Abstract Algebra

Groups, Rings and Fields, Advanced Group Theory,
Modules and Noetherian Rings, Field Theory

YOTSANAN MEEMARK

Semi-formal based on the graduate courses
2301613–4 Abstract Algebra I & II, offered at
Department of Mathematics and Computer Science,
Faculty of Science, Chulalongkorn University

Published by
Yotsanan Meemark
Department of Mathematics and Computer Science
Faculty of Science, Chulalongkorn University,
Bangkok, 10330 Thailand

First digital edition October 2013
First bound edition May 2014
Second bound edition August 2015

Available for free download at
http://pioneer.netserv.chula.ac.th/~myotsana/

Please cite this book as:
Y. Meemark, Abstract Algebra, 2015, PDF available at
http://pioneer.netserv.chula.ac.th/~myotsana/

Any comment or suggestion, please write to
yotsanan.m@chula.ac.th

Meemark, Yotsanan
Abstract Algebra / Yotsanan Meemark – 2nd ed. Bangkok:
Danex Intercorporation Co., Ltd., 2015. 195pp.
ISBN 978-616-361-389-9

Printed by Danex Intercorporation Co., Ltd., Bangkok, Thailand. www.protexts.com

Foreword

This book is written based on two graduate abstract algebra courses offered at Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University. It grows out of my lecture notes that I used while I was teaching those courses many times. My intention is to develop essential topics in algebra that can be used in research as illustrated some in the final chapter. Also, it can be served as a standard reference for preparing for a qualifying examination in Algebra. I have tried to make it self-contained as much as possible. However, it may not be suitable for reading it for the first course in abstract algebra. It hits and goes through many basic points quickly. A typically mathematical book style that begins with some motivation, definitions, examples and theorems, is used throughout. I try to pause with remarks to make readers have some thoughts before moving on.

The book also requires some background in undergraduate level linear algebra and elementary number theory. For example, I assume the readers to have known matrix theory over a field in which treatment can be found in most linear algebra books. My number theory lecture note is available on the web-page as well. However, some essential results are recalled in the first section. I give many examples to demonstrate new definitions and theorems. In addition, when the converse of a theorem may not hold, counter examples are provided. The major points are divided into five chapters as follows.

- **1 Groups** A *group* is a basic algebraic structure but it is a core in this course. I choose the approach via group actions. Although it is not quite elementary, it is an important aspect in dealing with groups. I also cover Sylow theorems with some applications on finite groups. The structure theorem of finite abelian groups is also presented.
- **2 Rings and Fields** The abstract treatments of rings and fields using groups are presented in the first section. *Rings discussed throughout this book always contain the identity.* Ideals and factorizations are discussed in detail. In addition, I talk about polynomials over a ring and which will be used in a construction of field extensions.
- **3 Advanced Group Theory** In this chapter, I give deeper theory of groups. Various kinds of series of a group are studied in the first three sections. I also have results on a linear group. Finally, I show how to construct a group from a set of objects and presentations and talk about a graphical representation called a *Cayley graph*.
- **4 Modules and Noetherian Rings** Modules can be considered as a generalization of vector spaces. I cover basic concepts of modules and work on free modules. Projective and injective modules are introduced. Moreover, I present the proof of the structure theorems for modules over a PID. Noetherian and Artinion rings are also explored. In the end, I demonstrate some aspects in doing research in algebra. The readers will see some applications of module theory, especially a free R-module over commutative rings, to obtain a structure theorem for finite dimensional symplectic spaces over a local ring. The symplectic graphs over a commutative ring is defined and studied.
- **5 Field Theory** I give more details on a construction of extension fields. Also, I prepare the readers to Galois theory. Applications of Galois theory are provided in proving fundamental theorem of algebra, finite fields, and cyclotomic fields. For the sake of completeness, I discuss some results on a transcendental extension in the final section.

The whole book is designed for a year course. Chapters 1 and 2 are appropriate for a first course and Chapters 3, 4 and 5 can be served as a more advanced course.

There are many topics that, in my opinion, they are worth mentioned. I try to break each topic in step-by-step and scatter it as a "Project" throughout this book. The projects consist of lengthy/generalization exercises, computations of numerical examples, programming suggestions, and research/open problems. This allows us to see that abstract algebra has many applications and is still an active subject. They are independent and can be skipped without any effects on the continuity of the reading.

The book would not have been possible without great lectures from my abstract algebra teachers—Ajchara Harnchoowong and Yupaporn Kemprasit at Chulalongkorn University, and Edward Formanek at the Pennsylvania State University. They initiate wonderful resources to compose each section in this book. I express my gratitude to them all.

I take full responsibility for typos/mistakes that may be found in the manuscript. If you catch ones or have any other suggestions, please write to me. I shall include and correct them in the more up-to-date version once a year on the website. Finally, I hope that the textbook will benefit many students, teachers and researchers in Algebra and Number Theory.

Yotsanan Meemark
Bangkok, Thailand

Contents

This page intentionally left blank

1 | Groups

We write $\mathbb{N}$ for the set of positive integers, $\mathbb{Z}$ for the set of integers, $\mathbb{Q}$ for the set of rational numbers, $\mathbb{R}$ for the set of real numbers and $\mathbb{C}$ for the set of complex numbers.

In this first chapter, we talk about a *group* which is a basic algebraic structure. However, it is a core in this course. Our approach here relies on group actions. Although it is not quite elementary, it is an important aspect in dealing with groups. We also discuss Sylow theorems with some applications and the structure of finite abelian groups.

1.1 Integers

As a number theorist, before I jump into the abstract part, let's lay down some foundations. My first undergraduate abstract algebra course started with elementary number theory—the study of integers. It contains many examples to bear in mind while we are studying the more general results in other abstract domains.

Theorem 1.1.1. [Division Algorithm] *Given integers a and b, with $b \neq 0$, there exist unique integers q and r satisfying*

$$a = qb + r, \qquad \text{where} \quad 0 \leq r < |b|.$$

The integers q and r are called, respectively, the **quotient** *and* **remainder** *in the division of a by b.*

Proof. To prove this theorem, we must use the well-ordering principle, namely, "*every nonempty set S of nonnegative integers contains a least element; that is, there is some integer a in S such that $a \leq b$ for all $b \in S$*".

Existence: First we shall assume that $b > 0$. Let $S = \{a - xb : x \in \mathbb{Z} \text{ and } a - xb \geq 0\} \subseteq \mathbb{N} \cup \{0\}$. We shall show that $S \neq \emptyset$. Since $b \geq 1$, we have $|a|b \geq |a|$, so

$$a - (-|a|)b = a + |a|b \geq a + |a| \geq 0,$$

Then $a - (-|a|)b \in S$, so $S \neq \emptyset$. By the well-ordering principle, S contains a least element, call it r. Then $a - qb = r$ for some $q \in \mathbb{Z}$. Since $r \in S$, $r \geq 0$ and $a = qb + r$. It remains to show that $r < b$. Suppose that $r \geq b$. Thus,

$$0 \leq r - b = a - qb - b = a - (q + 1)b,$$

so $r - b \leq r$ and $r - b \in S$. This contradicts the minimality of r. Hence, $r < b$.

Next, we consider the case in which $b < 0$. Then $|b| > 0$ and Theorem 2.5.2 gives $q', r \in \mathbb{Z}$ such that

$$a = q'|b| + r, \quad \text{where} \quad 0 \leq r < |b|.$$

Since $|b| = -b$, we may take $q = -q'$ to arrive at

$$a = qb + r, \quad \text{where} \quad 0 \leq r < |b|$$

as desired.

Uniqueness: Let $q, q', r, r' \in \mathbb{Z}$ be such that

$$a = qb + r \quad \text{and} \quad a = q'b + r',$$

where $0 \leq r, r' < |b|$. Then

$$(q - q')b = r' - r.$$

Since $0 \leq r, r' < |b|$, we have $|r' - r| < |b|$, so $|b||q - q'| = |r' - r| < |b|$. This implies that $0 \leq |q - q'| < 1$, hence $q = q'$ which also forces $r = r'$. □

Theorem 1.1.1 provides an important example in Section 2.4 where we discuss a more general domain called a Euclidean domain.

An integer b is said to be **divisible** by an integer $a \neq 0$, in symbols $a \mid b$, if there exists some integer c such that $b = ac$. We write $a \nmid b$ to indicate that b is not divisible by a. Note that $a \mid b \Leftrightarrow -a \mid b$, so we may consider only positive divisors. The next theorem contains elementary properties of divisibility.

Theorem 1.1.2. *For integers a, b and c, the following statements hold:*

1. $a \mid 0$, $1 \mid a$, $a \mid a$.
2. $a \mid 1$ *if and only if* $a = \pm 1$.
3. *If* $a \mid b$, *then* $a \mid (-b)$, $(-a) \mid b$ *and* $(-a) \mid (-b)$.
4. *If* $a \mid b$ *and* $c \mid d$, *then* $ac \mid bd$.
5. *If* $a \mid b$ *and* $b \mid c$, *then* $a \mid c$.
6. ($a \mid b$ *and* $b \mid a$) *if and only if* $a = \pm b$.
7. *If* $a \mid b$ *and* $b \neq 0$, *then* $|a| \leq |b|$.
8. *If* $a \mid b$ *and* $a \mid c$, *then* $a \mid (bx + cy)$ *for arbitrary integers* x *and* y.

An integer $p > 1$ is called a **prime number**, or simply a **prime**, if its only positive divisors are 1 and p. An integer greater than 1 which is not a prime is termed **composite**.

Example 1.1.1. $2, 3, 5, 11, 2011$ are primes. $6, 8, 12, 2558$ are composite numbers.

Let a and b be given integers, with at least one of them different from zero. The **greatest common divisor (gcd) of** a **and** b, denoted by $\gcd(a, b)$, is the positive integer d satisfying:

1. $d \mid a$ and $d \mid b$, 2. for all $c \in \mathbb{Z}$, if $c \mid a$ and $c \mid b$, then $c \leq d$.

Basic properties of $\gcd$ are collected in the next theorem.

Theorem 1.1.3. *Let a and n be integers not both zero.*

1. *If* $d = \min\{ax + ny > 0 : x, y \in \mathbb{Z}\}$, *then* $d = \gcd(a, n)$.
2. *If* $\gcd(a, n) = d$, *then* $\exists x, y \in \mathbb{Z}, ax + ny = d$.
3. $\gcd(a, n) = 1 \Leftrightarrow \exists x, y \in \mathbb{Z}, ax + ny = 1$.

Proof. (1) The given set contains $a^2 + n^2$, so it is not empty and d exists by the well-ordering principle. Then $d = ax + ny > 0$ for some $x, y \in \mathbb{Z}$. We shall prove that $d = \gcd(a, n)$. By the division algorithm, $\exists q, r \in \mathbb{Z}, a = dq + r$ with $0 \leq r \leq d$. If $r > 0$, then

$$0 < r = a - dq = a - (ax + ny)q = a(1 - xq) - nyq < d$$

which contradicts the minimality of d. Hence, $r = 0$ and $d \mid n$. Similarly, $d \mid n$. Since $d = ax + ny$, $\gcd(a, n) \mid d$, so $\gcd(a, n) \leq d$. But $d \mid a$ and $d \mid n$, so $d \leq \gcd(a, n)$. Hence, $d = \gcd(a, n)$. (2) follows from (1) and (3) follows from (2). The converse of (3) is immediate. □

Corollary 1.1.4. *Let a, b and c be integers and p a prime.*

1. *If $a \mid bc$ and $(a, b) = 1$, then $a \mid c$.*
2. *If $p \mid bc$ and $p \nmid b$, then $p \mid c$. More generally, if $a_1, a_2, \ldots, a_k$ are integers such that $p \mid a_1 a_2 \ldots a_k$, then $p \mid a_i$ for some i.*
3. *If $q_1, q_2, \ldots, q_n$ are all primes and $p \mid q_1 q_2 \ldots q_k$, then $p = q_i$ for some i.*

Proof. Since $\gcd(a, b) = 1$, we have $1 = ax + by$ for some $x, y \in \mathbb{Z}$. Then $c = acx + bcy$. Since $a \mid bc$, $a \mid c$. This proves (1). (3) follows from (2) and (2) follows from (1) and the fact that $p \nmid b \Leftrightarrow \gcd(p, b) = 1$. □

In my opinion, the next theorem is the most important and everyday use result in number theory. It also provides a core example when we study factorizations in Section 2.4. Its proof applies the results discussed above.

