \in \tilde{K} = \text{Fix}(\text{Aut}_K(L)) \). Hence, we must have \( K = \tilde{K} = \text{Fix}(\text{Aut}_K(L)) \). \( \square \)

If \( L \) fails to be a Galois extension of \( K \), then \( |\text{Aut}_K(L)| \leq [L : K] \). On the other hand, we have the following general result:

**Lemma 7.7.5** (Artin). Let \( G \) be a finite group of automorphisms of a field \( L \) and \( K = \text{Fix}(G) \). Then

\[
[L : K] \leq |G|.
\]

**Proof.** Let \( G = \{\phi_1 = \text{id}, \phi_2, \ldots, \phi_m\} \). We need to show that any \( n \) elements of \( L \) with \( n > m \) are linearly dependent over \( K \).

---

9Emil Artin (1898-1962) was among the greatest algebraists of the 20th Century. The modern approach to Galois Theory was formulated by him, together with Irving Kaplanski (1917-2006). Artin and Kaplanski were both members of the Bourbaki group.
Let \( u_1, \ldots, u_n \in L \) and consider the homogeneous linear system:

\[
\sum_{i=1}^{n} a_i \phi_j(u_i) = 0, \quad (j = 1, \ldots, m).
\]

Since there are more variables than equations, elementary Linear Algebra shows that it must have a non-trivial solution \( (a_1, \ldots, a_n) \in L^n \). After possibly some reordering, we can choose the solution to be of the form \( (a_1, \ldots, a_s, 0, \ldots, 0) \), where \( a_i \neq 0 \) for \( i = 1, \ldots, s \) and \( s \geq 2 \) is minimal (i.e., there is no non-trivial solution with a smaller \( s \)). Dividing by \( a_1 \), we can additionally assume that \( a_1 = 1 \). We claim that \( a_i \in K = \text{Fix}(G) \), for all \( i \).

In fact, if some \( a_i \not\in K = \text{Fix}(G) \), there exists \( \phi \in G \) such that \( \phi(a_i) \neq a_i \). Then the vector \( (1, \phi(a_2), \ldots, \phi(a_s), 0, \ldots, 0) \) is also a solution of the system and the difference vector \( (0, a_2 - \phi(a_2), \ldots, a_s - \phi(a_s), 0, \ldots, 0) \) is a non-trivial solution with more zeros than \( (1, a_2, \ldots, a_s, 0, \ldots, 0) \), contradicting the assumption that \( s \) was minimal.

We conclude that \( a_i \in K \), for all \( i \). When \( j = 1 \) the system gives the linear dependence relation

\[
\sum_{i=1}^{n} a_i u_i = 0.
\]

Finally we are able to state and prove the key result of Galois Theory.

**Theorem 7.7.6** (Fundamental Theorem of Galois Theory). Let \( L \) be a Galois extension of \( K \). There exists a bijective correspondence between the intermediate extensions \( K \subset \hat{K} \subset L \) and the subgroups \( H \) of the Galois group \( G \equiv \text{Aut}_K(L) \), given by:

\[
\hat{K} \mapsto \text{Aut}_K(L) \equiv H, \quad H \mapsto \text{Fix}(H) \equiv \hat{K}.
\]

If one writes \( H \leftrightarrow \hat{K} \), this correspondence satisfies:

(i) If \( H_1 \leftrightarrow \hat{K}_1 \) and \( H_2 \leftrightarrow \hat{K}_2 \), then \( H_2 \subset H_1 \) if and only if \( \hat{K}_2 \supset \hat{K}_1 \). In this case we have \( [H_1 : H_2] = [\hat{K}_2 : \hat{K}_1] \).

(ii) If \( H \leftrightarrow \hat{K} \), then \( H \) is a normal subgroup of \( G \) if and only if \( \hat{K} \) is a normal extension of \( K \). In this case, \( \hat{K} \) is a Galois extension with Galois group \( G/H \).
CHAPTER 7. GALOIS THEORY

Proof. First we check that the two assignments are inverse to each other:

Fix(Aut_{\hat{K}}(L)) = \hat{K}: Let L \supset \hat{K} \supset K be an extension. By Proposition 7.7.4 L is a finite-dimensional extension, normal and separable over K, hence also over \hat{K}. But then L is a Galois extension of \hat{K}, and we conclude that \hat{K} = Fix(H) with H = Aut_{\hat{K}}(L).

Aut_{\hat{K}}(L) = H: Let H \subset G \equiv Aut_K(L) and \hat{K} = Fix(H). By Artin’s Lemma, \vert H \vert \geq \vert [L : Fix(H)] = [L : \hat{K}] \vert. On the other hand, L is a Galois extension over \hat{K} and H \subset Aut_{\hat{K}}(L). Then by relation (7.7.1) we see that \vert H \vert \leq \vert Aut_{\hat{K}}(L) \vert = [L : \hat{K}]. Hence, \vert H \vert = \vert Aut_{\hat{K}}(L) \vert and we conclude that H = Aut_{\hat{K}}(L), with \hat{K} = Fix(H).

In order to check that (i) holds, notice that from Proposition 7.7.1 and the fact that the correspondence is bijective, we have \hat{K}_2 \supset \hat{K}_1 if and only if \hat{K}_2 \supset \hat{K}_1. Since L is a Galois extension over both \hat{K}_1 and \hat{K}_2, we find:

\vert \hat{K}_2 : \hat{K}_1 \vert = \frac{[L : \hat{K}_1]}{[L : \hat{K}_2]} = \frac{\vert Aut_{\hat{K}_1}(L) \vert}{\vert Aut_{\hat{K}_2}(L) \vert} = \vert \hat{K}_1 : \hat{K}_2 \vert = \vert H_1 : H_2 \vert.

In order to check that (ii) holds, assume that H ↔ \hat{K}. If \phi \in G, then the extension corresponding to \phi H \phi^{-1} is \phi(\hat{K}). Hence, we have:

(a) If H is a normal subgroup of G, then for all \phi \in G we have \phi(\hat{K}) \subset \hat{K}.

The map \phi \mapsto \phi|_{\hat{K}} is a surjective homomorphism of G onto Aut_K(\hat{K}) with kernel H. Hence, Aut_K(\hat{K}) \simeq G/H and

Fix(Aut_K(\hat{K})) = Fix(G/H) = Fix(G) = K.

We conclude that \hat{K} is a Galois extension with Galois group G/H.

(b) Conversely, assume that \hat{K} is a normal extension. By Theorem 7.3.9 \phi(\hat{K}) = \hat{K}, for every \phi in the Galois group G. Since \phi H \phi^{-1} ↔ \phi(\hat{K}), we conclude that \phi H \phi^{-1} = H. Hence, H is a normal subgroup of G.

Example 7.7.7.

We saw before that Q(i, \sqrt[4]{2}) is a Galois extension over Q, with Galois group

Aut_Q Q(i, \sqrt[4]{2}) = \{1, \tau, \tau^2, \tau^3, \sigma, \sigma \tau, \sigma \tau^2, \sigma \tau^3 \} \equiv G,

where \sigma and \tau are the Q-automorphisms of Q(i, \sqrt[4]{2}) defined by

\sigma(i) = -i, \quad \tau(i) = i,
\sigma(\sqrt[4]{2}) = \sqrt[4]{2}, \quad \tau(\sqrt[4]{2}) = i \sqrt[4]{2}.
This group has the following lattice of subgroups:

\[
\begin{array}{c}
G \\
\{1, \tau^2, \sigma \tau^3\} \\
\langle \sigma \tau \rangle \\
\langle \sigma \rangle \\
\langle 1 \rangle \\
\{1, \sigma, \tau \} \\
\langle \sigma \rangle \\
\langle \sigma \tau^2 \rangle \\
\langle \tau \rangle \\
\{1, \sigma, \tau \} \\
\end{array}
\]

The Galois correspondence yields a similar lattice of intermediate extensions of \( \mathbb{Q} \):

\[
\begin{array}{c}
\mathbb{Q} \\
\mathbb{Q}(i) \\
\mathbb{Q}(r^2) \\
\mathbb{Q}(i) \\
\mathbb{Q}(r^2) \\
\mathbb{Q}(i, r) \\
\mathbb{Q}(i, r^2) \\
\mathbb{Q}(i, r) \\
\mathbb{Q}(i, \sqrt{2}) \\
\mathbb{Q}(i, \sqrt{3}) \\
\mathbb{Q}(i, \sqrt{5}) \\
\end{array}
\]

where \( r = \sqrt{2} \). The extensions in the first row are normal extensions, being extensions of degree 2. They correspond to subgroups of index 2, therefore normal subgroups. In the second row, only the extension \( \mathbb{Q}(i, \sqrt{2}) \) is normal: it corresponds to the center \( C(G) = \{1, \tau^2\} \).

Exercises.

1. Determine the Galois correspondence for the extension \( \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}) \subset \mathbb{R} \) over \( \mathbb{Q} \).

2. Determine the Galois correspondence for the splitting field of the polynomial \( x^3 - 2 \) over \( \mathbb{Q} \) and over \( \mathbb{Z}_2 \).

3. Determine the Galois correspondence for the splitting field of the polynomial \( (x^3 - 2)(x^2 - 3) \in \mathbb{Q}[x] \).

4. Determine the Galois correspondence for the splitting field of the polynomial \( x^4 - 4x^2 - 1 \in \mathbb{Q}[x] \).
5. Let $p(x) \in \mathbb{Q}[x]$ be a polynomial of degree 3 with Galois group $G$. If $r_1, r_2, r_3 \in \mathbb{C}$ denote the roots of $p(x)$ and $\delta \equiv (r_1 - r_2)(r_1 - r_3)(r_2 - r_3)$, show that:

(a) $|G| = 1$ if and only if the roots of $p(x)$ belong to $\mathbb{Q}$;
(b) $|G| = 2$ if and only if $p(x)$ has exactly one rational root;
(c) $|G| = 3$ if and only if $p(x)$ has no rational roots and $\delta \in \mathbb{Q}$;
(d) $|G| = 6$ if and only if $p(x)$ has no rational roots and $\delta \notin \mathbb{Q}$.

6. If $p(x) \in K[x]$ is a polynomial of degree $n$ with roots $r_1, \ldots, r_n$, define the discriminant of $p(x)$ to be $\Delta = \delta^2$, where

$$\delta = \prod_{i<j}(r_i - r_j).$$

Assuming that $p(x)$ is separable and $K$ has characteristic different from 2, show that:

(a) $\Delta \in K$;
(b) $\Delta = 0$ if and only if $p(x)$ has a multiple root;
(c) $\Delta$ is a perfect square in $K$ if and only if the Galois group of $p(x)$ is contained in $A_n$.

7. If $p(x) = x^3 + px + q \in \mathbb{Q}[x]$, show that the discriminant (see the previous exercise) is $\Delta = -4p^3 - 27q^2$. What is the Galois group of $x^3 + 6x^2 - 9x + 3$?

8. Let $L$ be a Galois extension of $K$, and $L \supseteq \hat{K} \supseteq K$ an intermediate extension. Let $H \subset \text{Aut}_K(L)$ be the subgroup formed by the $K$-automorphisms which map $\hat{K}$ into itself. Show that $H$ is the normalizer of $\text{Aut}_{\hat{K}}(L)$ in $\text{Aut}_K(L)$.

7.8 Applications

We will now see several applications of Galois Theory. Our first application is to characterize all symmetric polynomials. Our second application is to find a criterion to decide if a complex number is constructible or not, completing our discussion in Section 7.2. Our third and final application is to the original question that motivated Galois: we will prove Galois criterion to decide if an algebraic equation is solvable by radicals.
7.8. APPLICATIONS

7.8.1 Symmetric Polynomials

Let \( x_1, \ldots, x_n \) be indeterminates. In the field of fractions \( K(x_1, \ldots, x_n) \) consider the polynomial

\[
p(x) = \prod_{i=1}^{n} (x - x_i) = x^n - s_1 x^{n-1} + s_2 x^{n-2} - \cdots + (-1)^n s_n.
\]

The coefficients \( s_i \) of this polynomial are certain polynomials in the indeterminates \( x_i \), called **elementary symmetric polynomials**. They can be explicit written as:

\[
\begin{align*}
s_1 &= \sum_i x_i, \\
s_2 &= \sum_{i<j} x_i x_j, \\
&\vdots \\
s_n &= x_1 \cdots x_n.
\end{align*}
\]

We write their general formula in the form:

\[
s_i = \sum_{j_1 < \cdots < j_i} x_{j_1} \cdots x_{j_i}, \quad i = 1, \ldots, n.
\]

The use of the term “symmetric” is justified by the fact that the polynomials remain the same under any permutation of the indices \( j_1, \ldots, j_n \). More generally, we can consider **symmetric rational expressions** in the variables \( x_i \), which can be formalized as follows. Every permutation \( \pi \in S_n \) determines a \( K \)-automorphism \( \phi_\pi \) of \( K(x_1, \ldots, x_n) \) which transforms \( x_i \) in \( x_{\pi(i)} \). A **symmetric rational expression** is any element of \( K(x_1, \ldots, x_n) \) which is fixed by the group \( G \equiv \{ \phi_\pi : \pi \in S_n \} \simeq S_n \). Notice that \( G \subset \text{Aut}_K(K(x_1, \ldots, x_n)) \) and, under the Galois correspondence, the symmetric rational expressions are precisely the elements of the intermediate extension \( K \subset \text{Fix}(G) \subset K(x_1, \ldots, x_n) \).