Theorem 1.1.5. [Fundamental Theorem of Arithmetic] *Every positive integer $m > 1$ can be expressed as a product of primes; this representation is unique, apart from the order in which the factors occur.*

Proof. Expressible: Assume on the contrary that there exists an integer $m > 1$ which is *not* a product of primes. By the well-ordering principle, there is a smallest n_0 such that n_0 is not a product of primes. Then n_0 is composite, so there exist integers $1 < d_1, d_2 < n_0$ such that $n_0 = d_1 d_2$. Since $d_1, d_2 < n_0$, d_1 and d_2 are products of primes, and so is n_0. This gives a contradiction. Hence, every positive integer $m > 1$ can be expressed as a product of primes.

Uniqueness: Assume that

$$m = p_1 p_2 \ldots p_s = q_1 q_2 \ldots q_t,$$

where $1 \leq s \leq t$ and p_i and q_j are prime such that

$$p_1 \leq p_2 \leq \cdots \leq p_s \quad \text{and} \quad q_1 \leq q_2 \leq \cdots \leq q_t.$$

Corollary 1.1.4 (3) tells us that $p_1 = q_k$ for some $k \in \{1, \ldots, t\}$. It makes $p_1 \geq q_1$. Similarly, $q_1 = p_l$ for some $l \in \{1, \ldots, s\}$. Then $q_1 \geq p_1$, so $p_1 = q_1$. Thus,

$$p_2 \ldots p_s = q_2 \ldots q_t.$$

Now, repeat the process to get $p_2 = q_2$, and we obtain

$$p_3 \ldots p_s = q_3 \ldots q_t.$$

Continue in this manner. If $s < t$, we would get

$$1 = q_{s+1} q_{s+2} \ldots q_t,$$

which is impossible. Hence, $s = t$ and

$$p_1 = q_1, p_2 = q_2, \ldots, p_s = q_s$$

as desired. □

Corollary 1.1.6. *Any positive integer $m > 1$ can be written uniquely in a canonical form*

$$m = p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r},$$

where, for $i = 1, 2, \ldots, r$, each k_i is a positive integer and each p_i is a prime, with $p_1 < p_2 < \cdots < p_r$.

To formulate an important example in group theory, we shall discuss about the set of integers modulo a positive integer.

Let n be a fixed positive integer. Two integers a and b are said to be **congruent modulo** n, symbolized by

$$a \equiv b \pmod{n} \quad \text{or} \quad a \equiv b \pmod{n}$$

if n divides the difference $a - b$; that is, provided that $a - b = kn$ for some integer k. The number m is called the **modulus of the congruence**. When $n \nmid (a - b)$, then we say that a is **incongruent to** b **modulo** n and in this case we write $a \not\equiv b \pmod{n}$.

Remark. If $n \mid a$, we may write $a \equiv 0 \pmod{n}$.

The first theorem is immediate.

Theorem 1.1.7. *If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then we have:*
1. *$ax + cy \equiv bx + dy \pmod{n}$ for all integers x and y,*
2. *$ac \equiv bd \pmod{n}$,*
3. *$a^m \equiv b^m \pmod{n}$ for every positive integer m.*

For $a \in \mathbb{Z}$, it follows from the division algorithm that there exist unique $q, r \in \mathbb{Z}$ such that $a = nq + r$, where $0 \leq r < n$. This implies $a \equiv r \pmod{n}$. Thus, we have shown:

Theorem 1.1.8. *For each integer a, there exists a unique integer r, with $0 \leq r < n$, such that $a \equiv r \pmod{n}$.*

In terms of congruence, Theorem 1.1.3 (3) may be restated as follows.

Corollary 1.1.9. [Inverse Modulo n] *Let a and n be integers with n positive. Then $\gcd(a, n) = 1$ if and only if there exists an integer x such that $ax \equiv 1 \pmod{n}$. We call x the* **inverse of** a **modulo** n.

We directly obtain the following corollary from Corollary 1.1.4. It gives a condition for canceling integers modulo n.

Corollary 1.1.10. [Cancellative of Integers Modulo n] *Let a, b, c and n be integers with n positive and let p be a prime.*
1. *If $ac \equiv bc \pmod{n}$ and $\gcd(n, c) = 1$, then $a \equiv b \pmod{n}$.*
2. *If $ac \equiv bc \pmod{p}$ and $p \nmid c$, then $a \equiv b \pmod{p}$.*

Finally, we shall define the set of integers modulo a positive integer n. It is not difficult to see that the congruence modulo n is an equivalence relation on $\mathbb{Z}$. That is, for $a, b, c \in \mathbb{Z}$, we have:
1. [reflexivity] $a \equiv a \pmod{n}$,
2. [symmetry] $a \equiv b \pmod{n}$ implies $b \equiv a \pmod{n}$,
3. [transitivity] $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ imply $b \equiv c \pmod{n}$.

For $a \in \mathbb{Z}$, the equivalence class of a is given by

$$\bar{a} = \{b \in \mathbb{Z} : b \equiv a \pmod{n}\} = \{kn + a : k \in \mathbb{Z}\}$$

It also follows from Theorem 1.1.8 that the set of all equivalence classes is

$$\{\overline{0}, \overline{1}, \ldots, \overline{n-1}\}.$$

This set is called the **set of integers modulo** n. In the next section, this set will be an important example of "groups" and will be denoted by $\mathbb{Z}_n$.

As we have discussed, integers have many interesting properties in a way that we can picture them. Next, we shall travel to the abstract parts of the course. In the first two chapters, especially in Chapter 2, we shall see many similarities and generalizations of the integers under operations $+$ and $\cdot$. The readers should keep this section in mind to avoid getting lost in this subject.

Exercises 1.1. 1. Let $d = \gcd(a, b)$. Prove that:
(a) $\gcd(a/d, b/d) = 1$ (b) $\gcd(a - bq, b) = d$ for all $q \in \mathbb{Z}$.

2. [Euclidean Algorithm] Let a and b be positive integers, with $b \leq a$. Repeatedly applications of the division algorithm to a and b give

$$\begin{aligned}
a &= bq_1 + r_1, && \text{where} \quad 0 < r_1 < b \\
b &= r_1q_2 + r_2, && \text{where} \quad 0 < r_2 < r_1 \\
r_1 &= q_3r_2 + r_3, && \text{where} \quad 0 < r_3 < r_2 \\
&\vdots \\
r_{n-2} &= q_nr_{n-1} + r_n, && \text{where} \quad 0 < r_n < r_{n-1} \\
r_{n-1} &= q_{n+1}r_n.
\end{aligned}$$

Prove that $r_n = \gcd(a, b)$. (*Hint*. Use 1. (b))

3. Let $a, b \in \mathbb{Z}$ and $n \in \mathbb{N}$. If $\gcd(a, n) = 1$, show that there exists an $x \in \mathbb{Z}$ such that $ax \equiv b \pmod{n}$.
4. The **least common multiple (lcm)** of two nonzero integers a and b, denoted by $\operatorname{lcm}(a, b)$, is the positive integer m satisfying: (1) $a \mid m$ and $b \mid m$, (2) if $a \mid c$ and $b \mid c$, with $c > 0$, then $m \leq c$. Prove that $\operatorname{lcm}(a, b)\gcd(a, b) = ab$ for all $a, b \in \mathbb{N}$.
5. Let a and b be two integers greater than 1 factored as

$$a = p_1^{a_1}p_2^{a_2}\dots p_r^{a_r} \quad \text{and} \quad b = p_1^{b_1}p_2^{b_2}\dots p_r^{b_r},$$

where for $i = 1, 2, \dots, r$, each p_i is a prime with $p_1 < p_2 < \cdots < p_r$, each a_i and b_i are nonnegative integers, and each a_i or b_i are positive. Prove that

$$\gcd(a, b) = p_1^{d_1}p_2^{d_2}\dots p_r^{d_r}, \text{ where } d_i = \min\{a_i, b_i\} \quad \text{for all } i = 1, 2, \dots, r$$

and

$$\operatorname{lcm}(a, b) = p_1^{c_1}p_2^{c_2}\dots p_r^{c_r}, \text{ where } c_i = \max\{a_i, b_i\} \quad \text{for all } i = 1, 2, \dots, r.$$

1.2 Groups

In order to study abstract senses of algebra, we shall begin with the definition of a group which occupies a very important seat in this course.

1.2.1 Definitions and Examples

For a nonempty set S, a function $\cdot : S \times S \to S$ is called a **binary operation** and image of (a, b) in $S \times S$ is denoted by $a \cdot b$ and it is said to be the **product of** a **and** b. A **groupoid** is a system $(S, \cdot)$ consisting of a nonempty set S with binary operation $\cdot$ on S. We may write S for $(S, \cdot)$ and ab for $a \cdot b$ where $a, b \in S$ if there is no ambiguity.

Let S be a groupoid. For nonempty subsets A and B of S and $x \in S$, let

$$AB = \{ab : a \in A \text{ and } b \in B\}, \quad xA = \{x\}A \quad \text{and} \quad Ax = A\{x\}.$$

If S satisfies the **associative law**, i.e., $\forall a, b, c \in S, (a\cdot b)\cdot c = a\cdot(b\cdot c)$, we say that S is a **semigroup**. Notice that if S is a semigroup, then any bracketing of $x_1, \dots, x_n$ gives the same product, so we can write $x_1 \cdots x_n$ for this product. In addition, for $a \in S$ and $m \in \mathbb{N}$, we may let $a^m = a \cdots a$ (m copies). A groupoid S is said to be **commutative** if $\forall a, b \in S, ab = ba$.

An element e of a groupoid S is a **two-sided identity** or **identity** if $\forall a \in S, ae = a = ea$. Clearly, S contains at most one identity (if e and e' are identity, then $e = ee' = e$). A **monoid** is a semigroup with (unique) identity. Let S be a monoid with identity e. If a and b in S are such that $ab = e = ba$, then b is called a **two-sided inverse** or **inverse** of a. We have that every element of S has at most one inverse. For, if b and b' are inverses of a, then $ab = e = ba$ and $ab' = e = b'a$, so $b = be = b(ab') = (ba)b' = eb' = b'$.

A **group** is a monoid G such that every element of G has an inverse, and for $a \in G$, let a^{-1} denote the (unique) inverse of a. A commutative group is also called an **abelian** group. The **order** of a group G is the cardinal number $|G|$.

Remark. For a nonempty set G with binary operation on G is a *group* if the following axioms are all satisfied:
(G1) [associativity] $\forall a, b, c \in G, (a \cdot b) \cdot c = a \cdot (b \cdot c)$
(G2) [identity] $\exists e \in G, \forall a \in G, ae = a = ea$
(G3) [inverse] $\forall a \in G, \exists b \in G, ab = e = ba$.

Let G be a group with identity e. For $a \in G$ and $m \in \mathbb{N}$, let $a^0 = e$ and $a^{-m} = (a^{-1})^m$.

Remarks. 1. For a group G, we have:
(a) $e^{-1} = e$ and $\forall a \in G, (a^{-1})^{-1} = a$,
(b) $\forall a \in G, \forall m, n \in \mathbb{Z}, a^m a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$, and
(c) $\forall a, b \in G, (ab)^{-1} = b^{-1}a^{-1}$ because $(ab)(b^{-1}a^{-1}) = e$.
2. In case G is abelian, we may choose to write G additively. This means:
(a) The binary operation is denoted by $+$.
(b) 0 denotes the identity element and $-a$ denotes the inverse of a.
(c) $\forall a \in G, \forall m \in \mathbb{N}, ma = a + \cdots + a$ (m copies).
3. A group G satisfies the **cancellation law**: $\forall a, b, c \in G$, $ab = ac$ (or $ba = ca$) $\Rightarrow b = c$.