**Examples 7.8.1.**

1. The polynomial \( p(x) = \prod_{i=1}^{n} (x - x_i) \) is clearly invariant under the action of the \( \phi_\pi \) (i.e., \( p(x) = p^{\phi_\pi}(x) \)), hence its coefficients \( s_i \) are fixed by \( \phi_\pi \) and we have \( s_i \in \text{Fix}(G) \). In other words, the \( s_i \) are symmetric rational expressions.
2. The fractions
\[ \frac{x_1 x_2}{x_3} + \frac{x_2 x_3}{x_1} + \frac{x_3 x_1}{x_2}, \quad \frac{1}{x_1^2} + \frac{1}{x_2^2} + \frac{1}{x_3^2}, \]
are symmetric rational expressions.

We can use the Galois correspondence to show that any symmetric rational expression can be expressed in terms of the elementary symmetric polynomials \( s_1, \ldots, s_n \). More precisely, we have:

**Theorem 7.8.2.** \( K(x_1, \ldots, x_n) \) is a Galois extension of \( K(s_1, \ldots, s_n) \) with Galois group \( G = \{ \phi_n : \pi \in S_n \} \simeq S_n \). In particular, \( \text{Fix}(G) = K(s_1, \ldots, s_n) \).

**Proof.** As we observed above, we have \( K(s_1, \ldots, s_n) \subset \text{Fix}(G) \) hence,
\[ G \subset \text{Aut}_{K(s_1, \ldots, s_n)} K(x_1, \ldots, x_n). \]
Note that \( K(x_1, \ldots, x_n) = K(s_1, \ldots, s_n, x_1, \ldots, x_n) \) is a splitting extension of \( p(x) = \prod_{i=1}^{n}(x - x_i) \in K(s_1, \ldots, s_n)[x] \). On the other hand, if \( \phi \in \text{Aut}_{K(s_1, \ldots, s_n)} K(x_1, \ldots, x_n) \), then \( \phi \) permutes the roots of \( p(x) \), hence
\[ \text{Aut}_{K(s_1, \ldots, s_n)} K(x_1, \ldots, x_n) \subset G. \]
We conclude that the Galois group of \( K(x_1, \ldots, x_n) \) over \( K(s_1, \ldots, s_n) \) is \( G \simeq S_n \). Since \( K(x_1, \ldots, x_n) \) is a normal and separable extension, it is a Galois extension and \( \text{Fix}(G) = K(s_1, \ldots, s_n) \).

**Example 7.8.3.**
The polynomial \( x_1^3 + x_2^3 + x_3^3 \in \mathbb{Q}(x_1, x_2, x_3) \) is a rational symmetric expression, hence it can be expressed as a rational fraction in \( s_1 = x_1 + x_2 + x_3, s_2 = x_1 x_2 + x_1 x_3 + x_2 x_3 \) and \( s_3 = x_1 x_2 x_3 \). Elementary degree considerations show that there exist rational numbers \( a_1, a_2 \) and \( a_3 \), such that
\[ x_1^3 + x_2^3 + x_3^3 = a_1 (x_1 + x_2 + x_3)^3 + a_2 (x_1 x_2 + x_1 x_3 + x_2 x_3) + a_3 x_1 x_2 x_3. \]

By choosing convenient values of \( x_1, x_2 \) and \( x_3 \), we can find the coefficients:
- \( 3 = 27a_1 + 9a_2 + a_3 \) (if we set \( x_1 = x_2 = x_3 = 1 \)),
- \( 2 = 8a_1 + 2a_2 \) (if we set \( x_1 = x_2 = 1, x_3 = 0 \)),
- \( 1 = a_1 \) (if we set \( x_1 = 1, x_2 = x_3 = 0 \)).

This linear system has the solution \( a_1 = 1, a_2 = -3, a_3 = 3 \). Hence, we conclude that:
\[ x_1^3 + x_2^3 + x_3^3 = (x_1 + x_2 + x_3)^3 - 3(x_1 x_2 + x_1 x_3 + x_2 x_3) + 3 x_1 x_2 x_3. \]
7.8. APPLICATIONS

7.8.2 Constructible Numbers

We have studied in Section 7.2 the subfield $\mathbb{C} \subset \mathbb{C}$ formed by the numbers that can be constructed using compass and straightedge. Using Galois Theory we obtain the following characterization of the constructible numbers:

**Theorem 7.8.4.** A complex number $z \in \mathbb{C}$ is constructible if and only if $z$ is algebraic over $\mathbb{Q}$, and the normal closure $\overline{\mathbb{Q}(z)}$ has dimension $2^s$, for some $s \in \mathbb{N}$.

**Proof.** Recall that a complex number $z \in \mathbb{C}$ is constructible if and only if $z$ belongs to a subfield $\mathbb{Q}(u_1, \ldots, u_r)$, with $u_1^2 \in \mathbb{Q}$ and $u_{m+1}^2 \in \mathbb{Q}(u_1, \ldots, u_m)$, for $m = 1, \ldots, r - 1$. It follows that if $z$ is a constructible, then $\overline{\mathbb{Q}(z)} \subset \overline{\mathbb{Q}(u_1, \ldots, u_r)}$.

Letting $G = \text{Aut}_\mathbb{Q}(\mathbb{Q}(u_1, \ldots, u_r))$, Theorem 7.3.9 shows that the field $\mathbb{Q}(u_1, \ldots, u_r)$ is generated by the images $\phi(\mathbb{Q}(u_1, \ldots, u_r))$, with $\phi \in G$. Hence, if $G = \{\phi_1, \ldots, \phi_n\}$ we find:

$\overline{\mathbb{Q}(u_1, \ldots, u_r)} = \overline{\mathbb{Q}(\phi_1(u_1), \ldots, \phi_r(u_r))}$.

Since $\phi_j(u_r)^2 = \phi_j(u_r^2)$, we conclude that $\overline{\mathbb{Q}(u_1, \ldots, u_r)}$ is an extension of the form $\mathbb{Q}(\tilde{u}_1, \ldots, \tilde{u}_l)$ with $\tilde{u}_1^2 \in \mathbb{Q}$ and $\tilde{u}_{m+1}^2 \in \mathbb{Q}(\tilde{u}_1, \ldots, \tilde{u}_m)$, for $m = 1, \ldots, l - 1$. We can now compute the algebraic degrees as follows:

$[\overline{\mathbb{Q}(u_1, \ldots, u_r)} : \mathbb{Q}] = \prod_{m=0}^{l-1} [\mathbb{Q}(\tilde{u}_1, \ldots, \tilde{u}_{m+1}) : \mathbb{Q}(\tilde{u}_1, \ldots, \tilde{u}_m)] = 2^t$.

It follows that

$[\overline{\mathbb{Q}(z)} : \mathbb{Q}] = \frac{[\overline{\mathbb{Q}(u_1, \ldots, u_r)} : \mathbb{Q}]}{[\overline{\mathbb{Q}(u_1, \ldots, u_r)} : \overline{\mathbb{Q}(z)}]} = 2^s$,

for some $s \in \mathbb{N}$.

Conversely, assume that $[\overline{\mathbb{Q}(z)} : \mathbb{Q}] = 2^s$. Then $\overline{\mathbb{Q}(z)}$ is a Galois extension of $\mathbb{Q}$ whose Galois group has order $|G| = 2^s$. By the results of Chapter 5 on $p$-groups, we know that there exists a normal tower $G = H_s \triangleright H_{s-1} \triangleright \cdots \triangleright H_1 \triangleright H_0 = \{e\}$, where $[H_m : H_{m-1}] = 2$. Galois correspondence, then gives intermediate extensions $\overline{\mathbb{Q}(z)} = K_s \supset K_{s-1} \supset \cdots \supset K_1 \supset \mathbb{Q}$.
where \([K_m : K_{m-1}] = 2\). Hence, for each \(1 \leq m \leq s\), there exist \(u_m \in K_m\) such that \(K_{m+1} = K_m(u_{m+1})\) and \(u_{m+1}^2 \in K_m\). We conclude that \(\mathbb{Q}(z)^m = \mathbb{Q}(u_1, \ldots, u_s)\), with \(u_{m+1}^2 \in \mathbb{Q}(u_1, \ldots, u_m)\) for \(0 \leq m \leq s - 1\). Hence, \(z\) is constructible.

Galois Theory gives more than a criterion to decide if a number is constructible: for constructible numbers it gives a method to construct the number! The is illustrated in the next example.

**Example 7.8.5.**

A regular polygon with 5 sides can be constructed with rule and compass: if the polygon is inscribed in the unit circle, then this is equivalent to the statement that a primitive root of \(x^5 - 1 = 0\) is constructible. In fact, if \(r = e^{2\pi i/5}\), then \(\mathbb{Q}(r) = \mathbb{Q}(e^{2\pi i/5})\), and this extension has algebraic degree \(4 = 2^2\).

In order to give an explicit description of a regular pentagon, we start by observing that \((x^5 - 1) = (x - 1)(x^4 + x^3 + x^2 + x + 1)\), hence \(r\) is a root of an irreducible polynomial of \(4\)th degree. The extension \(L = \mathbb{Q}(r)\) is a splitting extension of this polynomial, hence is a Galois extension. The Galois group has order \(|G| = 2^2\). The \(\mathbb{Q}\)-automorphism defined by \(\phi(r) = r^2\) is an element of the Galois group. If we label the roots as \(z_k = e^{2\pi ik/5} (k = 1, \ldots, 4)\), then

\[
\phi(z_1) = z_2, \quad \phi(z_2) = z_4, \quad \phi(z_4) = z_3, \quad \phi(z_3) = z_1,
\]

so we see that \(\phi\) corresponds to the permutation \((1243)\). This element has order 4, so the Galois group is \(G = \{I, \phi, \phi^2, \phi^3\}\). In terms of permutations of the roots:

\[
G = \{I, (1243), (14)(23), (3241)\}.
\]

This group admits the tower of \(p\)-subgroups:

\[
G \supset H \supset \{I\},
\]

where \(H = \{I, (14)(23)\}\). Since \(|G : H| = 2\), to \(H\) corresponds an extension \(\tilde{K}\) of degree 2 over \(\mathbb{Q}\). To determine this extension, we observe that an element \(u \in L\) is of the form \(u = a_1z_1 + a_2z_2 + a_3z_3 + a_4z_4\), where

\[
\phi^2(u) = a_1z_4 + a_2z_3 + a_3z_2 + a_4z_1,
\]

and \(u\) is fixed by \((14)(23)\) if and only if \(a_1 = a_4\) and \(a_2 = a_3\). Hence, \(\tilde{K} = \mathbb{Q}(\omega_1, \omega_2)\), where \(\omega_1 = z_1 + z_4\) and \(\omega_2 = z_2 + z_3\). It is easy to check that

\[
\omega_1 + \omega_2 = z_1 + z_4 + z_2 + z_3 = r + r^2 + r^3 + r^4 = -1
\]

\[
\omega_1\omega_2 = (r + r^4)(r^2 + r^3) = r + r^2 + r^3 + r^4 = -1,
\]

so we have:

\[
(x - \omega_1)(x - \omega_2) = x^2 + x - 1.
\]
Solving this equation, we see that

\[ \omega_1 = \frac{-1 + \sqrt{5}}{2}, \quad \omega_2 = \frac{-1 - \sqrt{5}}{2} \]

(these values are \(2 \text{Re}(z_1) = 2 \text{Re}(z_4)\) and \(2 \text{Re}(z_2) = 2 \text{Re}(z_3)\)).