Examples 1.2.1 (Examples of groups). 1. $(\mathbb{Z}, -)$ is a groupoid which is not a semigroup; $(\mathbb{N}, +)$ is a semigroup which is not a monoid; $(\mathbb{N}, \cdot)$ is a monoid which is not a group.
2. $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$, $(\mathbb{Q}^*, \cdot)$, $(\mathbb{R}^*, \cdot)$ and $(\mathbb{C}^*, \cdot)$ are infinite abelian groups. Here, A^* denotes the set of nonzero elements of A.
3. Let X be a set and $P(X)$ the power set of X. For subsets A and B of X, we define $A \triangle B = (A \setminus B) \cup (B \setminus A)$. Then $(P(X), \triangle)$ is an abelian group having the empty set as its identity and $A^{-1} = A$ for all $A \in P(X)$. Also, $(P(X), \cap)$ is a commutative monoid with identity X.
4. For $n \in \mathbb{N}$, let $\mathbb{Z}_n = \{\overline{0}, \overline{1}, \ldots, \overline{n-1}\}$ called the **set of integers modulo** n, where $\overline{a} = \{kn + a : k \in \mathbb{Z}\}$ for all $a \in \mathbb{Z}$. Define $+$ and $\cdot$ on $\mathbb{Z}_n$ by
$$\overline{a} + \overline{b} = \overline{a + b} \quad \text{and} \quad \overline{a} \cdot \overline{b} = \overline{a \cdot b} \quad \text{for all } a, b \in \mathbb{Z}.$$
These binary operations are well defined by Theorem 1.1.7. It follows that $(\mathbb{Z}_n, +)$ is an abelian group of order n. Moreover, $(\mathbb{Z}_n, \cdot)$ is a commutative monoid with identity $\overline{1}$.
5. For $n \in \mathbb{N}$ and $n \geq 2$, let $\mathbb{Z}_n^\times = \{\overline{a} : \gcd(a, n) = 1\}$. By Theorem 1.1.3 (3), we have $(\mathbb{Z}_n^\times, \cdot)$ is an abelian group. We write $\phi(n)$ for the order of $\mathbb{Z}_n^\times$. It is the **Euler ϕ-function**. Note that $\mathbb{Z}_n^\times = \mathbb{Z}_n \setminus \{\overline{0}\} \Leftrightarrow n$ is a prime.

Proof. If n is a prime, then $\mathbb{Z}_n^\times = \{\overline{1}, \overline{2}, \ldots, \overline{n-1}\} = \mathbb{Z}_n \setminus \{\overline{0}\}$. Conversely, assume that n is composite. Then $n = bc$ for some $1 < b, c < n$, so $\gcd(b, n)$ and $\gcd(c, n)$ are > 1. Thus, $\overline{b}, \overline{c} \notin \mathbb{Z}_n^\times$. Since $b, c < n$, we have $\overline{b}, \overline{c} \neq \overline{0}$. Hence, $\mathbb{Z}_n^\times \subsetneq \mathbb{Z}_n \setminus \{\overline{0}\}$. □

Remark. We recall some properties of the Euler's ϕ-function as follows.
(a) If p is a prime, then $\phi(p) = p - 1$ and $\phi(p^k) = p^k - p^{k-1}$ for all $k \in \mathbb{N}$.
(b) ϕ is multiplicative, namely, if $\gcd(m, n) = 1$, then $\phi(mn) = \phi(m)\phi(n)$.

6. Write F for any of $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ or other *fields*. Let $M_n(F)$ be the set of $n \times n$ matrices over F and $\mathrm{GL}_n(F)$ the set of matrices over F with nonzero determinants. Then $(M_n(F), +)$ is an abelian group and $\mathrm{GL}_n(F)$ is a group under multiplication which is not abelian if $n > 1$. The later group is called the **general linear group**.

7. For a nonempty set X, a function on X which is 1-1 and onto (a bijection on X) is said to be a **permutation of** X. Let $S(X)$ be the set of all permutations of X. Then under composition, $(S(X), \circ)$ is a group called the **symmetric group on** X; in case $X = \{1, 2, \ldots, n\}$, we write S_n and call S_n the **symmetric group on** n **letters**. It is a group of order $n!$.

Some equivalent definitions of groups are collected in the next theorem.

Theorem 1.2.1. [Criteria for Being a Group] *Let G be a semigroup. Then the following statements are equivalent.*

(i) G *is a group.*
(ii) *(a)* $\exists e \in G\, \forall a \in G, ea = a$ *and (b)* $\forall a \in G\, \exists b \in G, ba = e$.
(iii) *(a)* $\exists e \in G\, \forall a \in G, ae = a$ *and (b)* $\forall a \in G\, \exists b \in G, ab = e$.
(iv) $\forall a, b \in G\, \exists x, y \in G, ax = b$ *and* $ya = b$.
(v) $\forall a \in G, aG = G = Ga$.

Proof. If (i) holds, (ii)–(v) are clearly true. (iv) $\Leftrightarrow$ (v) is obvious.
(ii) $\Rightarrow$ (i). Assume (ii). Let $a \in G$. Then $\exists b \in G, ba = e$, and so $\exists c \in G, cb = e$. Thus,

$$ab = e(ab) = (cb)(ab) = c(ba)b = c(eb) = cb = e.$$

Moreover, $ae = a(ba) = (ab)a = ea = a$.
"(iii) $\Rightarrow$ (i)" is similar to "(ii) $\Rightarrow$ (i)".
(iv) $\Rightarrow$ (iii). Assume (iv). Let $a \in G$. Then $\exists e \in G, ae = a$. Let $b \in G$. Then $\exists c \in G, bc = e$ and $\exists y \in G, ya = b$. Thus, $be = (ya)e = y(ae) = ya = b$. □

Theorem 1.2.2. *If G is a finite cancellative semigroup, then G is a group.*

Proof. We shall show that $\forall a \in G, aG = G = Ga$. Let $a \in G$. Since G is cancellative, $|aG| = |G| = |Ga|$. Clearly, $aG \subseteq G$ and $Ga \subseteq G$. Since G is finite, $aG = G = Ga$. □

Remark. $(\mathbb{N}, +)$ is an infinite cancellative semigroup, but it is not a group.

1.2.2 Subgroups

Sometimes a group contains a nonempty subset that is closed under its operation. In this subsection, we discuss a small group in a bigger one with the same operation.

A nonempty subset H of a group G is said to be a **subgroup** of G if H is a group under the same operation of G and we write $H \leqslant G$. Observe that for $\emptyset \neq H \subseteq G$,

$$H \leqslant G \quad \Leftrightarrow \quad \forall a, b \in H, ab \in H \wedge a^{-1} \in H \quad \Leftrightarrow \quad \forall a, b \in H, ab^{-1} \in H.$$

Moreover, $\{e\}$ and G are always subgroups of G. Theorem 1.2.2 gives the following important fact.

Corollary 1.2.3. *If H is a finite nonempty subset of a group G which is closed under the operation of G, then H is a subgroup of G.*

Proof. Since G is a cancellative semigroup and $\emptyset \neq H \subseteq G$, H is a finite cancellative semigroup. Hence, H is a group by Theorem 1.2.2. □

Next, we investigate the *group of symmetries*. We begin with the following groups.

Examples 1.2.2 (Group of symmetries). 1. *The set of rotations about a point* 0 *in the plane; composition as usual.* If 0 is taken to be the origin, the rotation through an angle θ can be represented analytically as the map

$$(x, y) \mapsto (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

For $\theta = 0$, we get the identity map and the inverse of the rotation through the angle θ is the rotation through $-\theta$. It is called the **rotation group**.

2. *The set of rotations together with the set of reflections in the lines which passes through* 0 *with slope* $\tan\alpha$. The latter are given analytically by

$$(x, y) \mapsto (x\cos 2\alpha + y\sin 2\alpha, x\sin 2\alpha - y\cos 2\alpha) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{pmatrix}.$$

The product of two reflections is a rotation and the product in either order of a reflection and a rotation is a reflection.

3. Consider the *regular* n*-gon* (that is, the n-sided polygon in which the sides are all the same length and are symmetrically placed about a common center) inscribed in the unit circle in the plane, so that one of the vertices is $(1, 0)$. The vertices subtend angles of $0, 2\pi/n$, $4\pi/n, \ldots, 2(n-1)\pi/n$ radians with the positive x-axis. The subset of rotation maps which maps our figure into itself consists of the n rotations through angles of $0, 2\pi/n, 4\pi/n, \ldots, 2(n-1)\pi/n$ radians, respectively. These elements form a subgroup R_n of the rotation group defined in (1).

4. We now consider the set D_n of rotations and reflections which map the regular n-gon, as in (3), into itself. These form a subgroup of the group defined in (2). We shall call the elements of this group the **symmetries** of the regular n-gon. The reflection in the x-axis is one of our symmetries. Multiplying this on the right by the n rotational symmetries we obtain n distinct reflectional symmetries. These give them all, for if we let σ denote the reflection in the x-axis and τ denote any reflectional symmetry then $\sigma\tau$ is one of the n-rotational symmetries $\rho_1, \ldots, \rho_n$, say ρ_i. Since $\sigma^2 = 1, \sigma\tau = \rho_j$ gives $\tau = \sigma\rho_j$ which is one of those we counted. Thus, D_n consists of n rotations and n reflections and its order is $2n$. The group D_n is called the **dihedral group**. Note that $D_n = \{\sigma^i\rho_j : i \in \{0, 1\} \text{ and } j \in \{0, 1, 2, \ldots, n-1\}\}$.

Remark. It is easy to see that the intersection of a family of subgroups of a group G is a subgroup of G. If H and K are subgroups of a group G, then, in general, $H \cup K$ is not a subgroup of G. However, if H and K are subgroups of a group G with $G = H \cup K$, then $H = G$ or $K = G$.

Proof. Assume that there is an $x \in G \setminus H$ and a $y \in G \setminus K$. Since $G = H \cup K$, we have $x \in K$ and $y \in H$. Thus, $xy \notin H$ and $xy \notin K$, a contradiction. □

Let G be a group and A a subset of G. Define $\langle A\rangle$ to be the intersection of all subgroups of G containing A. It is the smallest subgroup of G containing A and is called the **subgroup of** G **generated by** A. The elements of A are called **generators**. Moreover, we have $\langle\emptyset\rangle = \{e\}$ and

$$\langle A\rangle = \{a_1^{n_1} \ldots a_k^{n_k} : a_i \in A \text{ and } n_i \in \mathbb{Z}\} \quad \text{if } A \neq \emptyset.$$

For $a_1, \ldots, a_m \in G$, we write $\langle a_1, \ldots, a_m\rangle$ for $\langle\{a_1, \ldots, a_m\}\rangle$. Then $\forall a \in G, \langle a\rangle = \{a^n : n \in \mathbb{Z}\} = \langle a^{-1}\rangle$ is called the **cyclic subgroup of** G **generated by** a and the **order of** a is $|\langle a\rangle|$ (finite or infinite) and denoted by $|a|$ or $o(a)$. If $G = \langle a\rangle$ for some $a \in G$, then G is said to be the **cyclic group generated by** a.

A subgroup N of G is **normal**, denoted by $N \trianglelefteq G$, if $\forall g \in G \forall x \in N, gxg^{-1} \in N$.

Examples 1.2.3 (Examples of subgroups and normal subgroups). 1. $\{e\}$ and G are normal subgroups of G. They are called the **trivial normal subgroups**.

2. Every subgroup of an abelian group is normal.
3. Let $\mathrm{SL}_n(F)$ be the set of matrices over F with determinant one. Then $\mathrm{SL}_n(F)$ is a normal subgroup of $\mathrm{GL}_n(F)$ because $\det(PQP^{-1}) = \det Q$ for all $P, Q \in \mathrm{GL}_n(F)$.
4. $R_n = \langle \rho_{2\pi/n} \rangle$ and $R_n \trianglelefteq D_n$ because $\sigma \rho_j \sigma = \rho_j^{-1}$ for all j.