We can now explain the traditional way of constructing a pentagon. In a unitary circle, we mark the four points \(A = (1,0), B = (0,1), C = (-1,0)\) and \(D = (0,-1)\). Splitting the segment \(OD\) into two equal parts, we obtain the point \(E\). The segment \(EA\) has length \(\sqrt{5}/2\). Centering the compass in \(E\) we obtain the arc \(AF\). The segment \(OF\) has length \(\omega_1 = \frac{-1 + \sqrt{5}}{2}\), and we can mark the point \(G\) in the horizontal axis, so that \(OG = \frac{AF}{2}\). The point \(G\) corresponds to the \(x\)-coordinate of \(z_1\) (it is slightly simple to observe that \(AF\) has the same length as the side of a pentagon).

![Construction of a regular pentagon](image)

Figure 7.8.1: Construction of a regular pentagon.

The Greeks knew how to construct regular polygons with 3, 5 and 15 sides and, given a regular polygon with \(n\) sides, how to obtain one with \(2n\) sides (obviously by bisecting the sides). Gauss, when he was only 19 years old, and before Galois Theory was discovered, found a way to construct a regular polygon with 17 sides! This discovery made him choose Mathematics over his study of Languages. In fact, he was so proud of this discovery, that later in his life he asked that a regular polygon with 17 sides be engraved in his tombstone. This never happened because the sculptor thought that a polygon with so many sides could be confused with a circle.

Galois Theory shows that a regular polygon with \(n\) sides is constructible if and only if \(n = 2^r p_1 \cdots p_s\) where the \(p_i\) are Fermat primes. Using Galois Theory, it is known how to construct regular polygons with 257 and 65 537 sides!
7.8.3 Solving algebraic equations by radicals

We will now comeback to the opening discussion in this chapter, concerning
the criterion found by Galois which allows one to decide if an algebraic
equation is solvable by radicals. In order to simplify the discussion, we will
assumed the fields have characteristic 0.

Definition 7.8.6. Let \( p(x) \in K[x] \) be a monic polynomial. We say that
an equation \( p(x) = 0 \) is solvable by radicals if there exists an extension
\( L \supset K \) which contains a splitting field of \( p(x) \) and which is of the form:

\[
L = K_{m+1} \supset \cdots \supset K_2 \supset K_1 = K,
\]

where \( K_{i+1} = K_i(d_i) \) and \( d_i^{n_i} \in K_i \).

Notice the meaning of this definition: any root of \( p(x) \) belongs to \( L \)
and can be expressed starting from elements of \( K \) by a sequence of rational
operations and by finding n-th roots.

Theorem 7.8.7 (Galois Criterion). Let \( p(x) \in K[x] \) be a monic polynomial.
The equation \( p(x) = 0 \) is solvable by radicals if and only if its Galois group
is solvable.

Example 7.8.8.

The equation \( x^5 - 4x + 2 = 0 \) is not solvable by radicals (in \( \mathbb{Q} \)). This can be
seen as follows. By Eisenstein’s Criterion, the polynomial \( p(x) = x^5 - 4x + 2 \)
is irreducible over \( \mathbb{Q} \). It is easy to check that this polynomial has three real
distinct roots \( r_1, r_2, \) and \( r_3, \) and two complex conjugate roots \( r_4 \) and \( r_5. \) Let
\( L = \mathbb{Q}(r_1, \ldots, r_5) \) be the splitting field of \( p(x) \) and denote by \( G \) its Galois group.
Since \( [\mathbb{Q}(r) : \mathbb{Q}] = 5 \), for any \( r \in \{r_1, \ldots, r_5\} \), we have that \( 5 \mid [L : \mathbb{Q}] = |G| \).
The Sylow Theorems, show that the Galois group \( G \subset S_5 \) contains an element
of order 5, i.e., a cycle \( (i_1, \ldots, i_5) \). On the other, complex conjugation \( a + \text{i}b \to a - \text{i}b \) restricted to \( L = \mathbb{Q}(r_1, \ldots, r_5) \) gives an element of \( G \) of order 2, i.e.,
a transposition. We leave it as an exercise to check that these two elements
generate \( S_5 \). Hence, \( G = S_5 \), by Galois Criterion the equation is not solvable
by radicals.

Before we prove Galois criterion, we deduce the following corollary:

Corollary 7.8.9 (Theorem of Abel-Rufini). There is no general algebraic
solution by radicals for polynomial equations of degree five or higher.
7.8. APPLICATIONS

Proof. A more precise way of formulating this result is that the **general equation**

\[ x^n - a_{n-1}x^{n-1} + \cdots + (-1)^n a_0 = 0, \]

*is not solvable by radicals, for \( n \geq 5 \). By “general equation” we mean that \( a_0, \ldots, a_{n-1} \) are indeterminates. Hence, we consider the polynomial \( p(x) = x^n - a_{n-1}x^{n-1} + \cdots + (-1)^n a_0 \) over the field \( K(a_0, \ldots, a_{n-1}) \) and we need to show that the Galois group of \( p(x) \) over this field is not solvable.

Let \( L \) be the splitting field of \( p(x) \) over \( K(a_0, \ldots, a_{n-1}) \), so that in \( L \) we have the factorization:

\[ p(x) = (x - r_1)(x - r_2) \cdots (x - r_n). \]

By comparison, we see that

\[
\begin{align*}
a_{n-1} &= \sum_i r_i, \\
a_{n-2} &= \sum_{i<j} r_ir_j, \\
&\vdots \\
a_0 &= r_1r_2 \cdots r_n.
\end{align*}
\]

Hence, \( L = K(a_0, \ldots, a_{n-1})(r_1, \ldots, r_n) = K(r_1, \ldots, r_n) \).

Consider a set of indeterminates \( x_1, \ldots, x_n \), and in the field \( K(x_1, \ldots, x_n) \) consider the subfield of rational symmetric expressions (see Section 7.8.1): this subfield is of the form \( K(s_1, \ldots, s_n) \), where \( s_1, \ldots, s_n \) are the elementary symmetric polynomials in the \( x_i \)'s and \( K(x_1, \ldots, x_n) \) is a splitting extension of \( q(x) = \prod_i (x - x_i) \) over \( K(s_1, \ldots, s_n) \). We also saw that the Galois group of this extension is \( S_n \).

If there is an isomorphism \( K(r_1, \ldots, r_n) \cong K(x_1, \ldots, x_n) \) which to the subfield \( K(a_0, \ldots, a_{n-1}) \) associates \( K(s_1, \ldots, s_n) \), then the Galois group of the general equation of degree \( n \) is \( S_n \), which is not solvable for \( n \geq 5 \), so we are done. Let us establish the existence of this isomorphism.

Consider the \( K \)-homomorphism \( \phi : K[a_0, \ldots, a_{n-1}] \to K[s_1, \ldots, s_n] \), associating \( a_i \mapsto s_{n-i} \): this homomorphism exists, because the \( a_0, \ldots, a_n \) are indeterminates. Similarly, we have a homomorphism \( \psi : K[x_1, \ldots, x_n] \to K[r_1, \ldots, r_n] \), and the following diagram commutes:

\[
\begin{array}{ccc}
K[a_0, \ldots, a_n] & \xrightarrow{\phi} & K[s_1, \ldots, s_n] \\
\downarrow & & \downarrow \\
K[r_1, \ldots, r_n] & \xleftarrow{\psi} & K[x_1, \ldots, x_n]
\end{array}
\]
In fact, we have:

$$
\psi(\phi(a_i)) = \psi(s_i) = \psi \left( \sum_{j_1 < \cdots < j_i} x_{j_1} \cdots x_{j_i} \right) = \sum_{j_1 < \cdots < j_i} r_{j_1} \cdots r_{j_i} = a_i.
$$

The diagram shows that $\phi$ must be a monomorphism. Since $\phi$ is clearly surjective, it follows that $\phi$ is an isomorphism. Extending this isomorphism to the corresponding fields of fractions, we obtain an isomorphism of fields

$$\tilde{\phi} : K(a_0, \ldots, a_{n-1}) \to K(s_1, \ldots, s_n).$$

This isomorphism associates to a polynomial $p(x) \in K(a_0, \ldots, a_{n-1})[x]$ the polynomial $p^{\tilde{\phi}}(x) = q(x) \in K(s_1, \ldots, s_n)[x]$. As we saw in Section 7.4, $\tilde{\phi}$ extends to an isomorphism of the corresponding splitting fields $K(r_1, \ldots, r_n) \simeq K(x_1, \ldots, x_n)$, as claimed.

The Abel-Ruffini Theorem shows that there exists no general formula to solve a polynomial equation of degree $n$, when $n \geq 5$. Still there are equations which can be solved by radicals, e.g., $x^5 - 2 = 0$. In fact, it could happen that a general formula did not exist, but still every equation could be solved by radicals. However, the example of $x^5 - 4x + 2 = 0$ shows that is not the case.

It remains to prove Galois Criterion. In this proof, the splitting fields of the equations $x^n - a = 0$ play an essential role. We saw before that the Galois group of $x^n - a = 0$ is cyclic if $K$ contains all the roots of unit of order $n$ and is abelian when $a = 1$. In general, an extension $L$ of $K$ whose Galois group is abelian (respectively, cyclic) is called an abelian extension (respectively, cyclic extension) of $K$.

We need the following preliminary result:

**Proposition 7.8.10.** Let $K$ be a field which contains the $p$ roots of $x^p - 1 = 0$ (p a prime). If $L$ is a cyclic extension cyclic of $K$ with $[L : K] = p$, then $L = K(r)$, where $r^p \in K$.

**Proof.** If $u \in L - K$, then $L = K(u)$, since $L \supseteq K(u) \supseteq K$ and $[K(u) : K] = p$. If $\text{Aut}_K(L) = \langle \phi \rangle$ and $\{z_1, \ldots, z_p\} \subset K$ are the $p$-roots of $x^p - 1 = 0$, consider the elements

$$r_i = u + \phi(u)z_i + \phi^2(u)z_i^2 + \cdots + \phi^{p-1}(u)z_i^{p-1}. \tag{7.8.1}$$

Since $\phi(r_i) = z_i^{-1}r_i$, we have $\phi(r_i^p) = r_i^p$, and we conclude that $r_i^p \in K$. We can always write $u$ as a linear combination of the $r_i$’s by solving the
linear system (7.8.1) for the unknowns $u, \phi(u), \ldots, \phi^{p-1}(u)$ (this is possible because the corresponding determinant is a Van der Monde determinant). Hence, $L = K(r_1, \ldots, r_p)$, and for some $k_0$ we must have $r_{k_0} \notin K$. Setting $r = r_{k_0}$, we have $L = K(r)$, with $r^p \in K$.

Proof of the Galois Criterion. Let $p(x) \in K$ denote by $G = \text{Aut}_K(L)$ its Galois group. We will show both implications:

(i) If $p(x) = 0$ is solvable by radicals, then $G$ is solvable: If $p(x) = 0$ is solvable by radicals there exists an extension $L$ of $K$, which contains a splitting extension of $p(x)$, and which admits a tower of subfields

$$L = K_{i+1} \supset \cdots \supset K_2 \supset K_1 = K,$$

(7.8.2)

where $K_{i+1} = K_i(d_i)$, with $d_i^{n_i} = a_i \in K_i$. The normal closure $\tilde{L}^n$ of $L$ is generated by the $\phi(L)$, with $\phi \in \text{Aut}_K(\tilde{L}^n)$. Hence, if $\text{Aut}_K(\tilde{L}^n) = \{\text{id}, \phi_1, \ldots, \phi_r\}$, we obtain

$$\tilde{L}^n = K(d_1, \ldots, d_i, \phi_1(d_1), \ldots, \phi_1(d_i), \ldots, \phi_r(d_1), \ldots, \phi_r(d_i)).$$

Let $m = \text{lcm}(n_1, \ldots, n_i)$. We can extend the tower (7.8.3) to $\tilde{L}^m(z)$, where $z$ is a primitive root of unit of order $m$. Since $\tilde{L}^n$ the splitting field of a polynomial $p(x)$, it follows that $\tilde{L}^m(z)$ the splitting field of $p(x)(x^m - 1)$, and we conclude that $\tilde{L}^m(z)$ is normal. Reordering terms, we obtain the new tower:

$$\tilde{L}^m(z) = K_i(z) \supset \cdots \supset \tilde{K}_3 = K_2(z) \supset \tilde{K}_2 = K(z) \supset \tilde{K}_1 = K.$$

This new tower satisfies $\tilde{K}_{i+1} = \tilde{K}_i(d_i)$, with $d_i^{n_i} \in \tilde{K}_i$, for all $i$.

Let $G$ be the Galois group of $p(x)$ and $H = \text{Aut}_K(\tilde{L}^m(z))$. Then $\tilde{K}_i$ is an abelian extension of $\tilde{K}_{i-1}$. If the subgroup $H_i \subset H$ corresponds to the intermediate extension $\tilde{K}_i$, we have $H_{i-1} \triangleright H_i$ and $H_{i-1}/H_i$ is isomorphic to the Galois group of $\tilde{K}_i$ over $\tilde{K}_{i-1}$, i.e., is abelian. We conclude that $H$ admits an abelian tower, so it is a solvable group. Since $\tilde{L}^m(z)$ contains a splitting field of $p(x)$, $G$ is a subgroup of $H$ and it must be solvable.

(ii) If $G$ is solvable, then $p(x) = 0$ is solvable by radicals: Let $L$ be a splitting field of $p(x) = 0$ and $n = |G| = [L : K]$. If $K_1 = K$ and $K_2 = K(z)$, where $z$ is a primitive root of $x^n - 1 = 0$, then $M = L(z)$ has Galois group over $K_2$ isomorphic to a subgroup $H$ of $G$. Hence, $H$ is solvable and has a composition series $H = H_1 \triangleright H_2 \triangleright \cdots \triangleright e$, where each $H_i/H_{i+1}$ is cyclic of prime order. By the Galois correspondence, there exists a tower of subfields $K_2 \subset K_3 \subset \cdots \subset M$ where each $K_{i+1}$ is a normal extension over $K_i$ with Galois group cyclic of order prime $p_i$. Since $p_i | n$ and $K_i$ contains a
primitive root of $x^n - 1 = 0$, we see that $K_i$ contains the $p_i$-th roots of unit. By Proposition 7.8.10 we conclude that $K_{i+1} = K_i(d_i)$, with $d_i^{p_i} \in K_i$. This shows that the equation $p(x) = 0$ is solvable by radicals.

Exercises.

1. Show that \( \frac{1}{x^1} + \frac{1}{x^2} + \frac{1}{x^3} \) is a symmetric rational expression and determine its representation in terms of elementary symmetric polynomials.

2. Determine the integeres $1 \leq n \leq 10$ for which a regular polygon of $n$ sides admits a compass-and-straightedge construction.

3. Let $G \subset S_p$, with $p$ prime, be a subgroup which contains a cycle of length $p$ and a transposition. Show that one must have $G = S_p$.

4. Let $G$ be any finite group. Show that there exists fields $L$ and $K$ such that $L$ is an extension of $K$, with Galois group $G$. (HINT: By Cayley’s Theorem, one can assume that $G \subset S_n$ for some $n$.)

5. Using Galois Criterion, show that the Galois group of the equation $x^n - a = 0$ (over $\mathbb{Q}$) is solvable.
Chapter 8

Commutative Algebra (not yet available)
Appendix A

Set Theory

The notion of set is the most important of all mathematical notions, and the foundations of all mathematics rest upon it. The reader is for sure familiar with the informal notion of set and element of a set, as well as some elementary constructions with sets, such as (unions, intersections, complements, etc.). On the other hand, statements such as:

- two sets are equal if and only if they have the same elements,
- given two sets, there exists a set containing them,
- given any set, there exists a set formed by all its subsets,

are usually accepted as obvious. However, to fully justified them it is necessary a deeper look into the foundations of Set Theory, which is beyond the aim of this book. For example, the famous Russel paradox, concerning the existence of the set of all sets, can only be solved via an axiomtization of Set Theory. The interested reader can, e.g., consult the concise book by Paul Halmos escreveu a este respeto[1]. In this appendix we will limit ourselves to a few elementary notions and results of Set Theory which are essentially to the study of abstract algebra.

A.1 Relations and Functions

The notions of binary relation and map are directly related with the notion of ordered pair and, in fact, can be defined from the latter notion. From

a practical point of view, the most basic property of an ordered pair is the equivalence:

\[(A.1.1) \quad (x_1, y_1) = (x_2, y_2) \iff x_1 = x_2 \text{ and } y_1 = y_2.\]

The properties that an ordered pair should satisfy can be expressed in terms of even more basic notions of Set Theory, and lead to the previous equivalence:

**Definition A.1.1.** If \(X\) and \(Y\) are sets, \(x \in X\) and \(y \in Y\), the **ordered pair** \((x, y)\) is the set \((x, y) = \{x, \{x, y\}\}\). The elements \(x\) and \(y\) are called the **components** of the pair \((x, y)\). The set of all ordered pairs \((x, y)\), where \(x \in X\) and \(y \in Y\), are called the **cartesian product** of \(X\) and \(Y\), and is denoted by \(X \times Y\).

Note that \((A.1.1)\) is an immediate logical consequence of Definition A.1.1, and in particular the fact that \((x, y)\) is equal to \((y, x)\) if and only if \(x = y\). One can use the notion of ordered pair to formalize another important notion which will lead to the notion of map:

**Definition A.1.2.** A relation between \(X\) and \(Y\) is a subset \(R \subset X \times Y\). We will often write “\(xRy\)” instead of “\(x, y \in R\)”, and when \(X = Y\), we say that \(R\) is a **binary relation** on \(X\).

It is immediate to check that a relation \(R\) between \(X\) and \(Y\) has an associated relation between \(Y\) and \(X\), which is obtained by “switching” the components of every ordered pair in \(R\).

**Definition A.1.3.** If \(R\) is a relation between \(X\) and \(Y\), the **inverse or opposite relation** of \(R\), denoted by \(R^{op}\), is defined by setting \[R^{op} = \{(y, x) : (x, y) \in R\}.\]

There are several important classes of relations. In particular, the notions of **order relations**, **equivalence relations** and **maps**, play an essential role in many of our considerations.

**Definition A.1.4.** Let \(R\) be a relation on the set \(X\). We say \(R\) is an order relation on \(X\) if:

(i) **Transitivity**: For \(x, y, z \in X\), if \(xRy\) and \(yRz\), then \(xRz\).

(ii) **Anti-symmetry**: For \(x, y \in X\), if \(xRy\) and \(yRx\), then \(x = y\).
There are also various kinds of order relations, which are distinguished by adding one of the terms strict/non-strict e total/partial:

**Definition A.1.5.** Let $R$ be an order relation on $X$.

(i) $R$ is called a **total order relation** (the opposite of a partial order relation) if it satisfies the trichotomy property: For any $x, y \in X$, one of the following holds: $xRy$ or $yRx$ or $x = y$.

(ii) $R$ is called a **strict order relation** (the opposite of a non-strict order relation) if it satisfies the anti-reflexivity property: For $x, y \in X$, if $xRy$ then $x \neq y$.

The following examples illustrate all the various possibilities:

**Examples A.1.6.**

1. The usual relation “$>$” (larger than) between real numbers is a strict order relation and also a total order relation.

2. The usual relation “$\geq$” (larger or equal) between real numbers is a non-strict order relation and a total order relation.

3. The relation “$\supseteq$” (contains) between subsets of a given fixed set is a non-strict order relation and a partial order relation.

4. The relation “$\supset$” (strictly contains) between subsets of a given fixed set is a strict order relation and a partial order relation.

Given a partial ordered set $X$, with order relation denoted by “$\leq$”, any subset $Y \subset X$ is also partially ordered with the order relation induced on $X$ (which we will still denote by “$\leq$”). Obviously, the order relation induced on $X$ may have properties that the order relation on $X$ does not satisfy. For example, it may very well happened that $X$ is partially ordered while the induced ordered in $Y \subset X$ is a total order. In this case, the set $Y$ is called a chain in $X$.

**Example A.1.7.**

In $\mathbb{R}^2$ consider the partial order relation defined by:

\[(x_1, y_1) \leq (x_2, y_2) \text{ if and only if } y_1 = y_2 \text{ and } x_1 \leq x_2,\]

where the last inequality is the usual order relation for real numbers. In this case, any subset $\{(x, y) \in \mathbb{R}^2 : y = c\}$, with $c \in \mathbb{R}$ fixed (i.e., an horizontal line recta), is a chain. Note, also, that $\mathbb{R}^2$, with this order relation is not totally ordered.
The next class of binary relations is the following:

**Definition A.1.8.** Let \( R \) be a binary relation on the set \( X \). We call \( R \) an **equivalence relation** on \( X \) if it satisfies:

(i) **Reflexivity:** For all \( x \in X \), \( xRx \).

(ii) **Symmetry:** For \( x, y \in X \), if \( xRy \), then \( yRx \).

(iii) **Transitivity:** For \( x, y, z \in X \), if \( xRy \) and \( yRz \), then \( xRz \).

Here are some simple examples of equivalence relations.

**Examples A.1.9.**

1. The relation of parallelism between lines in the plane is an equivalence relation on the set of all lines.

2. The relation of equality between elements of any set \( X \) is an equivalence relation on \( X \).

3. The relation de congruence modulo \( m \) is an equivalence relation on the set of integers.

Any equivalence relation \( R \) on a set \( X \) determines an important collection of subsets of \( X \).

**Definition A.1.10.** Given an equivalence relation \( R \) on \( X \) and \( x \in X \), the **equivalence class** of \( x \), denoted by \( [x] \) or \( x \), is the set

\[
[x] = \{y \in X : xRy\}.
\]

The set of all equivalence classes \( [x] \) is called the **quotient set** of \( X \) by \( R \) and is denoted by \( X/R \), so that:

\[
X/R = \{[x] : x \in X\} = \{\{y \in X : xRy\} : x \in X\}.
\]

**Examples A.1.11.**

1. For the relation of parallelism between the lines in the plane, the equivalence class of a line \( L \) is formed by all lines parallel to \( L \).

2. If \( R \) is the relation of equality on the set \( X \), the quotient set is \( X/R = \{\{x\} : x \in X\} \).

3. If \( R \) is the congruence relation modulo \( m \) in the set of integers \( \mathbb{Z} \) and \( a \in \mathbb{Z} \), then \( [a] = \{a + km : k \in \mathbb{Z}\} \).
It is a basic fact that two equivalence classes either coincide or are disjoint sets. The proof is elementary.

**Proposition A.1.12.** Let $R$ be an equivalence relation on $X$. For $x, y \in X$, the following statements are equivalent:

(i) $xRy$;
(ii) $x = y$;
(iii) $x \cap y \neq \emptyset$.

Finally, we consider the following important class of equivalence relations, usually called “maps”.

**Definition A.1.13.** A relation $f$ between $X$ and $Y$ is called a map from $X$ to $Y$, denoted $f : X \to Y$, of:

(i) for each $x \in X$ there exists $y \in Y$ such that $xfy$, and
(ii) if $xfy$ and $xfy'$, then $y = y'$.

Because of (ii), we will write $y = f(x)$ instead of $xfy$. We then call $X$ the **domain** and $Y$ the **image** (or **range**) of the map $f$.

Other common names to denote a map are **function** and **transformation**.

**Examples A.1.14.**

1. If $X$ is a set, $I_X : X \to X$, given by $I_X(x) = x$, is called the **identity map** in $X$.
2. If $X \supset Y$ are sets, $i_Y : Y \to X$, given by $i_Y(y) = y$, is called the **inclusion map** of $Y$ in $X$.
3. If $R$ is an equivalence relation on $X$, the map $\pi_{X/R} : X \to X/R$, given by $\pi_{X/R}(x) = [x]$, is called the **quotient map** of $X$ in $X/R$.

In general, if $X \supset X'$ and $Y \supset Y'$, given a map $f : X \to Y$, we set:

$$ f(X') \equiv \{ f(x) : x \in X' \}, \quad \text{and} \quad f^{-1}(Y') \equiv \{ x \in X : f(x) \in Y' \}. $$

We say that $f(X')$ is the **direct image** of $X'$ by $f$, and $f^{-1}(Y')$ is an **inverse image** of $Y'$ by $f$. In particular, the **image** of $f$ the set $f(X)$, which we will often denote by $\text{Im} \ f$. 
Example A.1.15.

If \( f : \mathbb{R} \to \mathbb{R} \) is the map \( \cos x \), its image is the set \( f(\mathbb{R}) = [-1, +1] \). The inverse image of the set \( \{0\} \) is \( f^{-1}(\{0\}) = \{\frac{2n+1}{2} \pi : n \in \mathbb{Z}\} \).

As we all know certain kinds of maps have special designations:

Definition A.1.16. If \( f : X \to Y \) is a map, then \( f \) is called:

(i) **surjective** or **onto** if for all \( y \in Y \) there exists \( x \in X \) such that \( y = f(x) \);

(ii) **injective** if \( f(x) = f(x') \iff x = x' \);

(iii) **bijective** or a **bijection** if it is both injective and surjective. In this case, we say that sets are **equipotent** or **isomorphic**.

Examples A.1.17.