Remarks. 1. $N \trianglelefteq G$ if and only if ($N \leq G$ and $\forall g \in G, gNg^{-1} = N$).
2. If N is a subgroup of a group G, then $(\forall g \in G, gNg^{-1} = N) \Leftrightarrow (\forall g \in G, gN = Ng)$.

Let G be a group and X a nonempty subset of G. The **centralizer of** X is the set

$$C_G(X) = \{g \in G : \forall x \in X, gx = xg\}$$

and the **normalizer of** X is the set

$$N_G(X) = \{g \in G : gX = Xg\}.$$

We call $Z(G) = C_G(G) = \{z \in G : \forall x \in G, zx = xz\}$, the **center of** G.

Remarks. 1. The centralizer and normalizer of X are subgroups of G.

Proof. Since $\forall x \in G, ex = xe$, we have $e \in C_G(X)$. Let $g, h \in C_G(X)$. Let $x \in X$. Then $gx = xg$ and $hx = xh$, so $xh^{-1} = h^{-1}x$ and

$$(gh^{-1})x = g(h^{-1}x) = g(xh^{-1}) = (gx)h^{-1} = (xg)h^{-1} = x(gh^{-1}).$$

Thus, $gh^{-1} \in C_G(X)$. We leave $N_G(X)$ as an exercise. □

2. $Z(G) \trianglelefteq G$ and $Z(G) = \bigcap_{x \in G} C_G(\{x\})$.

Proof. By (1), $Z(G)$ is a subgroup of G. To see it is normal, let $g \in G$ and $z \in Z(G)$. Let $x \in G$. Then $zg = gz$ and $zx = xz$, so

$$(gzg^{-1})x = (zgg^{-1})x = zx = xz = x(zgg^{-1}) = x(gzg^{-1}).$$

Hence, $gzg^{-1} \in Z(G)$. Finally, for $z \in G$,

$$z \in Z(G) \Leftrightarrow \forall x \in G, xz = zx \Leftrightarrow \forall x \in G, z \in C_G(\{x\}) \Leftrightarrow z \in \bigcap_{x \in G} C_G(\{x\}).$$

This proves the second statement. □

3. G is abelian $\Leftrightarrow Z(G) = G$. (This is clear.)
4. If $K \leq G$, then $K \trianglelefteq N_G(K)$ and $N_G(K)$ is the largest subgroup of G containing K in which K is normal (this means $\forall H \leq G, K \trianglelefteq H \Rightarrow H \subseteq N_G(K)$).

Proof. If $g \in K$, then $gK = K = Kg$, so $K \leq N_G(K)$. To see K is normal in $N_G(K)$, let $x \in N_G(K)$ and $g \in K$. Since $Kx = xK$, we have $xg \in Kx$, so $xg = kx$ for some $k \in K$. Thus $xgx^{-1} = k \in K$. Next, let H be a subgroup of G such that K is normal in H. Then $\forall h \in H, Kh = hK$ which implies $H \subseteq N_G(K)$. □

1.2.3 Homomorphisms

We now study a function between two groups that is required to preserve group operations.

Let G and H be two groups. A **homomorphism** from G to H is a map $\varphi : G \to H$ which satisfies

$$\varphi(xy) = \varphi(x)\varphi(y) \quad \text{for all } x, y \in G.$$

An injective homomorphism is called a **monomorphism** and a surjective homomorphism is called an **epimorphism**. An **isomorphism** is a homomorphism which is both injective and surjective. We write $G \cong H$ if $\exists \varphi : G \to H, \varphi$ is an isomorphism. An **endomorphism of** G is a homomorphism on G and an **automorphism of** G is an isomorphism of G onto itself.

Examples 1.2.4 (Examples of group homomorphisms). 1. $\varphi_1 : \mathbb{Z} \to \mathbb{Z}_n$ given by $\varphi(a) = \bar{a}$ is a homomorphism.
2. $\varphi_2 : (\mathbb{R}, +) \to (\mathbb{R}^\times, \cdot)$ given by $\varphi(x) = 2^x$ is a homomorphism.
3. $\varphi_3 : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{R}$ given by $\varphi(A) = \det A$ is a homomorphism.

The **kernel** of a homomorphism $\varphi : G \to H$ is given by the set

$$\ker \varphi = \{g \in G : \varphi(g) = e_H\}.$$

Example 1.2.5. $\ker \varphi_1 = n\mathbb{Z}$, $\ker \varphi_2 = \{0\}$ and $\ker \varphi_3 = \mathrm{SL}_n(\mathbb{R})$.

Remarks. Let $\varphi : G \to H$ be a homomorphism of groups.
1. $\varphi(e_G) = e_H$ and $\varphi(a^{-1}) = (\varphi(a))^{-1}$ for all $a \in G$.

   Proof. Note that $e_H\varphi(e_G) = \varphi(e_G) = \varphi(e_G e_G) = \varphi(e_G)\varphi(e_G)$. By cancellation in H, we have $e_H = \varphi(e_G)$. Next, let $a \in G$. Then $\varphi(a)\varphi(a^{-1}) = \varphi(aa^{-1}) = \varphi(e_G) = e_H$. Hence, $\varphi(a^{-1}) = \varphi(a)^{-1}$. □

2. φ is 1-1 $\Leftrightarrow \ker \varphi = \{e_G\}$.

   Proof. Assume that φ is 1-1. By (1), $e_G \in \ker \varphi$. Let $x \in \ker \varphi$. Then $\varphi(x) = e_H = \varphi(e_G)$. Since φ is 1-1, we have $x = e_G$. Conversely, suppose that $\ker \varphi = \{e_G\}$. Let $x, y \in G$ be such that $\varphi(x) = \varphi(y)$. Then $e_H = \varphi(x)^{-1}\varphi(y) = \varphi(x^{-1})\varphi(y) = \varphi(x^{-1}y)$. Thus, $x^{-1}y \in \ker \varphi$, so $x^{-1}y = e_G$. Hence, $x = y$. □

Theorem 1.2.4. *Let φ be a homomorphism of a group G into a group H.*
1. *$\ker \varphi$ is a normal subgroup of G.*
2. *$\mathrm{im}\, \varphi = \{\varphi(g) : g \in G\} = \varphi(G)$ is a subgroup of H.*

Proof. Since $\varphi(e_G) = e_H$, $e_G \in \ker \varphi$ and $e_H \in \mathrm{im}\, \varphi$. Let $x, y \in \ker \varphi$. Then $\varphi(x) = e_H = \varphi(y)$, so

$$\varphi(xy^{-1}) = \varphi(x)\varphi(y^{-1}) = \varphi(x)\varphi(y)^{-1} = e_H e_H^{-1} = e_H.$$

Thus, $xy^{-1} \in \ker \varphi$. Hence, $\ker \varphi$ is a subgroup of G. Next, let $g \in G$ and $x \in \ker \varphi$. Then $\varphi(x) = e_H$ and

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

Thus, $gxg^{-1} \in \ker \varphi$, and so $\ker \varphi$ is normal. Finally, let $y, z \in \mathrm{im}\, \varphi$. Then $\exists x_1, x_2 \in G, \varphi(x_1) = y$ and $\varphi(x_2) = z$. Thus,

$$yz^{-1} = \varphi(x_1)\varphi(x_2)^{-1} = \varphi(x_1)\varphi(x_2^{-1}) = \varphi(x_1 x_2^{-1}).$$

Since $x_1 x_2^{-1} \in G$, $yz^{-1} \in \mathrm{im}\, \varphi$. Hence, $\mathrm{im}\, \varphi$ is a subgroup of H. □

Remark. If $\varphi : G \to H$ is a homomorphism and $H_1 \leq H$, then $\ker \varphi \subseteq \varphi^{-1}(H_1) \leq G$.

The next result is clear. It gives another way to construct a group for the Cartesian product of groups.

Theorem 1.2.5. [Cartesian Product of Groups] *If G_1 and G_2 are groups, then the Cartesian product*

$$G_1 \times G_2 = \{(x, y) : x \in G_1, y \in G_2\}$$

is a group under coordinatewise multiplication $(x_1, y_1)(x_2, y_2) = (x_1x_2, y_1y_2)$ for all $x_1, x_2 \in G_1$ and $y_1, y_2 \in G_2$.

Exercises 1.2. 1. Let G be the set of pairs of real numbers (a, b) with $a \neq 0$ and define:

$$(a, b)(c, d) = (ac, ad + b) \quad \text{and} \quad 1 = (1, 0).$$

Verify that this defines a group.

2. Let $H = \{\sigma \in S_4 : \{\sigma(1), \sigma(2)\} = \{1, 2\} \text{ or } \{\sigma(1), \sigma(2)\} = \{3, 4\}\}$. Prove that H is a subgroup of S_4 and find $|H|$. Is H normal in S_4? Justify your answer.
3. Let G be a semigroup such that $\forall a \in G, \exists b \in G, a = aba$ and $\exists! e \in G, e^2 = e$. Prove that G is a group.
4. Let G be a semigroup such that $\forall a, b \in G, a^2b = b = ba^2$. Prove that G is an abelian group.
5. A certain multiplicative operation on a nonempty set G is associative and allows cancellations on the left, and there exists $a \in G$ such that $x^3 = axa$ for all $x \in G$. Prove that G endowed with this operation is an abelian group.
6. Let G be a group with the following properties:
(i) G has no element of order 2 and (ii) $(xy)^2 = (yx)^2$, for all $x, y \in G$.
Prove that G is abelian. If (i) fails, give an example to support that "G may not be abelian".
7. If H and K are subgroups of a group G, prove that $HK \leq G$ if and only if $HK = KH$.
8. Let $\varphi : G \to \bar{G}$ be a group homomorphism, and let N and $\bar{N}$ be a normal subgroup of G and $\bar{G}$, respectively. Show that $\varphi[N]$ is a normal subgroup of $\operatorname{im}\varphi$ and $\varphi^{-1}[\bar{N}]$ is a normal subgroup of G.
9. Let G be a group with identity e and $\varphi : G \to G$ a function such that

$$\varphi(g_1)\varphi(g_2)\varphi(g_3) = \varphi(h_1)\varphi(h_2)\varphi(h_3)$$

whenever $g_1g_2g_3 = e = h_1h_2h_3$. Prove that there exists an element $a \in G$ such that $\psi(x) = a\varphi(x)$ for all $x \in G$ is a homomorphism.
10. Let D_n be the dihedral group of order $2n$ where $n > 2$. Show that the center of D_n has one or two elements according as n is odd or even.

Project 1 (Quaternions)**.** Consider the eight objects ± 1, $\pm i$, $\pm j$ and $\pm k$ with multiplication rules:

$$\begin{gathered} ij = k \qquad jk = i \qquad ki = j \\ ji = -k \qquad kj = -i \qquad ik = -j \\ i^2 = j^2 = k^2 = -1 \end{gathered}$$

where the minus signs behave as expected and 1 and -1 multiply as expected. (For example, $(-1)j = -j$ and $(-i)(-j) = ij = k$.) Show that these objects form a group containing only one element of order 2. This group is called the **quaternion group** and is denoted by Q_8. [*Hint*. The difficulty is to show that the operation is associative. One may transform the elements and operation into 2×2 matrices and matrix products, respectively.]

Project 2 (Associativity)**.** One of the required properties for $(G, *)$ to be a group is associativity. However, this is the hardest one to check as one can see from the previous item. Consider the set $\{a, b, c\}$ of three distinct elements and the operation $*$ given by

$*$	a	b	c
a	a	b	c
b	b	a	c
c	c	c	a

To check associativity, we must check every possible instance of the equation $(x * y) * z = x * (y * z)$. That means we must think of every possible combination of what x, y, and z could be. After a while, we find that $(b * c) * c = a$ but $b * (c * c) = b$. Hence, the set $\{a, b, c\}$ under the operation $*$ is not associative. Be careful! Many students make the mistake of concluding that a set is associative by checking just a few examples. We cannot do this! To determine whether or not a set is associative, you must check every single combination of 3 elements, unless you have a good general argument for why all combinations will be associative or you have a good reason (such as the existence of an identity element) for limiting the number of cases you must check. (Notice that even when you have the existence of an identity element, you still have to check ALL the cases which do not include the identity element.)

Now, for the set of two elements $\{a, b\}$ the number of different binary operations on this set is 16. However, the number of associative binary operation on that set is only 8. This can be checked by writing all out and counting. Unfortunately, for the set of three elements $\{a, b, c\}$, there are $3^{(3\cdot3)} = 19683$ binary operations. How to know how many associative binary operations there are on a set of three elements? (There are 113 operations.) Create an efficient algorithm to solve this task. How about the set of n elements?

1.3 Group Actions

Let G be a group. For each $g \in G$, define the left multiplication function $\ell_g : G \to G$ by

$$\ell_g(x) = gx \quad \text{for all } x \in G.$$

By the left cancellation on G, ℓ_g is a bijection and so $\ell_g \in S(G)$, the symmetric group on G. It is easy to see that the map $g \mapsto \ell_g$ is an injective group homomorphism from G into $S(G)$. This proves an important result in group theory.

Theorem 1.3.1. [Cayley] *Every group G is isomorphic to a subgroups of $S(X)$ for some set X.*

Allowing an abstract group to behave like a group of permutations, as happened in the proof of Cayley's theorem, is a useful tool. In this section, we talk about how a group acting on a set, called a *group action*. There are many nice results and applications as we shall see in the following sections and chapters.

Let G be a group with identity element e and X a nonempty set. We say that G **acts on** X or X is a G**-set** if there is a mapping $G \times X \to X$ (denoted by $g \cdot x$ or gx) which satisfies:

1. $\forall\, x \in X, e \cdot x = x$ and
2. $\forall g, h \in G, \forall x \in X, g \cdot (h \cdot x) = (gh) \cdot x$.

Assume that a group G act on a set X. Then each $g \in G$ determines a set map $\phi_g : X \to X$ by

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

Moreover, $\forall g \in G, \phi_g$ is a bijection (1-1 and onto). Hence, $\phi_g \in S(X)$, the symmetric group on X. The map $g \overset{\theta}{\mapsto} \phi_g$ defines a group homomorphism from G to $S(X)$ (i.e., $\phi_{gh} = \phi_g \circ \phi_h$ for all $g, h \in G$). Its kernel is the set

$$\{g \in G : gx = x \text{ for all } x \in X\}.$$

1. If $g = e$ is the only element of G such that $gx = x$ for all $x \in X$, then G is said to **act faithfully on** X. In this case, $G \hookrightarrow S(X)$.
2. If $x \in X$, the set $Gx = \{gx : g \in G\}$ is called the **orbit of** x.
3. If $Gx = X$ for some (and hence all) $x \in X$, then G is said to **act transitively on** X.
4. If $Y \subseteq X$, the set $\{g \in G : gY = Y\}$ is called the **stabilizer of** Y, denoted by $\mathrm{Stab}_G Y$. The stabilizer of Y is a subgroup of G.

   Proof. Clearly, $e \in \mathrm{Stab}_G Y$. Let $g, h \in \mathrm{Stab}_G Y$. Then $gY = Y$ and $hY = Y$. Thus, $(gh)Y = g(hY) = gY = Y$ and $g^{-1}Y = g^{-1}(gY) = eY = Y$. □

Examples 1.3.1 (Examples of group actions). 1. If X is a set, $S(X)$ acts naturally on X by $f \cdot x = f(x)$ for all $f \in S(X)$ and $x \in X$. This action is faithfully if $|X| > 1$. In particular, S_n acts on $\{1, 2, \ldots, n\}$. The orbit of each $i \in \{1, 2, \ldots, n\}$ is all of $\{1, 2, \ldots, n\}$, thus S_n acts transitively on $\{1, 2, \ldots, n\}$. If $Y \subseteq \{1, 2, \ldots, n\}$, the stabilizer of Y is isomorphic to $S(Y) \times S(Z) \cong S_k \times S_{n-k}$ where $Z = \{1, 2, \ldots, n\} \setminus Y$. Hence, the stabilizer of $\{n\}$ is isomorphic to S_{n-1}.

2. Let G be any group and let $X = G$, considered as a set. Let G act on X by left multiplication

$$g \cdot x = gx.$$

This action is called the **left regular representation**. It is faithful and transitive.

3. $\mathrm{GL}_n(F)$ acts faithfully on F^n, the set of $n \times 1$ column vectors by left multiplication. The orbit of $\vec{0}$ is itself and $\mathrm{GL}_n(F)$ acts transitively on the nonzero vectors.
4. Let G be any group and let X be any set. Let G act on X by $g \cdot x = x$ for all $g \in G$ and $x \in X$. This is called the **trivial** G **action**. Assuming $g \neq e$ and X has more than one element, this action is not faithful and not transitive. All orbits are singleton and G is the stabilizer of every subset of X.
5. Let G be a group and let $X = G$, considered as a set. Let G act on X by **conjugation**

$$g \cdot x = gxg^{-1}.$$

This action may not be faithful. The center of G acts trivially. The orbit of $x \in G$ is the set of conjugates of x, that is

$$g \cdot X = \{gxg^{-1} : g \in G\},$$

called the **conjugacy class of** x. If $|G| > 1$, then this action is not transitive. The number of orbits of the number of conjugacy classes. If Y is a subset of G, the stabilizer of Y under the action is its normalizer, i.e., $\mathrm{Stab}_G Y = N_G(Y)$.

6. Let G be a group and let H be a subgroup of G. Let H act on G by left multiplication. This action is faithful and the orbit of $x \in G$ is

$$H \cdot x = \{hx : h \in H\} = Hx.$$

The action is not transitive unless $H = G$. Moreover, we can let H act on G by $h \cdot g = gh^{-1}$ for all $h \in H$ and $g \in G$. This action is also faithful and the orbit of $x \in G$ is

$$H \cdot x = \{xh^{-1} : h \in H\} = \{xh : h \in H\} = xH.$$

7. Let $X = \mathbb{C} \cup \{\infty\}$, a set that becomes the **Riemann sphere** in complex analysis. The group $\mathrm{GL}_2(\mathbb{C})$ acts on X by the **linear fractional transformation**

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot z = \frac{az+b}{cz+d},$$

the understanding being the image of ∞ is a/c and the image of $-d/c$ is ∞, just as if we were to pass to a limit in each case.

8. Let $\mathrm{SL}_2(\mathbb{R})$ be the subgroup of real matrices in $\mathrm{GL}_2(\mathbb{R})$ of determinant one, and let $\mathcal{H}$ be the subset of $\mathbb{C} \cup \{\infty\}$ in which $\operatorname{Im} z > 0$, called the **Poincaré upper half plane**. Then $\mathrm{SL}_2(\mathbb{R})$ acts on $\mathcal{H}$ by linear fractional transformations.

Now, we present another way to view the orbits.

Theorem 1.3.2. *Let G be a group and suppose G acts on a nonempty set X. Define a relation $\sim$ on X by*

$$x \sim y \quad \Leftrightarrow \quad \exists g \in G, y = g \cdot x.$$

Then

1. *$\sim$ is an equivalence relation on X.*
2. *The equivalence class of $x \in X$ under $\sim$ is $Gx = \{gx : g \in G\}$, the orbit of x. Thus, X is a disjoint union of orbits under the action of G.*

Proof. It is routine to show that $\sim$ is an equivalence relation on X. The equivalence class of x is

$$[x]_\sim = \{y \in X : x \sim y\} = \{y \in X : \exists g \in G, y = gx\} = \{gx : x \in X\} = Gx$$

and hence $X = \bigcup_{x\in X} Gx$. □

If H is a subgroup of G and $x \in G$, the set

$$Hx = \{hx : h \in H\} \quad \text{and} \quad xH = \{xh : h \in H\}$$

are called a **left coset of H in G** and a **right coset of H in G**, respectively.

From Example 1.3.1 (6), we can let H act on G in two ways. Then Hx [xH] is an orbit of x, and so G is a *disjoint* union of left [right] cosets of H in G. If we choose a subset $\{x_\alpha\}$ of G such that G is the disjoint union of the left cosets Hx_α, then $\{x_\alpha\}$ is called a **right transversal or system of left coset representatives of H in G** and if we choose a subset $\{x_\alpha\}$ of G such that G is the disjoint union of the right cosets $y_\alpha H$, then $\{y_\alpha\}$ is called a **left transversal or system of right coset representatives of H in G**.

Remarks. 1. By Theorem 1.3.2,
   (a) $G = \bigcup_{x\in G} Hx$ [$G = \bigcup_{x\in G} xH$],
   (b) $\forall x, y \in G, Hx = Hy$ or $Hx \cap Hy = \emptyset$ [$\forall x, y \in G, xH = yH$ or $xH \cap yH = \emptyset$],
   (c) $\forall x, y \in G, Hx = Hy \Leftrightarrow xy^{-1} \in H$ [$\forall x, y \in G, xH = yH \Leftrightarrow y^{-1}x \in H$].
2. $\forall a \in G, |H| = |Ha| = |aH|$ by cancellation on H.
3. The map $aH \mapsto Ha^{-1}$ for all $a \in G$ is a 1-1 correspondence between the sets $\{xH : x \in G\}$ and $\{Hx : x \in G\}$.

   Proof. For $a, b \in G$, $aH = bH \Leftrightarrow b^{-1}a \in H \Leftrightarrow a^{-1}b \in H \Leftrightarrow Ha^{-1} = Hb^{-1}$. Then this map is 1-1, well defined and clearly onto. □

The **index of H in G**, denoted by $[G : H]$, is the cardinal number of distinct right (or left) cosets of H in G, that is,

$$[G : H] = |\{Hx : x \in G\}| = |\{xH : x \in G\}|.$$

Next, we show that a subgroup of index two is always a normal subgroup.

Theorem 1.3.3. *If H is a subgroup of G of index two, then H is normal in G.*

Proof. Since $[G : H] = 2$, G has exactly two right (or left) cosets. Then $Hg = G \setminus H$ and $gH = G \setminus H$ for all $g \in G$ not in H. Hence, $\forall g \in G, Hg = gH$, so H is normal in G. □

Let I and A be sets. Define $A_i = \{(i, a) : a \in A\}$ for all $i \in I$. Then $|A_i| = |A|$ for all $i \in I$, $A_i \cap A_j = \emptyset$ if $i \neq j$, and

$$\left|\bigcup_{i\in I} A_i\right| = \sum_{i\in I} |A_i| = \sum_{i\in I} |A| = |I||A|.$$

Lagrange observed an important property of subgroups of G, namely, its order must be a divisor of the order of G.

Theorem 1.3.4. [Lagrange] *If H is a subgroup of G, then $|G| = [G : H]|H|$. In particular, if G is finite and $H < G$, then $|H|$ divides $|G|$, and so $|a| = |\langle a\rangle|$ divides $|G|$ for all $a \in G$.*

Proof. Since G is a disjoint union of distinct left cosets Hx_α, $\alpha \in \Lambda$, and $|\Lambda| = [G : H]$,

$$|G| = \left| \bigcup_{\alpha \in \Lambda} Hx_\alpha \right| = \sum_{\alpha \in \Lambda} |Hx_\alpha| = \sum_{\alpha \in \Lambda} |H| = |\Lambda||H| = [G : H]|H|.$$

If G is finite, then $|H|$ divides $|G|$. In addition, $\forall a \in G, \langle a \rangle \leq G$, so $|a| = |\langle a \rangle|$ divides $|G|$ for all $a \in G$. □

Corollary 1.3.5. *If G is a group of prime order, then $\{e\}$ and G are the only two subgroups of G and G must be cyclic.*

Proof. Let $H \leq G$. Then $|H|$ divides $|G| = p$, so $|H| = 1$ or $|H| = p$. Thus, $H = \{e\}$ and $H = G$. Also, if $a \neq e$, then $\langle a \rangle \neq \{e\}$. Hence, $\langle a \rangle = G$ and so G is cyclic. □

A relationship between the stabilizer of x in a group G and the number of elements in the orbit $G \cdot x$ is recorded in the next theorem.