1. The identity map \( I_X : X \to X \) is bijective.
2. The inclusion map \( i_Y : Y \to X \) is injective.
3. The quotient map \( p_{X/R} : X \to X/R \) is surjective.
4. The map \( \cos : \mathbb{R} \to \mathbb{R} \) is neither injective nor surjective.

Given maps \( f : X \to Y \) and \( g : Y \to Z \), the **compose** of \( f \) and \( g \) is the map \( g \circ f : X \to Z \) (read “\( g \) after \( f \)”) given by \( (g \circ f)(x) = g(f(x)) \). We leave for the exercises to verify that:

Proposition A.1.18. If \( X, Y, Z, \) and \( W \) are sets. Then:

(i) **Associativity**: If \( f : X \to Y, \) \( g : Y \to Z \) and \( h : Z \to W \) are maps, then \( (h \circ g) \circ f = h \circ (g \circ f) \);

(ii) **Left Inverse**: \( f : X \to Y \) is injective if and only if there exists \( g : Y \to X \) such that \( g \circ f = I_X \);

(iii) **Right Inverse**: \( f : X \to Y \) is surjective if and only if there exists \( g : Y \to X \) such that \( f \circ g = I_Y \);

(iv) **Inverse**: \( f : X \to Y \) is bijective if and only if there exists \( g : Y \to X \) such that \( f \circ g = I_Y \) and \( g \circ f = I_X \). In this case, \( g \) is called the **inverse map** of \( f \) and is denoted by \( f^{-1} \).
Exercises.

1. Use Definition (A.1.1) to prove that (A.1.1) holds.

2. Describe the anti-symmetry and trichotomy properties of a relation $R$ in terms of $R$ and its inverse $R^{op}$.


4. Show that, if $f : X \to Y$ is a map and $\{Y_i\}_{i \in I}$ is a family of subsets of $Y$, then the following identities hold:

   $$f^{-1}(\bigcup_{i \in I} Y_i) = \bigcup_{i \in I} f^{-1}(Y_i), \quad \text{and} \quad f^{-1}(\bigcap_{i \in I} Y_i) = \bigcap_{i \in I} f^{-1}(Y_i).$$

   Are these still true if we assume that $\{X_i\}_{i \in I}$ is a family of subsets of $X$ and we replace $f^{-1}$ por $f$?

5. Let $f : X \to Y$ be a map.

   (a) Can there exist more that one map $g : Y \to X$ such that $f \circ g = I_Y$?

   (b) Can there exist more that one map $g : Y \to X$ such that $g \circ f = I_X$?

   (c) Can there exist more that one map $g : Y \to A$ such that $f \circ g = I_Y$ and $g \circ f = I_X$?

6. Prove items (i) and (ii) of Proposition (A.1.13).

7. Show that the inverse $f^{-1}$ of a map $f : X \to Y$ is:

   (a) a map $f^{-1} : f(X) \to X$ if $f$ is injective;

   (b) a map $f^{-1} : Y \to X$ if $f$ is bijective.

8. Verify that, if $f : X \to Y$ is a bijective map, then $g = f^{-1}$ is the only map such that $f \circ g = I_Y$ and $g \circ f = I_X$.

9. Show that, if $X \neq \emptyset$ and $Y = \emptyset$, there are no maps $f : X \to Y$.

10. If $X = \emptyset$ and $f : X \to Y$, then $f$ is injective, and moreover $f$ is surjective if and only if $Y = \emptyset$.

\[\text{The proofs of items (iii) and (iv) require the Axiom of Choice which will be mentioned later.}\]
A.2 Axiom of Choice, Zorn’s Lemma and Induction

We have mentioned several times the notion of cartesian product of sets. Using the notion of map, we can introduce define the cartesian product of any arbitrary number of sets.

The set of of integers will play a relevant role in the discussion and given a positive integer \( n \in \mathbb{N} \) we will denote in this section by \( I_n = \{ k \in \mathbb{N} : k \leq n \} \) the set consisting of the first \( n \) natural numbers. In particular, \( I_0 = \emptyset \).

Given a set \( X \), a map \( f : I_2 \rightarrow X \) determines uniquely an ordered pair with components in \( X \), namely the pair \((f(1), f(2))\), which we will also write as \((f_1, f_2)\). Hence, the set of all maps \( f : I_2 \rightarrow X \) is isomorphic to the set of all ordered pairs with components in \( X \).

A similar observation holds if we consider maps \( f : I_2 \rightarrow X \cup Y \) such that \( f(1) \in X \) and \( f(2) \in Y \). The set of all such maps \( f : I_2 \rightarrow X \cup Y \) is isomorphic to the set of ordered pairs \((x, y)\), with \( x \in X \) and \( y \in Y \), i.e., to the cartesian product \( X \times Y \). This leads us to define arbitrary cartesian products as sets of maps.

Definition A.2.1. Let \( X_1, X_2, \ldots, X_n \) be sets. The cartesian product \( \prod_{i=1}^n X_i \) is the set of all maps \( f : I_n \rightarrow \bigcup_{i=1}^n X_i \) such that \( f(k) \in X_k \).

If \( f \in \prod_{i=1}^n X_i \) we write \( f = (f(1), f(2), \ldots, f(n)) \), or \( f = (f_1, f_2, \ldots, f_n) \). Also, when the sets \( X_i \) all coincide with the same set \( X \), we will write \( X^n \) instead of \( \prod_{i=1}^n X_i \) and called it the power \( n \) of \( X \). In this case, the elements of \( X^n \) are called \( n \)-tuples of elements of \( X \).

One can apply the method used in Definition A.2.1 to define the cartesian product of an arbitrary number of sets. For that, we replace the family of sets \( X_1, X_2, \ldots, X_n \), indexed by a natural number \( 1 \leq k \leq n \), by a family \( \{X_i : i \in I\} \), indexed by a parameter \( i \) belonging to an arbitrary set \( I \).

Definition A.2.2. Given a family \( \{X_i : i \in I\} \), the cartesian product \( \prod_{i \in I} X_i \) is given by

\[
\prod_{i \in I} X_i = \{ f : I \rightarrow \bigcup_{i \in I} X_i, \text{ with } f(i) \in X_i \text{ para qualquer } i \in I \}. 
\]

The canonical projection \( \pi_k : \prod_{i \in I} X_i \rightarrow X_k \) is the map associating to an element \( f \in \prod_{i \in I} X_i \) the element \( f(k) \in X_k \).

\[\text{Recall that given some class of sets } X, \text{ a family indexed by } i \in I \text{ is just a map } f : I \rightarrow X. \text{ We write } X_i \text{ instead of } f(i), \text{ just like we write, for example, } x_n \text{ instead of } f(n) \text{ when we deal with a sequence of real numbers.}\]

\[\text{We will often write } f_k, \text{ instead of } f(k).\]
Examples A.2.3.

1. Given a set $X$, the cartesian product $\prod_{n \in \mathbb{N}} X$ is just the set of all sequences with values in $X$.

2. The cartesian product $\prod_{n \in \mathbb{N}} I_n$ is the set of sequences of natural numbers $f : \mathbb{N} \to \mathbb{N}$ such that $f(n) \leq n$, for all $n \in \mathbb{N}$.

3. The cartesian product $\prod_{x \in \mathbb{R}} [x-1, x+1]$ is the set of all functions $f : \mathbb{R} \to \mathbb{R}$ such that $x - 1 \leq f(x) \leq x + 1$.

If we consider all maps with a fixed domain $X$ and fixed range $Y$, this is just the cartesian product $\prod_{x \in X} Y$, but it is common to denote it by the symbol “$Y^X$”.

Definition A.2.4. $Y^X = \{ f : X \to Y \}$ is the set of all maps from $X$ to $Y$.

Examples A.2.5.

1. $X^\mathbb{N}$ is the set consisting of all sequences in $X$.

2. $\mathbb{R}^\mathbb{R}$ denotes the set of all real valued functions of one variable.

3. $\mathbb{R}^{\mathbb{R} \times \mathbb{R}}$ is the set consisting of all real valued functions of two variables.

4. In general, $X^{X \times X}$ is the set of all binary operations on $X$.

5. One can also denote the set $X^n$ by $X^I$.

In general, the cartesian product $\prod_{i \in I} X_i$ is a strict subset of $(\bigcup_{i \in I} X_i)^I$, since not all maps $f : I \to \bigcup_{i \in I} X_i$ satisfy the condition $f(i) \in X_i$, for all $i \in I$. However, when $X_i = X$ for every $i \in I$, it is obvious that

$$\prod_{i \in I} X_i = \bigcap_{i \in I} X = X^I.$$ 

Hence, the set $X^I$ is a special type of cartesian product and this justifies the use of the “power” notation.

One shows easily (by induction) that the product $\prod_{i=1}^n X_i$ of any finite number of sets is non-empty, as long as the sets $X_i$ are non-empty. The very same statement, but for an arbitrary number of sets, is a fundamental axiom of Set Theory:

Axiom I (Axiom of Choice). If $I \neq \emptyset$ and $X_i \neq \emptyset$, for all $i \in I$, then $\prod_{i \in I} X_i \neq \emptyset$. 
The reason for the designation for this axiom is easy to explain. An element \( f \) of the cartesian product \( \prod_{i \in I} X_i \) amounts to a “choice” of an element in each of the sets \( X_i \). The axiom states that given an arbitrary family of sets, there exists always a map which “chooses” exactly one element from each set.

The next example illustrates the kind of issue that the Axiom of Choice allows to bypass.

**Example A.2.6.**

Assume that each \( X_i \) is formed a single pair of shoes. Then we can decide to choose, for example, the right shoe of each pair. In this case, the Axiom of Choice is useless. On the other hand, if each \( X_i \) is formed by a pair of socks, then there is no criteria for a choice of an element in \( X_i \), and we can invoke the Axiom of Choice to claim the existence of a set formed by exactly one sock from each set \( X_i \).

Many results concerning the existence of some set, or of an element in a set, satisfying some property, can be reformulated in terms of the existence of a maximal element for an appropriate order relation. In this respect, the following result often plays a crucial role:

**Theorem A.2.7** (Zorn’s Lemma). Let \( X \) be a partially ordered, non-empty, where every chain possesses an upper bound. Then \( X \) contains a maximal element.

One can show that Zorn’s Lemma is equivalent the Axiom of Choice (see P. Halmos book cited at the beginning of this appendix). The next example illustrates how Zorn’s Lemma can be applied.

**Example A.2.8.**

Let \( A \) be a ring, \( I \subset A \) an ideal, and denoted by \( X \) the set of all proper ideals of \( A \) that contains \( I \). Since \( I \) belongs to \( X \), this set is non-empty. We consider on \( X \) the relation of inclusion, which is obviously a partial order relation. If \( \{ I_j : j \in J \} \) is a chain in \( X \), then \( \bigcup_{j \in J} I_j \) is an upper bound (you should check that this is indeed an ideal of \( A \) which contains \( I \) and hence belongs to \( X \)). We conclude from Zorn’s Lemma that \( X \) has a maximal element. This shows in an arbitrary ring \( A \), every proper ideal is contain in a maximal ideal.

---

5The notions of upper/lower bound, supremum/infimum and maximal/minimal element for subsets of partial ordered sets is discussed in Section 23.
One can imagine an alternative proof of the result in this example, as follows: given a proper ideal \( \subset A \), if \( I \) is not maximal, then there exists an ideal \( I_0 \) containing \( I \). Then, if \( I_0 \) is not maximal, there exists an ideal \( I_1 \) which contains \( I_0 \), and so on. The problem, of course, is that "so on" may not terminate. Zorn's Lemma is useful precisely to avoid this kind of problem.

For basically the same reason, a partial ordered set \( X \) may fail to contain a minimal element. Even if a minimal element exists, a subset \( Y \subset X \) may fail to have a minimal element. The following definition attempts to prevent this.

**Definition A.2.9.** A partial ordered set \( X \) is said to be well-ordered if every non-empty subset \( S \subset X \) has a minimal element.

Obviously, the order relation in a well-ordered set \( X \) is a total order\(^{6}\) if \( x, y \in X \), the set \( \{x, y\} \) has a minimal element, hence either \( x \leq y \) or \( y \leq x \).

**Examples A.2.10.**

1. The set \( \mathbb{N} \), with the usual order relation, is a well-ordered set.
2. The set \( \mathbb{Z} \), with the usual order relation, is not a well-ordered set since, e.g., the subset \( \{n \in \mathbb{Z} : n \leq 0\} \) does not have a minimal element.

Another result which is equivalent to the Axiom of Choice and, hence, also to Zorn's Lemma, is the following:

**Theorem A.2.11 (Well-Ordering Principle).** Every set can be well-ordered.

Again we refer to the book of P. Halmos for a proof of this result.

**Example A.2.12.**

We observed above that the set of integers \( \mathbb{Z} \), with the usual order relation, is not well-ordered. However, we have, for example, the following well-ordering of \( \mathbb{Z} \):

\[0, 1, -1, 2, -2, \ldots, n, -n, \ldots\]

The main feature of well-ordered sets is the possibility of generalizing the usual induction method to these sets, as we now explain. Given a partial

\(^{6}\)From now on, unless otherwise mentioned, we denote a (non-strict) order relation on set \( X \) by the symbol "\( \leq \)".
ordered set let us denote by $s(x)$ the set of all elements strictly less than $x$, also called the “segment” ending at $x$:

$$s(x) = \{y \in X : y \leq x, \text{ and } y \neq x\}.$$ 

We then have:

**Theorem A.2.13** (Transfinite Induction). Let $X$ be an well-ordered set, and $S \subset X$ a subset such that:

$$\forall x \in X, s(x) \subset S \Rightarrow x \in S.$$ 

Then $S = X$.

**Proof.** If $X - S$ is non-empty, denote by $x$ its minimal element. Then the segment $s(x)$ is contained in $S$. By the “induction hypothesis”, $x \in S$. Since $x$ cannot belong at the same time to $S$ and to $X - S$, we must have $X - S$ empty, and $X = S$. \qed

The Transfinite Induction method has a large range of application, as it follows from the Well-Ordering Principle. For now, we will use it to study recursive definitions, including transfinite recursive definitions.

If $A$ is a set and $n \in \mathbb{N}$, consider the set $A^n$ of all maps $f : I_n \to A$, i.e., all the $n$-tuples $(x_1, x_2, \cdots, x_n)$ in $A$. Consider also the class $\Phi = \bigcup_{n \in \mathbb{N}} A^n$.

**Definition A.2.14.** A map $F : \Phi \to A$ is called a recursive formula.

The justification for this name is that a map $F : \Phi \to A$ allows to associate to a $n$-tuple $f_n = (x_1, x_2, \cdots, x_n)$ the $(n+1)$-tuple given by $f_{n+1} = (x_1, x_2, \cdots, x_n, x_{n+1})$, where $x_{n+1} = F(f_n) = F(x_1, x_2, \cdots, x_n)$.

**Theorem A.2.15** (Recursive Definitions). Given an element $x_1 \in A$ and a recursive formula $F : \Phi \to A$, there exists a unique sequence $\phi : \mathbb{N} \to A$ such that

(i) $\phi(1) = x_1$, and

(ii) $\phi(k + 1) = F(\phi|_{I_k})$, for all $k \in \mathbb{N}$.

where we have denoted by $\phi|_{I_k}$ the restriction of $\phi$ to the set $I_k$.

**Proof.** Using induction, we prove first that for every $n \in \mathbb{N}$ there exists $f_n \in A^n$ satisfying:

(1) $f_n(1) = x_1$, and
A.2. AXIOM OF CHOICE, ZORN’S LEMMA AND INDUCTION

(2) \( f_n(k + 1) = F(f_n|_{I_k}) \), for all \( k < n \).

The result is obvious for \( n = 1 \), if we set \( f_1(1) = x_1 \) and if we observe that in this case condition (2) is empty. Assume the result holds for some \( n \geq 1 \): there exists an \( n \)-tuple \( f_n = (x_1, x_2, \cdots, x_n) \in A^n \) satisfying (1) and (2). Then we define \( f_{n+1} = (x_1, x_2, \cdots, x_n, x_{n+1}) \in A^{n+1} \), where \( x_{n+1} = F(f_n) = F(x_1, x_2, \cdots, x_n) \). One checks immediately that \( f_{n+1} \) automatically satisfies (1), and it also satisfies (2) for \( k < n + 1 \).

Assume now that \( f_n \in A^n \) and \( f_m \in A^m \) satisfy both (1) and (2). We assume, without loss of generality, that \( n < m \), and we prove that \( f_n \) is the restriction of \( f_m \) to \( I_n \). For that, consider the set \( D = \{ k \in I_n : f_n(k) \neq f_m(k) \} \). Assuming \( D \) non-empty, let \( r + 1 \) be its minimal element, and note that \( r \geq 1 \), because by assumption \( f_n(1) = f_m(1) = x_1 \). Hence, we have that \( f_n(k) = f_m(k) \) for all \( k \leq r \), so the restrictions \( f_n|_{I_r} \) and \( f_m|_{I_r} \) are equal. But then we must have \( f_n(r + 1) = F(f_n|_{I_r}) = F(f_m|_{I_r}) = f_m(r + 1) \), contradicting that \( r + 1 \in D \).

In order to conclude the proof, since we already know that for any \( n \in \mathbb{N} \) there exists exactly one element \( f_n \in A^n \) satisfying (1) and (2), we define \( \phi : \mathbb{N} \to A \) by \( f(n) = f_n(n) \). One checks easily that this is the only sequence that satisfies (i) and (ii).

The previous result maybe generalized, replacing \( \mathbb{N} \) by any well-ordered set \( X \). In this case, the sets \( I_n \) are replaced by the segments \( s(x) = \{ y \in X : y < x \} \), \( A_x = A^{s(x)} \) is then the set of all maps \( f : s(x) \to A \) and \( \Phi = \bigcup_{x \in X} A_x \). A TRANSFINITE RECURSIVE FORMULA is again a map \( F : \Phi \to A \).

The proof of the following generalization of Theorem A.2.15 is left for the exercises:

**Theorem A.2.16 (Transfinite Recursive Definitions).** Given a transfinite recursive formula \( F : \Phi \to A \), there exists a unique map \( f : X \to A \) such that \( f(x) = F(f|_{s(x)}) \), for all \( x \in X \).

**Exercises.**

1. Show that \( X \times Y \) is isomorphic to \( Y \times X \). Under what conditions does the equality \( X \times Y = Y \times X \) hold?

2. Show that \( X \times (Y \times Z) \) is isomorphic to \( (X \times Y) \times Z \).

3. Describe the sets \( X^\emptyset \), \( \emptyset^X \) and \( \emptyset^\emptyset \).

4. Show that the sets \( X^{Y \cup Z} \) and \( X^Y \times X^Z \) are isomorphic.
5. Show that the sets $X^Y \times Z$ and $(X^Y)^Z$ are isomorphic, whenever $Y \cap Z = \emptyset$.

6. Show that $(X \times Y)^Z$ is isomorphic to $X^Z \times Y^Z$.

7. Use the Axiom of Choice to prove that $f : X \to Y$ is surjective if and only if there exists $g : Y \to X$ such that $f \circ g = I_Y$.

8. Show that, if $I \neq \emptyset$ and $X_i \neq \emptyset$ for all $i \in I$, then the canonical projections $\pi_k : \prod_{i \in I} X_i \to X_k$ are surjective.

9. Use Zorn’s Lemma to show that for an arbitrary group, every proper subgroup is contained in a maximal subgroup.

10. Prove the following statements:
    (a) Every partially ordered set has a maximal chain;
    (b) Every chain in a partially ordered set is contained in a maximal chain.

11. Show that the Principle of Transfinite Induction is equivalent to the usual Principle de Induction (see Chapter 2) when $X = \mathbb{N}$.

12. Give an example of a well-ordering of $\mathbb{Q}$.

13. Show that a totally ordered set $X$ is well-ordered if and only if for all $x \in X$ the segment $s(x)$ is well-ordered.

14. Prove Theorem A.2.16. Explain why the statement of this result does not mention an element $x_1$ as was the case with Theorem A.2.15.

### A.3 Finite Sets

Intuition says that a set $X$ is finite if one can “count” its elements. The prototype of a finite set with $n \geq 0$ elements is given by the set formed by the first $n$ natural numbers:

$$I_n = \{1, 2, 3, \ldots, n\} = \{k \in \mathbb{N} : k \leq n\}.$$

When $n = 0$, we have the empty set $I_0 = \emptyset$. A moment’s thought shows that by “counting” we really mean establishing a bijective correspondence (map) between $X$ and $I_n$. More formally, we define:

**Definition A.3.1.** A set $X$ is called **finite** if it is isomorphic to $I_n$, for some $n \geq 0$. If $X$ is not isomorphic to any $I_n$, then $X$ is called **infinite**.
Examples A.3.2.

1. The set $I_n$ is obviously isomorphic to itself, hence it is a finite set.

2. We leave as an exercise to check that a subset $X \subset \mathbb{Z}$ is finite if and only if it is bounded.

We have mentioned in Chapter 1 that $X$ is infinite if and only if there exists an injective map $\phi : X \to X$ which is not surjective. This will be established in the next section, where we study in some detail infinite sets. For now, we limit ourselves to finite sets. Let us look first at the sets $I_n$.

**Lemma A.3.3.** If $\phi : I_n \to I_n$ is injective, then $\phi$ is surjective.

*Proof.* We use induction. When $n = 0$, there is nothing to prove.

Assuming that the result holds for some $n$, let $\phi : I_{n+1} \to I_{n+1}$ be an injective map, and let $\alpha = \phi(n+1)$. Consider (see Figure A.3.1) the bijective map $\psi : I_{n+1} \to I_{n+1}$ given by $\psi(n+1) = \alpha$, $\psi(\alpha) = n + 1$, and $\psi(x) = x$ in all other cases ($\psi$ “exchanges” the natural numbers $\alpha$ and $n + 1$, and is the identity if $x \neq \alpha, n + 1$, but this last fact is irrelevant for the proof).

Now we define $\phi^* = \psi \circ \phi$ and observe that $\phi^*$ is injective, since it is the composition of injective maps. Also, it satisfies $\phi^*(n+1) = n + 1$, $\alpha = \phi(n+1)$.

![Figure A.3.1: The maps $\phi$, $\phi^*$ and $\psi$.](image)

A map $f : X \to Y$ is just a set of ordered pairs with some special properties. It is possible that $f$ is the empty set, and this happens exactly when $X$ is also the empty set. In this case, $f$ is necessarily injective, and it is surjective if $Y$ is also the empty set.
by the definition of $\psi$. It follows that if $x \in I_n$, then \(\phi^*(x) \neq n + 1\), so that $\phi^*(x) \in I_n$, or equivalently $\phi^*(I_n) \subseteq I_n$.

This shows that the restriction of $\phi^*$ to $I_n$ is an injective map from $I_n$ into $I_n$. By the induction hypothesis, this restriction is surjective: $\phi^*(I_n) = I_n$. Since $\phi^*(n+1) = n + 1$, conclude that $\phi^*(I_{n+1}) = I_{n+1}$, i.e., $\phi^*$ is a surjective map from $I_{n+1}$ onto $I_{n+1}$.

Finally, we conclude that $\phi = \psi^{-1} \circ \phi^*$ is surjective, since it is the composition of surjective maps.

**Proposition A.3.4.** If $X$ is finite and $\phi : X \to X$ is injective, then $\phi$ is surjective.

**Proof.** Let $\Psi : I_n \to X$ be a bijection. Note that $\phi^* = \Psi^{-1} \circ \phi \circ \Psi : I_n \to I_n$ is injective, since it is the composition of injective maps (see Figure A.3.2). According to Lemma A.3.3, $\phi^*$ is necessarily surjective.

\[
\begin{array}{ccc}
X & \xrightarrow{\phi} & X \\
\downarrow{\Psi} & & \downarrow{\Psi^{-1}} \\
I & \xleftarrow{\phi^* = \Psi^{-1} \circ \phi \circ \Psi} & I
\end{array}
\]

Figure A.3.2: The maps $\phi$, $\phi^*$ and $\Psi$.

It follows that $\phi = \Psi \circ \phi^* \circ \Psi^{-1}$ is a composition of surjective maps, and hence it is surjective.

**Corollary A.3.5.** If $\phi : X \to X$ is injective but not surjective, then $X$ is an infinite set.

**Example A.3.6.**

In Chapter 2, we observed that the map $f : \mathbb{N} \to \mathbb{N}$ given by $f(n) = n + 1$ is injective and not surjective. From the previous result, we conclude that $\mathbb{N}$ is infinite.

Another consequence of the proposition above is that a finite set cannot be isomorphic to a proper subset of it. The example above shows that this
fails for infinite sets. We state this result as follows and leave the proof for the exercises:

**Corollary A.3.7.** If $X$ is finite, $X \supseteq Y$ and $\phi : X \to Y$ is injective, then $X = Y$.

It seems obvious that $X$ is isomorphic to $I_n$ if and only if $X$ has $n$ elements, in which case $X$ cannot be isomorphic to $I_m$ for $m \neq n$. In fact, this is a direct consequence of Proposition A.3.4.