Theorem 1.3.6. [Orbit-Stabilizer Theorem] *Let a group G act on a set X and suppose $x \in X$. Then $[G : \operatorname{Stab}_G x] = |G \cdot x|$, that is, the index of the stabilizer of x in G is the number of elements in the orbit of x.*

Proof. Let $x \in X$. Note that for all $g_1, g_2 \in G$,

$$g_1 x = g_2 x \Leftrightarrow (g_2^{-1} g_1)x = x \Leftrightarrow g_2^{-1} g_1 \in \operatorname{Stab}_G\{x\} \Leftrightarrow g_1 \operatorname{Stab}_G\{x\} = g_2 \operatorname{Stab}_G\{x\}.$$

Then $|\{gx : g \in G\}| = |\{g \operatorname{Stab}_G x : g \in G\}|$. Hence, $|G \cdot x| = [G : \operatorname{Stab}_G x]$. □

This theorem is most useful when this index is finite but it is true in general. We see some applications of this theorem in the following results.

Theorem 1.3.7. *Let G be a group and $x \in G$. Then the following statements hold.*

1. *$|\{gxg^{-1} : g \in G\}| = [G : C_G(x)]$, i.e., the number of conjugates of x is $[G : C_G(x)]$.*
2. *If G is finite, then the number of conjugates of x is a divisor of $|G|$.*

Proof. It follows directly from the Orbit-Stabilizer Theorem if we consider the action of G on G by conjugation. □

Observe that for each $x \in G$,

$$|\{gxg^{-1} : g \in G\}| = 1 \Leftrightarrow \{gxg^{-1} : g \in G\} = \{x\} \Leftrightarrow \forall g \in G, gx = xg \Leftrightarrow x \in Z(G).$$

Corollary 1.3.8. [Class Equation] *Let G be a finite group and let $x_1, \ldots, x_s$ represent the conjugacy classes of G which contains more than one element. Then*

$$|G| = |Z(G)| + \sum_{i=1}^{s} [G : C_G(x_i)].$$

Now, let G act on the set of all subsets of G by conjugation, i.e., if $Y \subset G$, then $g \cdot Y = gYg^{-1}$. Under this action the stabilizer of Y is $\{g \in G : gYg^{-1} = Y\} = N_G(Y)$, the normalizer of Y, and the orbit of Y is $\{gYg^{-1} : g \in G\}$, the set of conjugates of Y. Thus, the number of conjugates of Y is the index of the normalizer of Y. Hence, we have shown:

Theorem 1.3.9. *Let G be a finite group and Y a subset of G. Then the number of conjugates of Y is $[G : N_G(Y)]$ where $N_G(Y)$ is the normalizer of Y. In particular, the number of conjugates of Y divides the group order.*

Remark. If H is a subgroup of G, then $H \triangleleft N_G(H) < G$. Hence, if G is finite, then the number of conjugates of H is

$$[G : N_G(H)] = \frac{[G : H]}{[N_G(H) : H]} \leq [G : H].$$

Burnside's theorem gives the number of orbits in X under the action of a finite group G.

Theorem 1.3.10. [Burnside] *Let a finite group G act on a finite set X. For each $g \in G$, let $X_g = \{x \in X : gx = x\}$, the set of points in X fixed by g. Then the number of orbits in X is*

$$N = \frac{1}{|G|} \sum_{g \in G} |X_g|.$$

Proof. Let U be the subset of $G \times X$ defined by

$$U = \{(g, x) \in G \times X : gx = x\}.$$

For $h \in G$, let

$$U(h) = U \cap (\{h\} \times X) = \{(h, x) : x \in X \text{ and } hx = x\}.$$

For $y \in X$, let

$$U[y] = U \cap (G \times \{y\}) = \{(g, y) : g \in G \text{ and } gy = y\} = (\operatorname{Stab}_G\{y\}) \times \{y\}.$$

Now, $U = \bigcup_{g \in G} U(g) = \bigcup_{x \in X} U[x]$ and these unions are "disjoint". Note that for each $g \in G$, $|U(g)| = |X_g|$ and for each $x \in X$, $|U[x]| = |\operatorname{Stab}_G\{x\}| = [G : G \cdot x] = |G|/|G \cdot x|$. Thus,

$$\sum_{g \in G} |X_g| = \sum_{g \in G} |U(g)| = |U| = \sum_{x \in X} |U[x]| = \sum_{x \in X} \frac{|G|}{|G \cdot x|} = |G| \sum_{x \in X} \frac{1}{|G \cdot x|} = |G|N$$

as desired. □

Corollary 1.3.11. *Let a finite group G act transitively on a finite set X. Then $|G| = \sum_{g \in G} |X_g|$. Moreover, if $|X| > 1$, then there exists a $g \in G$ fixing no point of X.*

Proof. Since G acts transitively on X, $N = 1$, and so $|G| = \sum_{g \in G} |X_g|$. Assume that $|X| > 1$ and no $g \in G$ fixing no point of X. Then $\forall g \in G, \exists x \in X, gx = x$ which implies that $|X_g| \geq 1$ for all $g \in G$. Thus,

$$|G| \leq \sum_{g \in G} |X_g| = |G|.$$

This forces that $|X_g| = 1$ for all $g \in G$. But $|X_e| = |X| > 1$, a contradiction. Hence, there exists a $g \in G$ fixing no point of X. □

We have known from Lagrange's theorem that the order of any subgroups of a group G is a divisor of $|G|$. The next theorem implies that if $|G|$ has a prime divisor p, then G has a subgroup of order p. Its proof is another application of group actions.

Theorem 1.3.12. [Cauchy] *Suppose G is a finite group and a prime p divides $|G|$. Then the number of solutions of $g^p = e$ in G is a multiple of p. Hence, G contains an element of order p. In particular, if G is a finite group and a prime p divides $|G|$, then G has a subgroup of order p.*

Proof. Consider the set $Y = G \times G \times \cdots \times G$ (p copies) and let $X = \{(g_1, g_2, \ldots, g_p) \in Y : g_1 g_2 \cdots g_p = e\}$. Then $|X| = |G|^{p-1}$ since $g_p = (g_1 g_2 \ldots g_{p-1})^{-1}$. Let $R_p = \langle \rho_{2\pi/p} \rangle$ act on X by

$$\rho_{2\pi/p}(g_1, g_2, \ldots, g_p) = (g_2, g_3, \ldots, g_p, g_1).$$

Note that the orbit of $(g_1, g_2, \ldots, g_p) \in X$ under this action has either
(a) length p, in case $g_1, g_2, \ldots, g_p$ are not all equal, or
(b) length 1, in case $g_1 = g_2 = \cdots = g_p = g$, i.e., $g^p = e$.
Thus, # of orbits of length 1 = # of solutions of $g^p = e$ in G. By Theorem 1.3.2 (2),

$$|G|^{p-1} = |X| = p(\#\text{ of orbits of length } p) + 1(\#\text{ of orbits of length } 1).$$

Since p divides $|G|$, it follows that $|\{g \in G : g^p = e\}| > 1$ is a multiple of p. □

Exercises 1.3.

1. Let G act on S, H act on T and assume $S \cap T = \emptyset$. Let $U = S \cup T$ and define $(g, h)s = gs$ and $(g, h)t = ht$ for all $g \in G, h \in H, s \in S, t \in T$. Show that this gives an action of the group $G \times H$ on U.
2. Let H and K be subgroups of a group G.
   (a) If H and K are finite, then HK is a finite *set* and $|HK| = \dfrac{|H||K|}{|H \cap K|}$.
   (b) For x and y in G, prove that $xH \cap yK$ is empty or is a coset of $H \cap K$.
   (c) Deduce from (b) that if H and K have finite index in G, then so does $H \cap K$.
   (d) If $[G : H]$ and $[G : K]$ are finite and relatively prime, prove that $G = HK$.
3. Let α be an automorphism of a finite group G which leaves only the identity fixed. Prove that $G = \{x^{-1}\alpha(x) | x \in G\}$.
4. Let a group G act on a set X transitively. Prove that
   (a) $\forall x, y \in X, \exists g \in G, gx = y$, and
   (b) $\forall x, y \in X, \exists g \in G, gG_x g^{-1} = G_y$, i.e., all stabilizers are conjugate.
5. Let H be a subgroup of a group G and $N = \bigcap_{x \in G} xHx^{-1}$. Prove that
   (a) N is a normal subgroup of G, and (b) if $[G : H]$ is finite, then $[G : N]$ is finite.
6. Determine the number of conjugacy classes in a non-abelian group G of order p^3 where p is a prime.
7. Let S and T be sets and let $M(S, T)$ denote the set of all functions of S into T. Let G be a finite group acting on S. For each map $f : S \to T$ and $x \in G$ define the map $\pi_x f : S \to T$ by $(\pi_x f)(s) = f(x^{-1}s)$.
   (a) Prove that $x \mapsto \pi_x$ is an action of G on $M(S, T)$.
   (b) Assume that S and T are finite. Let $n(x)$ denote the number of orbits of the cyclic group $\langle x \rangle$ on S. Prove that the number of orbits of G in $M(S, T)$ is equal to
   $$\frac{1}{|G|} \sum_{x \in G} |T|^{n(x)}.$$
8. Two actions of a group G on sets X and Y are called **equivalent** if there is a bijection $f : X \to Y$ such that $f(gx) = gf(x)$ for all $g \in G$ and $x \in X$. Let H and K be subgroups of a group G. Let G act by left multiplication on the sets of left cosets G/H and G/K. Prove that these actions are equivalent if and only if H and K are conjugate (i.e., $K = aHa^{-1}$ for some $a \in G$).

Project 3 (Semi-direct product)**.** A group H is said to *act on a group K by automorphisms* if we have an action of H on K and for every $h \in H$ the map $k \mapsto hk$ of K is an automorphism. Suppose this is the case and let G be the product set $K \times H$. Define a binary operation in $K \times H$ by

$$(k_1, h_1)(k_2, h_2) = (k_1(h_1 k_2), h_1 h_2)$$

and define $1 = (1, 1)$ – the units of K and H, respectively. Verify that this defines a group such that $h \mapsto (1, h)$ is a monomorphism of H into $K \times H$ and $k \mapsto (k, 1)$ is a monomorphism of K into $K \times H$ whose image is a normal subgroup. This group is called a **semi-direct product of K by H** and is denoted by $K \rtimes H$.

1.4 Quotient Groups and Cyclic Groups

This section contains a construction of a new group using a normal subgroup, called a *quotient group*. We also work on an important kind of subgroups of G which are generated by a single element. We conclude this section by studying the group of automorphisms of G.

1.4.1 Quotient Groups

Suppose G is any group and N is a normal subgroup of G. Then for any $g \in G$,

$$N = gNg^{-1} \quad \text{or} \quad Ng = gN.$$

In other words, every left coset of N in G is also a right coset of N in G. If we have two left cosets of N in G;

$$Nx = \{ax : a \in N\} \quad \text{and} \quad Ny = \{by : b \in N\},$$

then $NxNy = \{axby : a, b \in N\} = N(xN)y = N(Nx)y = Nxy$ is again a left coset of N in G. Thus

$$NxNy = Nxy$$

defines a binary operation on the set $G/N = \{Nx : x \in G\}$ of left cosets of N in G.

Theorem 1.4.1. [Quotient Groups] *Suppose G is a group and N is a normal subgroup of G. Let $G = \bigcup Nx_\alpha$ be a decomposition of G as a disjoint union of left (or right) cosets. Then the binary operation*

$$Nx_\alpha Nx_\beta = Nx_\alpha x_\beta$$

makes the set of left cosets of N into a group, called the **quotient or factor group of G by the normal subgroup N.** *This group is denoted by G/N. The map $\pi : G \to G/N$ defined by*

$$\pi(x) = Nx$$

is a group homomorphism whose kernel is N, called the **canonical projection**.

We have the following observations on the above construction:

1. If H is a subgroup of G which is *not* normal, then the set of left cosets of H in G does not form a group in any natural way. For example, if $G = S_3$ and $H = \langle(12)\rangle = \{(1), (12)\}$, then H is not normal in G and $\{H, H(13), H(23)\}$ is not a group because

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

   which is not one of the cosets.
2. $|G/N| = [G : N]$, the index of N in G.
3. If G is abelian written additively, then

   $$G/N = \{N + x : x \in G\}$$

   and the binary operation on G/N is given by

   $$(N + x) + (N + y) = N + (x + y)$$

   for all $x, y \in G$.

We now present three group isomorphism theorems.

Theorem 1.4.2. [First Isomorphism Theorem] *Suppose $\varphi : G \to H$ is a group homomorphism. Then $G/(\ker \varphi) \cong \operatorname{im} \varphi$.*

Proof. By Theorem 1.2.4, $\operatorname{im}\varphi$ is a subgroup of H and $\ker\varphi$ is a normal subgroup of G, and so $G/(\ker\varphi)$ is a group. Let $\overline{\varphi} : G/\ker\varphi \to \operatorname{im}\varphi$ by $\overline{\varphi} : x(\ker\varphi) \mapsto \varphi(x)$. Then for all $x, y \in G$,

$$\overline{\varphi}(x(\ker\varphi)) = \overline{\varphi}(y(\ker\varphi)) \Leftrightarrow \varphi(x) = \varphi(y) \Leftrightarrow xy^{-1} \in \ker\varphi \Leftrightarrow x(\ker\varphi) = y(\ker\varphi),$$

so $\overline{\varphi}$ is well defined and 1-1. In addition, for all $x, y \in G$,

$$\overline{\varphi}(x(\ker\varphi)y(\ker\varphi)) = \overline{\varphi}(xy(\ker\varphi)) = \varphi(xy) = \varphi(x)\varphi(y) = \overline{\varphi}(x(\ker\varphi))\overline{\varphi}(y(\ker\varphi)).$$

Moreover, $\overline{\varphi}$ is clearly onto. Hence, φ is an isomorphism. □

Examples 1.4.1. $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$, $D_n/R_n \cong \mathbb{Z}_2$, $\mathbb{R}/\mathbb{Z} \cong \mathbf{S}^1$ and $\mathrm{GL}_n(F)/\mathrm{SL}_n(F) \cong F \setminus \{0\}$.

Let $G \xrightarrow{\theta} H \xrightarrow{\varphi} K$ be a sequence of group homomorphisms. We say that it is **exact at** H if $\operatorname{im}\theta = \ker\varphi$. A **short exact sequence of groups** is a sequence of groups and homomorphisms

$$1 \longrightarrow G \xrightarrow{\theta} H \xrightarrow{\varphi} K \longrightarrow 1$$

which is exact at G, H and K. That is, θ is injective, φ is surjective and $\operatorname{im}\theta = \ker\varphi$. Here, 1 stands for the smallest group of order one.

Remark. If N is a normal subgroup of G, then

$$1 \longrightarrow N \xrightarrow{\iota} G \xrightarrow{\pi} G/N \longrightarrow 1$$

is exact. Here ι denotes the inclusion map. On the other hand, if $N \leq G$ and

$$1 \longrightarrow N \xrightarrow{\iota} G \longrightarrow H \longrightarrow 1$$

is exact, then N is normal in G and $H \cong G/N$. *Thus short exact sequences are just another notation for normal subgroups and factor groups.*

Theorem 1.4.3. [Second Isomorphism Theorem] *Suppose G is a group and H and N are subgroups of G, with N normal in G. Then $HN = NH$ is a subgroup of G in which N is normal, $H \cap N$ is normal in H and $H/(H \cap N) \cong HN/N$.*

Proof. Since N is normal in G, $hN = Nh$ for all $h \in H$, so $HN \subseteq NH$ and $NH \subseteq HN$. Thus, $NH = HN$. It is routine to show that NH is a subgroup of G. Since $N \trianglelefteq G$, $N \trianglelefteq NH$. The theorem follows from exactness of the sequence

$$1 \longrightarrow H \cap N \xrightarrow{\iota} H \xrightarrow{\varphi} HN/N \longrightarrow 1\,,$$

where the homomorphism $\varphi : h \mapsto hN$ for all $h \in H$ and $\ker\varphi = \{h \in H : hN = N\} = \{h \in H : h \in N\} = H \cap N$. □

Remark. If H and N are not normal in G, then HN may not be a subgroup of G. E.g., in S_3, the subgroups $H = \{(1), (12)\}$ and $N = \{(1), (13)\}$ are not normal in S_3 and $HN = \{(1), (12), (13), (132)\}$ is not a subgroup of S_3.

Theorem 1.4.4. [Third Isomorphism Theorem] *Suppose G is a group and N is a normal subgroup of G. Then the map*

$$\theta : H \mapsto H/N$$

gives a 1-1 correspondence

$$\{\text{ subgroups of } G \text{ containing } N\} \longleftrightarrow \{\text{ subgroups of } G/N\}.$$

This correspondence carries normal subgroups to normal subgroups. Moreover, if H is normal in G containing a subgroup N, then

$$G/H \cong (G/N)/(H/N).$$

Proof. Let H_1 and H_2 be subgroups of G containing N and assume that $H_1/N = H_2/N$. Let $x \in H_1$. Then $Nx \in H_1/N = H_2/N$, so $Nx = Ny$ for some $y \in H_2$. Thus, $xy^{-1} \in N \subseteq H_2$. Since $y \in H_2$, $x \in H_2$. Hence, $H_1 \subseteq H_2$. By symmetry, $H_2 \subseteq H_1$. Therefore, $H_1 = H_2$ and θ is 1-1. Next, let $\mathcal{H} \leq G/N$. Then $\{N\} \subseteq \mathcal{H}$. Choose $H = \bigcup \mathcal{H}$, the union of cosets in $\mathcal{H}$. Thus, $N \subseteq H$. Let $x, y \in H$. Then $Nx, Ny \in \mathcal{H}$, so $Nxy^{-1} \in \mathcal{H}$ which implies $xy^{-1} \in H$. Thus, H is a subgroup of G containing N and $\mathcal{H} = H/N$. Hence, θ is onto.

Assume that H is a normal subgroup of G containing N. Let $g \in G$ and $x \in H$. Then $gxg^{-1} \in H$, so $gNxNg^{-1}N = gxg^{-1}N \in H/N$. Hence, H/N is normal subgroup of G/N.

The final isomorphism follows from exactness of the sequence

$$1 \longrightarrow H/N \xrightarrow{\iota} G/N \xrightarrow{\varphi} G/H \longrightarrow 1 \ ,$$

where the homomorphism $\varphi : gN \mapsto gH$ for all $g \in G$ which is well defined because $N \subseteq H$, and $\ker \varphi = \{gN : g \in G \text{ and } gH = H\} = \{gN : g \in H\} = H/N$. □

Remark. The above theorem is useful for obtaining subgroups and normal subgroups of a quotient group. It plays an important role in the study of normal series of groups and nilpotent groups as we shall see in Chapter 3.

1.4.2 Cyclic Groups

Recall that a cylic subgroup of G is a subgroup of G generated by a singleton. It has a simple structure and is easy to construct. Its properties depend mostly on the group of integer modulo n. We shall go deep inside the groups $\mathbb{Z}_n$ and $\mathbb{Z}_n^\times$ in this section. We recall from Examples 1.2.1 that:

Theorem 1.4.5. *Let $n \geq 2$ and $\mathbb{Z}_n^\times = \{\bar{a} : \gcd(a, n) = 1\}$. Then $(\mathbb{Z}_n^\times, \cdot)$ is an abelian group of order $\phi(n)$, the* **Euler ϕ-function**.

Example 1.4.2. $\mathbb{Z}_{10}^\times = \{\bar{1}, \bar{3}, \bar{7}, \bar{9}\}$ and $\mathbb{Z}_p^\times = \{\bar{1}, \bar{2}, \ldots, \overline{p-1}\}$ where p is a prime.

Now we study cyclic subgroups of a group G. Recall that if G is a group and $a \in G$, then $\langle a \rangle = \{a^m : m \in \mathbb{Z}\}$ and the order of a is $|a| = |\langle a \rangle|$.

Theorem 1.4.6. *Let G be a group and $a \in G$. Then*

1. $\forall n \in \mathbb{N}, a^n = e \Rightarrow \langle a \rangle = \{e, a, a^2, \ldots, a^{n-1}\}$.
2. *$|a|$ is finite* $\Leftrightarrow \exists i, j \in \mathbb{Z}(i \neq j \wedge a^i = a^j) \Leftrightarrow \exists n \in \mathbb{N}, a^n = e$.
3. *$|a|$ is infinite* $\Leftrightarrow \forall i, j \in \mathbb{Z}(i \neq j \rightarrow a^i \neq a^j) \Leftrightarrow \langle a \rangle \cong \mathbb{Z}$.
4. *If $G = \langle a \rangle$ is infinite, then a and a^{-1} are only two generators of G.*
5. *If G is finite, then $|a| = \min\{n \in \mathbb{N} : a^n = e\}$ and $\langle a \rangle = \{e, a, a^2, \ldots, a^{|a|-1}\} \cong \mathbb{Z}_{|a|}$.*
6. $\forall n \in \mathbb{Z}$, $a^n = e \Rightarrow |a|$ *divides* n.
7. *If G is finite, then $a^{|G|} = e$.*

Proof. (1)–(3) are clear.
(4) Assume that $G = \langle a \rangle$ is infinite and a^m is a generator of G for some $m \in \mathbb{Z}$. Then $\langle a^m \rangle = \langle a \rangle$, so $a = (a^m)^k$ for some $k \in \mathbb{Z}$. Since $|a|$ is infinite, $mk = 1$. Thus, $m \mid 1$, so $m = \pm 1$.
(5) Assume that G is finite. Then $a^n = e$ for some $n \in \mathbb{N}$. Choose n_0 to be the smallest such n. Thus, $a^{n_0} = e$. We shall show that $\langle a \rangle = \{e, a, a^2, \ldots, a^{n_0-1}\}$. Clearly, $\{e, a, a^2, \ldots, a^{n_0-1}\} \subseteq \langle a \rangle$. Let $j \in \mathbb{Z}$. Then $j = n_0 q + r$ for some $q, r \in \mathbb{Z}$ and $0 \leq r < n_0$, so $a^j = a^{n_0 q + r} = a^r \in \{e, a, a^2, \ldots, a^{n_0-1}\}$. Hence, $|a| = n_0 = \min\{n \in \mathbb{N} : a^n = e\}$ and $a^{|a|} = e$. Finally, an isomorphism is given by $a^j \mapsto \bar{j}$ for all $j \in \mathbb{Z}$.
(6) Let $n \in \mathbb{Z}$ and $a^n = e$. By the division algorithm, $n = |a|q + r$ for some $q, r \in \mathbb{Z}$ and $0 \leq r < |a|$.

If $r > 0$, then $e = a^n = a^{|a|q+r} = a^r$ which contradicts the minimality of $|a|$. Hence, $r = 0$ and so $|a|$ divides n.
(7) By Lagrange, $|a|$ divides $|G|$, so $|G| = |a|q$ for some $q \in \mathbb{Z}$. Then $a^{|G|} = a^{|a|q} = e$. □

The above results for the group $\mathbb{Z}_n^\times$ yield famous results below.

Corollary 1.4.7. *1.* [Euler] *If $a \in \mathbb{Z}$ and $\gcd(a, n) = 1$, then $a^{\phi(n)} \equiv 1 \pmod{n}$.*
2. [Fermat] *If p is a prime, then $a^p \equiv a \pmod{p}$ for all $a \in \mathbb{Z}$.*

Proof. (1) Apply the above theorem to $G = \mathbb{Z}_n^\times$.
(2) If $\gcd(a, p) = 1$, then by (1), $a^{p-1} \equiv 1 \pmod{p}$, so $a^p \equiv a \pmod{p}$. If $\gcd(a, p) > 1$, then $p \mid a$, so $p \mid (a^p - a)$. Hence, $a^p \equiv a \pmod{p}$ for all $a \in \mathbb{Z}$. □

Theorem 1.4.8. *Any two cyclic groups of the same orders (finite or infinite) are isomorphic.*

Proof. Assume that G is cyclic. Then $G = \langle a \rangle$ for some $a \in G$. By Theorem 1.4.6, if G is infinite, then $G \cong \mathbb{Z}$, and if G is finite, then $|G| = |a|$ and $G = \{e, a, \ldots, a^{|a|-1}\} \cong \mathbb{Z}_{|a|}$. □

Next, we study subgroups of a cyclic group.