**Corollary A.3.8.** If $\phi : I_n \to X$ and $\psi : I_m \to X$ are both bijections, then $n = m$.

**Proof.** Assume, without loss of generality, that $m \leq n$, so that $I_m \subset I_n$. Then the map $\Psi = \psi^{-1} \circ \phi : I_n \to I_m$ gives an injective map from $I_n$ into its subset $I_m$. By Corollary A.3.7 we must have $I_m = I_n$, i.e., $n = m$.

Hence, if $X$ is finite there exists a unique non-negative integer $n$ such that $X$ is isomorphic to $I_n$. In this case, we say that $X$ has $n$ elements, and we call $n$ the **cardinal number** or **cardinality** of $X$. We will use often the symbol $\# X$ to denote the cardinality of $X$.

We will list a few properties of the cardinality of finite sets. We only sketch the proofs, leaving the details for the exercises.

**Proposition A.3.9.** If $Y$ is a subset of the set $X$, then:

(i) If $X$ is finite, then $Y$ is also finite and $\# Y \leq \# X$.

(ii) If $X$ is finite and $\# Y = \# X$, then $X = Y$.

(iii) If $Y$ is infinite, then $X$ is also infinite.

The proof uses the following two lemmas, the first of which has an elementary proof and so it is left as an exercise.

**Lemma A.3.10.** If $\phi : I_n \to X$ is injective but not surjective, then there exists $\phi^* : I_{n+1} \to X$ injective.

**Lemma A.3.11.** If $\phi : X \to I_n$ is injective, then $X$ is finite and $\# X \leq n$.

**Proof.** Let $M(X)$ be the set of integers $m \geq 0$ for which there exists an injective map $\Psi_m : I_m \to X$. Note that $0$ belongs to $M(X)$, and we claim that $M(X)$ is bounded above.
If $\Psi_m : I_m \to X$ is injective, the composition $\phi \circ \Psi : I_m \to I_n$ is also injective. According to Corollary A.3.7, we cannot have $n < m$, i.e., $n$ is an upper bound for $M(X)$.

Hence, $M(X)$ has a maximum $k$. By Lemma A.3.10, we have $k = \#X$. Since $n$ is an upper bound for $M(X)$, it follows that $\#X = k \leq n$. 

Our last proposition expresses in a precise form some of our intuition about the meaning of the addition and the product of natural numbers.

**Proposition A.3.12.** If $X$ and $Y$ are finite sets, then $X \cup Y$ and $X \times Y$ are finite sets and we have:

(i) $\#(X \cup Y) \leq \#X + \#Y$;

(ii) $\#(X \cup Y) = \#X + \#Y$, if $X$ and $Y$ are disjoint;

(iii) $\#(X \times Y) = (\#X)(\#Y)$.

**Proof.** We prove only item (ii), leaving the remaining statements for the exercises.

Let $\phi : I_n \to X$ and $\psi : I_m \to Y$ be bijections, so that $\#X = n$ and $\#Y = m$. Define a map $\Psi : I_{n+m} \to X \cup Y$ by:

$$\Psi(k) = \begin{cases} 
\phi(k) & \text{if } 1 \leq k \leq n, \\
\psi(k-n) & \text{if } n + 1 \leq k \leq n + m.
\end{cases}$$

Obviously, $\Psi$ is surjective and injective (since $X$ and $Y$ are disjoint). Hence,

$$\#(X \cup Y) = n + m.$$

**Exercises.**

In these exercises, the symbols $X$ and $Y$ denote arbitrary sets.

1. Prove Corollary A.3.7

2. Prove Lemma A.3.10

3. Let $\phi : I_n \to X$ be a surjective map. Show that $\Psi : X \to I_n$, given by $\Psi(x) = \max\{k \in I_n : \phi(k) = x\}$ is injective.

4. Show that the following statements are equivalent:
A.4. INFINITE SETS

(a) $X$ is finite and $\#X \leq n$.
(b) There exists an injective map $\phi : X \rightarrow I_n$.
(c) There exists a surjective map $\psi : I_n \rightarrow X$.

5. Let $X$ be a finite set. Show that if either $\phi : X \rightarrow Y$ is surjective or $\psi : Y \rightarrow X$ is injective, then $Y$ is finite and $\#Y \leq \#X$.


7. Show that if $X$ is finite, then $\#X = \#(X - Y) + \#(X \cap Y)$.

8. Let $X$ and $Y$ be finite. Show that $X \cup Y$ is also finite and
\[
\#(X \cup Y) = \#X + \#Y - \#(X \cap Y).
\]

9. Let $X$ and $Y$ be finite. Show that $X \times Y$ is finite and $\#(X \times Y) = (\#X)(\#Y)$.

10. Let $X_1, X_2, \ldots, X_n$ be finite sets. Show that:
   (a) $\#(\bigcup_{k=1}^{n} X_k) \leq \sum_{k=1}^{n} \#X_k$;
   (b) $\#(\bigcup_{k=1}^{n} X_k) = \sum_{k=1}^{n} \#X_k$, if the sets $X_k$’s are disjoint;
   (c) $\#(\prod_{k=1}^{n} X_k) = \prod_{k=1}^{n} \#X_k$.

11. Assume that $\#X = n$ and $\#Y = n$. Find the cardinality of the following sets:
   (a) The set $Y^X$ of all maps $f : X \rightarrow Y$.
   (b) The set of all injective maps $f : X \rightarrow Y$.
   (c) The set of all surjective maps $f : X \rightarrow Y$.

12. Assuming that $\#X = n$, show that:
   (a) $X$ has $2^n$ distinct subsets;
   (b) $X$ has $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ subsets with exactly $k$ elements.

13. Let $X \subset \mathbb{Z}$. Show that $X$ is finite if and only if $X$ is bounded.

A.4 Infinite Sets

We have already seen some elementary results on infinite sets. In particular, we have seen that $\mathbb{N}$ is an infinite set. We will now study some more properties of infinite sets. We start by showing that $\mathbb{N}$ is the “smallest” infinite set.
Lemma A.4.1. $X$ is infinite if and only if $X$ contains a subset isomorphic to $\mathbb{N}$.

Proof. If there exists a subset $Y \subset X$ and a bijection $\phi : \mathbb{N} \to Y$, it follows that $Y$ is infinite, and hence $X$ is also infinite, according to Proposition A.3.9 (iii).

Conversely, assume that $X$ is infinite. We claim that there exists an injective map $\phi : \mathbb{N} \to X$. Taking $Y = \phi(\mathbb{N})$, this completes the proof. We will define the map $\phi$ recursively.

Since $X \neq \emptyset$, there exists $x_1 \in X$ and we set $\phi(1) = x_1$. Assume that $\phi$ has been defined and is injective in $\{1, 2, \ldots, n\}$. The set $Z_n = X - \{\phi(1), \ldots, \phi(n)\}$ is non-empty, for otherwise $X$ would be finite. Choose $z$ some element of $Z_n$, and set $\phi(n + 1) = z$. This defines an injective map $\phi : \mathbb{N} \to X$ recursively. □

The previous result allows to complete Corollary A.3.5 and to justify the characterization of infinite sets mentioned in the exercises of Section 2.1.

Theorem A.4.2. $X$ is infinite if and only if there exists $\phi : X \to X$ injective and not surjective.

Proof. Corollary A.3.5 shows that if there exists $\phi : X \to X$ injective and not surjective, then $X$ is infinite. It remains to show that if $X$ is infinite, then there exists such a map.

If $X$ is infinite, Lemma A.4.1 yields an injective map (sequence) $\psi : \mathbb{N} \to X$. Let $Y = \psi(\mathbb{N})$ so that $\psi : \mathbb{N} \to Y$ is a bijection. Now we define $\phi : X \to X$ as follows:

$$\phi(x) = \begin{cases} x & \text{if } x \notin Y, \\ \psi(\psi^{-1}(x) + 1) & \text{if } x \in Y. \end{cases}$$

One checks easily that $\phi$ is injective and not surjective. □

All the properties of infinite sets that we have studied so far are, perhaps, not surprising. Our next result is less obvious: it states that Lemma A.4.1 cannot be improve, i.e., there are infinite sets which are not isomorphic to $\mathbb{N}$. In what follows we will denote by $\mathcal{P}(X)$ the set consisting of the subsets of $X$.

Theorem A.4.3 (Cantor). Let $\Psi : X \to \mathcal{P}(X)$ be any map. Then $\Psi$ is not surjective.
Proof. The proof is by contradiction, and resembles the ideas behind Rus- sel’s paradox, that we mentioned in Chapter ??.

Let \( \Psi : X \to \mathcal{P}(X) \) and define a subset \( Y \subset X \) by:

\[
Y = \{ x \in X : x \notin \Psi(x) \}.
\]

Since \( Y \in \mathcal{P}(X) \), if \( \Psi \) is surjective there exists an element \( y \in X \) such that \( Y = \Psi(y) \). Obviously, either \( y \in Y \) or \( y \notin Y \). We claim that both cases lead to a contradiction:

(i) If \( y \in Y = \Psi(y) \), then the definition of \( Y \) gives that \( y \notin Y \), a contra-
diction;

(ii) If \( y \notin Y = \Psi(y) \) it follows from the definition of \( Y \), that \( y \in Y \), which
is also a contradiction.

It follows that there is no \( y \in X \) such that \( Y = \Psi(y) \), hence \( \Psi \) is not surjective.

Examples A.4.4.

1. In order to illustrate the proof above, let \( X = \{0, 1\} \), so that:

\[
\mathcal{P}(X) = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}.
\]

If \( \Psi : X \to \mathcal{P}(X) \) is given by \( \Psi(0) = \{1\} \) and \( \Psi(1) = \{0, 1\} \), we have that

\[
Y = \{ x \in X : x \notin \Psi(x) \} = \{0\},
\]

so obviously \( Y \notin \Psi(X) = \{\{1\}, \{0, 1\}\} \).

2. Consider the set \( \mathcal{P}(\mathbb{N}) \) whose elements are the sets containing only natural numbers. The result above says that this set is not isomorphic to \( \mathbb{N} \). On the other hand, the map \( \Phi : \mathbb{N} \to \mathcal{P}(\mathbb{N}) \) given by \( \Phi(n) = \{n\} \) is obviously injective, and hence \( \mathcal{P}(\mathbb{N}) \) is infinite.

Definition A.4.5. A set \( X \) is called countable if it is finite or isomorphic
to \( \mathbb{N} \). Otherwise, \( X \) is called (infinite) uncountable or

Therefore, as we saw above, \( X \) is an infinite set if and only if it contains an infinite countable subset, but there are infinite uncountable sets, such as \( \mathcal{P}(\mathbb{N}) \). On the other hand, there are also uncountable sets, not isomorphic to \( \mathbb{N} \), and also not isomorphic between themselves: for example, the sets \( \mathcal{P}(\mathbb{N}), \mathcal{P}(\mathcal{P}(\mathbb{N})), \mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N}))), \) etc.
APPENDIX A. SET THEORY

There are many infinite uncountable sets which we are used to, such as \( \mathbb{R} \) or \( \mathbb{C} \). To check that \( \mathbb{R} \) is indeed uncountable, we will use the following proposition which is usually presented as a consequence of the Supremum Axiom.

**Proposition A.4.6.** Every bounded monotone sequence of real numbers converges.

For the proof we refer to Chapter \([\text{I}]\) where the real numbers are introduced in a constructive way.

**Theorem A.4.7.** \( \mathbb{R} \) is an uncountable set.

*Proof.* Let \( \phi : \mathbb{N} \to \mathbb{R} \) be any sequence of real numbers. We need to show that \( \phi \) is not surjective, i.e., that there exists \( x \in \mathbb{R} \) such that \( x \not\in \phi(\mathbb{N}) \).

There exist real numbers \( a_1 \) and \( b_1 \) such that \( a_1 < b_1 < \phi(1) \). In particular, \( \phi(1) \not\in [a_1, b_1] \). One can define recursively sequences \( a_n \) and \( b_n \) which are, respectively, increasing and decreasing, and such that:

\[
a_n < b_n, \quad \phi(n) \not\in [a_n, b_n].
\]

Obviously, both sequences are bounded by \( a_1 \) and \( b_1 \). It follows from Proposition \([\text{A.4.6}]\) that both sequences converge, and we denote their limits by \( a \) and \( b \), respectively. It is also clear that \( a_n \leq a \leq b \leq b_n \), and we conclude immediately that \( a \neq \phi(n) \), for all \( n \in \mathbb{N} \). Hence, \( \phi \) is not surjective. \( \Box \)

**Example A.4.8.**

*Consider the set \( \{0, 1\}^\mathbb{N} \), consisting of all sequences in \( \{0, 1\} \), (called binary sequences) is also uncountable. In this case, it is easy to show that \( \{0, 1\}^\mathbb{N} \) is isomorphic to \( \mathcal{P}(\mathbb{N}) \). For that, observe that a binary sequence \( \phi : \mathbb{N} \to \{0, 1\} \) is completely determined by its support, i.e., by the set of natural numbers \( n \) for which \( \phi(n) \neq 0 \). In other words, the map \( \Psi : \{0, 1\}^\mathbb{N} \to \mathcal{P}(\mathbb{N}) \), defined by \( \Psi(\phi) = \{ n \in \mathbb{N} : \phi(n) \neq 0 \} \), is a bijection.*

One can use the bijection in the last example, together with some elementary facts concerning binary representations of real numbers (base two), to show that both \( \mathcal{P}(\mathbb{N}) \) and \( \{0, 1\}^\mathbb{N} \) are isomorphic to \( \mathbb{R} \) (see the exercises).

We will not define “\#X” when \( X \) is an infinite set. Instead, we will use the symbol “\(|X|\)”, also called cardinal of \( X \), where \( X \) denotes any set (finite or infinite), as part of the expressions “\(|X| = |Y|\)” “\(|X| \leq |Y|\)” and “\(|X| < |Y|\)”, with the following meaning:
A.4. INFINITE SETS

Definition A.4.9. We will write:

(i) \(|X| = |Y|\) if \(X\) and \(Y\) are isomorphic, i.e., if there exists a bijection \(\phi: X \to Y\);

(ii) \(|X| \leq |Y|\) if \(X\) is isomorphic to a subset of \(Y\), i.e., if there exists an injective map \(\phi: X \to Y\);

(iii) \(|X| < |Y|\) if \(|X| \leq |Y|\), and \(X\) is not isomorphic to \(Y\).

The equality “\(|X| = |Y|\)” and the inequality “\(|X| \leq |Y|\)” are analogues of the equalities and inequalities between numbers “\(#(X) = #(Y)\)” and “\(#(X) \leq #(Y)\)”, when \(X\) and \(Y\) are finite sets. For this reason, we could write \(#X = |X|\). Even when \(X\) and \(Y\) are infinite sets, the cardinal has some properties similar to the equalities and inequalities between numbers. For example, the following two properties are immediate to prove:

Proposition A.4.10. Let \(X\), \(Y\) and \(Z\) be sets.

(i) If \(|X| = |Y|\) and \(|Y| = |Z|\), then \(|X| = |Z|\);

(ii) If \(|X| \leq |Y|\) and \(|Y| \leq |Z|\), then \(|X| \leq |Z|\).

It may look obvious that

\(|X| \leq |Y|\) and \(|Y| \leq |X| \iff |X| = |Y|\).

However, the proof of this fact is not at all obvious for infinite sets. In order to prove this we start by showing that the following lemma holds. Note that we already know it holds for finite sets, in which case we can add to the conclusion that “\(X = Y\)” (but not for infinite sets!).

Lemma A.4.11. If \(Y \subset X\) and \(\phi: X \to Y\) is injective, then \(|X| = |Y|\).

Proof. We define recursively two sequences of sets as follows: \(X_1 = X\), \(Y_1 = Y\), and para \(n > 1\), \(X_n = \phi(X_{n-1})\) and \(Y_n = \phi(Y_{n-1})\). We define also \(Z_n = X_n - Y_n\), and

\[Z = \bigcup_{n=1}^{\infty} Z_n.\]

Note that \((X - Z) \subset Y\), since if \(x \in X\) and \(x \notin Y = Y_1\), then \(x \in Z_1\), so \(x \in Z\). Moreover, if \(x \in Z\), i.e., if there exists \(n\) such that \(x \in Z_n\), then \(\phi(x) \in Z_{n+1}\), so that \(\phi(x) \in Z\).
Now, define \( \Psi : X \rightarrow X \) by

\[
\Psi(x) = \begin{cases} 
\phi(x), & \text{if } x \in Z, \\
x, & \text{if } x \not\in Z.
\end{cases}
\]

Notice that \( \Psi(X) \subseteq Y \), since if \( x \in Z \), then \( \Psi(x) = \phi(x) \in Y \), and if \( x \not\in Z \), then \( x \in Y \) and \( \Psi(x) = x \). We claim that \( Y \subseteq \Psi(X) \), so it follows that \( \Psi(X) = Y \). To prove the claim, consider \( x \in Y \) and observe that if \( x \not\in Z \) then \( x = \Psi(x) \in \Psi(X) \). On the other hand, if \( x \in Z \cap Y \), then \( x \in Z_n \) for some \( n > 1 \) (obviously \( Y \) does not contain any element of \( Z_1 \)), and hence \( x \in \phi(Z_{n-1}) \), so that \( x \in \Psi(X) \) (see Figure A.4.1).

Since \( \Psi(X) = Y \), in order to complete the proof, we only need to show that \( \Psi \) is injective. Let \( x, y \in X \) and assume that \( x \neq y \). Then:

(i) If \( x, y \in Z \), then \( \Psi(x) \neq \Psi(y) \), since \( \Psi = \phi \) in \( Z \), and \( \phi \) is injective.

(ii) If \( x, y \not\in Z \), then \( \Psi(x) \neq \Psi(y) \), since \( \Psi \) is the identity in \( X - Z \).

(iii) If \( x \in Z \) and \( y \not\in Z \), then \( \Psi(x) \neq \Psi(y) \), since \( \Psi(y) = y \not\in Z \) and \( \Psi(x) = \phi(x) \in Z \).

We conclude that \( X \) and \( Y \) are isomorphic.

The construction of the set \( Y \) in the previous proof seems ingenious, but it is actually very simple, as we illustrate in the following example.

![Figure A.4.1: The sets \( Z_1, Z_2, Z_3, \ldots \)](image-url)
Example A.4.12.

Let \( X = \mathbb{N}_0 = \{ n \in \mathbb{Z} : n \geq 0 \} \) and \( Y = \mathbb{N} \) and \( \phi(x) = x + 2 \). Then we find immediately that \( X_n = \{ k \in \mathbb{Z} : k \geq 2(n-1) \} \) and \( Y_n = \{ k \in \mathbb{Z} : k \geq 2(n-1) + 1 \} \). It follows that \( Z_n = \{ 2(n-1) : n \in \mathbb{N} \} \), so \( Z \) is the set of even positive integers. The map \( \Psi : \mathbb{N}_0 \to \mathbb{N} \) in the previous proof is then given by:

\[
\Psi(x) = \begin{cases} 
  x + 2, & \text{if } x = 2n, \\
  x, & \text{if } x = 2n + 1.
\end{cases}
\]

This is obviously a bijection from \( \mathbb{N}_0 \) onto \( \mathbb{N} \).

We can now show:

Theorem A.4.13 (Schroeder-Bernstein). If \( |X| \leq |Y| \) and \( |Y| \leq |X| \), then \( |X| = |Y| \).

Proof. Let \( \phi : X \to Y \) and \( \psi : Y \to X \) be injective maps and \( Z = \psi(Y) \). Note that \( \psi \circ \phi : X \to Z \) is injective and since \( Z \subseteq X \), the previous lemma shows that there exists a bijection \( \Psi : Z \to X \). The composition \( \Psi \circ \psi : Y \to X \) is the desired bijection. \( \square \)

The Schroeder-Bernstein theorem is often used to show that two sets are isomorphic, without having to exhibit an explicit bijection between the sets. This is illustrated in the next example.


Consider the sets \( X = \mathbb{N} \times \mathbb{N} \) and \( Y = \mathbb{N} \). The map \( \phi : \mathbb{N} \to \mathbb{N} \times \mathbb{N} \) given by \( \phi(n) = (n, 1) \) is injective. According to the Fundamental Theorem of Arithmetic the map \( \psi : \mathbb{N} \times \mathbb{N} \to \mathbb{N} \) given by \( \psi(n, m) = 2^n 3^m \) is also injective. It follows from the Schroeder-Bernstein theorem that \( \mathbb{N} \) and \( \mathbb{N} \times \mathbb{N} \) are isomorphic.

The following result follows easily from \( |\mathbb{N} \times \mathbb{N}| = |\mathbb{N}| \) and is left as an exercise:

Proposition A.4.15. Let \( X, X_n \) and \( Y \) be countable sets.

(i) If \( Y \neq \emptyset \) and \( X \) is infinite, then \( |X \times Y| = |X| \), i.e., \( X \times Y \) is countable.

(ii) The set \( \bigcup_{n=1}^{\infty} X_n \) is countable.
This proposition shows that the behavior of the notion of cardinal under forming unions and products of infinite sets is slightly peculiar. We state two general results concerning unions and products, illustrating this. We will omit the proofs of these statements, since we do not wish to get even more involved with technical issues in Set Theory (see, however, the exercises at the end of the section).

**Theorem A.4.16.** If $X$ is infinite and $|Y| \leq |X|$, then:

(i) $|X \cup Y| = |X|$;

(ii) $|X \times Y| = |X|$, if $Y \neq \emptyset$.

**Exercises.**

1. Show that $\mathbb{Z}$ and $\mathbb{Q}$ are countable.

2. Show that the intervals $[0,1]$, $]0,1[$, and $[0,1[$ (in $\mathbb{R}$) are all isomorphic to $\mathbb{R}$.

3. Let $X$ be a countable (finite or infinite) set. For each of the following examples determine if the set is countable (finite or infinite), justifying your answer.

   (a) The set $X^{\{0,1\}}$ of all maps $f : \{0,1\} \to X$.

   (b) The set $X^{I_n}$ of all maps $f : I_n \to X$.

   (c) The set $Y = \bigcup_{n=1}^{\infty} X^{I_n}$.

   (d) The set $X^\mathbb{N}$ of all maps $f : \mathbb{N} \to X$.

   (e) The set $\{0,1\}^X$ of all maps $f : X \to \{0,1\}$.

   (f) The set of sequences $f : \mathbb{N} \to X$ which are "eventually constant", i.e., for each sequence there exists $N \in \mathbb{N}$ such that $n > N \Rightarrow f(n) = x \in X$.

   (g) The set $P_{\text{fin}}(X)$ of finite subsets of $X$.

4. Show that the set of polynomials with rational coefficients is countable.

5. Let $X_n$, $n = 1, 2, 3, \ldots$ be countable sets. Show that the sets $\bigcup_{n=1}^{N} X_n$ and $\bigcup_{n=1}^{\infty} X_n$ are also countable.

6. Let $X_n$, $n = 1, 2, 3, \ldots$ be countable sets. Show that $\prod_{n=1}^{N} X_n$ is countable. When is $\prod_{n=1}^{\infty} X_n$ countable?

7. Show that if $X$ is infinite, $Y \subset X$ and $Y$ is finite, then $|X| = |X - Y|$.

8. Show that if $X$ is infinite, $Y \subset X$, $X - Y$ is infinite and $Y$ is countable, then $|X| = |X - Y|$. Conclude that if $X$ is uncountable and $Y$ is countable, then $|X \cup Y| = |X|$.
9. Let \( \{0, 1\}^\mathbb{N} \) be set of all binary sequences. Show that \( \{0, 1\}^\mathbb{N} \) is isomorphic to the interval \([0, 1] \subset \mathbb{R}\), and hence isomorphic to \( \mathbb{R} \).

10. Let \( \{0, 1\}^\mathbb{N} \) be set of all binary sequences. Show that \( \{0, 1\}^\mathbb{N} \times \{0, 1\}^\mathbb{N} \) is isomorphic to \( \{0, 1\}^\mathbb{N} \), and conclude that \( \mathbb{R}^n \) is isomorphic to \( \mathbb{R} \), for any natural number \( n \).

11. Assume that \( |X_n| \leq |\mathbb{R}| \). Show that \( |\bigcup_{n=1}^{\infty} X_n| \leq |\mathbb{R}| \).

12. Consider the sequence of sets \( X_1 = \mathbb{N}, X_{n+1} = \mathcal{P}(X_n) \). Show that there exists a set \( Y \) such that \( |Y| > |X_n| \), for every \( n \in \mathbb{N} \).

13. Let \( X \) be an infinite set and let \( \mathcal{P}_{\text{fin}}(X) \) be set of finite subsets of \( X \). Show that \( |X| = |\mathcal{P}_{\text{fin}}(X)| \).
References (not available yet)