Theorem 1.4.9. [Subgroups of a Cyclic Group] *Let G be a cyclic group generated by a, and let H be a subgroup of G. Then H is also a cyclic group generated by a^k where $k = \min\{m \in \mathbb{N} : a^m \in H\}$ or $H = \{e\}$. Consequently, every subgroup of a cyclic group is cyclic.*

Proof. Since $a^k \in H$, $\langle a^k \rangle \subseteq H$. Let $x \in H$. Then $x \in G$, so $x = a^t$ for some $t \in \mathbb{Z}$. By the division algorithm, $t = kq + r$ for some $q, r \in \mathbb{Z}$ and $0 \leq r < k$. Thus,

$$a^r = a^{t-kq} = a^t a^{-kq} = x(a^k)^{-q} \in H.$$

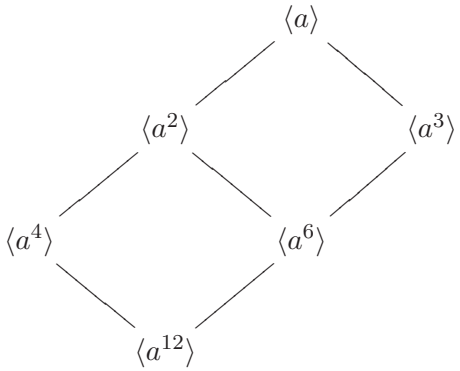
But $r < k$, so $r = 0$. Hence, $x = a^{kq} = (a^k)^q \in \langle a^k \rangle$. □

Corollary 1.4.10. *All distinct subgroups of $\mathbb{Z}$ are $k\mathbb{Z} = \{kq : q \in \mathbb{Z}\}$ where $k \in \mathbb{N} \cup \{0\}$.*

Theorem 1.4.11. [Generators of a Subgroup of a Finite Cyclic Group] *Let G be a finite cyclic group of order n. Then G has exactly one subgroup H of order d for each divisor d of n, and no other subgroups. Moreover, if G is generated by a, then H is generated by $a^{n/d}$.*

Proof. Let $d \mid n$. Since $(a^{n/d})^d = e$, $|a^{n/d}| \leq d$. If $|a^{n/d}| = r < d$, then $a^{nr/d} = e$ and $nr/d < n$ which contradicts $|a| = n$. Thus, $|a^{n/d}| = d$. Let H be a subgroup of G of order d. If $d = 1$, then $H = \{e\}$. Assume that $d > 1$. By Theorem 1.4.9, $H = \langle a^k \rangle$, where $k = \min\{m \in \mathbb{N} : a^m \in H\}$. Since $|H| = d$, $(a^k)^d = e$, so $n \mid kd$ which implies $\frac{n}{d} \mid k$. Thus, $k = \frac{n}{d}q$ for some $q \in \mathbb{Z}$. Hence, $a^k = (a^{n/d})^q \in \langle a^{n/d} \rangle$. It follows that $H \subseteq \langle a^{n/d} \rangle$. However, $|H| = d = |\langle a^{n/d} \rangle|$, so $H = \langle a^{n/d} \rangle$. □

Example 1.4.3. All subgroups of the cyclic group $G = \langle a \rangle$ of order 12 are shown in the following diagram.



The order of an element in a cyclic group and its generators are studied in the next theorem.

Theorem 1.4.12. [Order of an Element] *Let G be a finite cyclic group of order n generated by a and $m \in \mathbb{Z}$. Then*

1. $\langle a^m \rangle = \langle a^d \rangle$, *where* $d = \gcd(m, n)$.
2. $|a^m| = \dfrac{n}{\gcd(m, n)}$.
3. a^m *is a generator of* $G \Leftrightarrow \gcd(m, n) = 1$, *and so* G *contains precisely* $\phi(n)$ *elements of order* n.

Proof. (1) Since $d \mid m$, $\langle a^m \rangle \subseteq \langle a^d \rangle$. Since $d = \gcd(m, n)$, $d = mx + ny$ for some $x, y \in \mathbb{Z}$, so

$$a^d = a^{mx+ny} = a^{mx} a^{ny} = a^{mx} \in \langle a^m \rangle.$$

(2) $|a^m| = |\langle a^m \rangle| = |\langle a^d \rangle| = |\langle a^{n/(n/d)} \rangle| = \frac{n}{d}$ from Theorem 1.4.11.
(3) $\langle a^m \rangle = G \Leftrightarrow |a^m| = n \Leftrightarrow \frac{n}{d} = n \Leftrightarrow d = 1$. □

Remark. Since $\mathbb{Z}_n = \langle \overline{1} \rangle$ and $m \cdot \overline{1} = \overline{m}$, we have $\langle \overline{m} \rangle = \mathbb{Z}_n \Leftrightarrow \gcd(m, n) = 1 \Leftrightarrow \overline{m} \in \mathbb{Z}_n^\times$.

Theorem 1.4.13. *Let G be a group and $a \in G$. Then $|a| = n \Leftrightarrow (\forall k \in \mathbb{N}, a^k = e \Leftrightarrow n \mid k)$.*

Proof. Assume that $\forall k \in \mathbb{N}, a^k = e \Leftrightarrow n \mid k$. Since $n \mid n$, $a^n = e$. Let $k \in \mathbb{N}$ be such that $a^k = e$. Then $n \mid k$, so $n \leq k$. Hence, $|a| = n$. Another direction follows from Theorem 1.4.6 (6). □

Recall that an automorphism of a group G is an isomorphism on G. The set of all automorphisms of a group G is denoted by $\operatorname{Aut} G$ and is called the **automorphism group of** G. We shall close this section by studying the group of automorphisms of G and determining the automorphism group of cyclic groups.

Theorem 1.4.14. [Inner Automorphisms]

1. *With group operation composition of functions,* $\operatorname{Aut} G$ *is a group.*
2. *Each $g \in G$ determines an automorphism $\phi_g : G \to G$ defined by*

$$\phi_g(x) = gxg^{-1} \text{ for all } x \in G,$$

   and ϕ_g is called an **inner automorphism**. *The subgroup of* $\operatorname{Aut} G$ *consisting of the $\{\phi_g : g \in G\}$ is called the* **inner automorphism group of** G *and is denoted by* $\operatorname{Inn} G$.
3. *The map $\theta : g \mapsto \phi_g$ is a group homomorphism from G into* $\operatorname{Aut} G$.
4. *The kernel of θ is $Z(G)$, the center of G, and the image of θ is* $\operatorname{Inn} G$. *Consequently,*

$$G/Z(G) \cong \operatorname{Inn} G \leq \operatorname{Aut} G.$$

Example 1.4.4. $\operatorname{Aut} \mathbb{Z} \cong \mathbb{Z}_2$ and $\operatorname{Aut} \mathbb{Z}_n \cong \mathbb{Z}_n^\times$.

Proof. Let $\varphi \in \operatorname{Aut} \mathbb{Z}$. Note that for each $k \in \mathbb{N}$, $\varphi(k) = \varphi(k \cdot 1) = \varphi(\underbrace{1 + \cdots + 1}_{k}) = k \cdot \varphi(1)$ and $\varphi(-k) = -\varphi(k) = -(k \cdot \varphi(1))$, so φ is completely determined by $\varphi(1)$. Since φ is onto, $\operatorname{im} \varphi = \varphi(1)\mathbb{Z} = \mathbb{Z}$. Thus, $\varphi(1) \mid 1$, so $\varphi(1) = \pm 1$. Hence, $\operatorname{Aut} \mathbb{Z} = \{\pm \mathrm{id}\} \cong \mathbb{Z}_2$.

Let $\varphi \in \operatorname{Aut} \mathbb{Z}_n$. Similarly, φ is completely determined by $\varphi(\overline{1})$. Since φ is onto, $\operatorname{im} \varphi = \langle \varphi(\overline{1}) \rangle = \mathbb{Z}_n$. By Remark after Theorem 1.4.12, $\varphi(1) \in \mathbb{Z}_n^\times$. Therefore, $\operatorname{Aut} \mathbb{Z}_n \cong \mathbb{Z}_n^\times$ with isomorphism $\varphi \mapsto \varphi(\overline{1})$. □

Exercises 1.4.
1. Prove that if G is a group for which $G/Z(G)$ is cyclic, then G is abelian.
2. Let G be a group of order $2k$ where k is odd. Show that G contains a subgroup of index 2.
3. Let H be a proper subgroup of a finite group G. Show that $G \neq \bigcup_{g\in G} gHg^{-1}$.
4. Let G be a group and $a \in G$. If $\langle a\rangle \triangleleft G$ and $H < \langle a\rangle$, prove that H is normal in G.
5. Let G be a group and N a subgroup contained in the center of G. Suppose that G/N is cyclic. Prove that G is necessarily abelian.
6. Let G be a group. If $a, b \in G$ are of finite order such that $ab = ba$ and $\forall m \in \mathbb{N}, a^m b^m = e \Rightarrow a^m = b^m = e$, prove that $|ab| = \operatorname{lcm}(|a|, |b|)$.
7. Let m and n be integers. Prove the following statements.
   (a) $m\mathbb{Z} + n\mathbb{Z} = d\mathbb{Z}$ and $m\mathbb{Z} \cap n\mathbb{Z} = l\mathbb{Z}$ where $d = \gcd(m, n)$ and $l = \operatorname{lcm}(m, n)$.
   (b) If $\gcd(m, n) = 1$, then $\mathbb{Z}_{mn} \cong \mathbb{Z}_m \times \mathbb{Z}_n$. This is called the "Chinese remainder theorem". Is the converse true?
8. Let G be a group, K a normal subgroup of G of index r, and let $g \in G$ be an element of order n. Prove that if r and n are relatively prime, then $g \in K$.
9. Prove the following statements.
   (a) If $\gcd(m, n) = 1$, then $\operatorname{Aut}(\mathbb{Z}_m \times \mathbb{Z}_n) \cong \mathbb{Z}_m^\times \times \mathbb{Z}_n^\times$.
   (b) $\operatorname{Aut}(\mathbb{Z}_p \times \mathbb{Z}_p) \cong \mathrm{GL}_2(\mathbb{Z}_p)$.
10. Prove Theorem 1.4.14.
11. Let $H < G$. Prove that $C_G(H) \triangleleft N_G(H)$ and $N_G(H)/C_G(H)$ is isomorphic to a subgroup of $\operatorname{Aut} H$.
12. If G is a group for which $\operatorname{Aut}(G) = \{1\}$, prove that $|G| \leq 2$.

Project 4 (Generalization of Fermat's little theorem)**.** The project is based on I. M. Isaacs and M. R. Pournaki [26]. It gives a generalization of Fermat's little theorem using group actions.

(a) Let G be a finite group. For each positive integer a, let $[a]^G$ be the set of functions from G to $\{1, 2, \ldots, a\}$. Prove that

$$(g \cdot f)(h) = f(g^{-1}h) \quad \text{for all } g, h \in G \text{ and } f \in [a]^G$$

defines an action of G on the set $[a]^G$.

(b) Show that

$$\sum_{g\in G} a^{|G|/|g|} \equiv 0 \pmod{|G|}.$$

[*Hint*. Use Burnside's Theorem with the action in (a) to conclude that $\frac{1}{|G|}\sum_{g\in G} a^{|G|/|g|}$ is a positive integer.]

(c) Taking $G = \mathbb{Z}_m$, deduce that $\sum_{k=1}^{m} a^{\gcd(k,m)} \equiv 0 \pmod{m}$.

1.5 The Symmetric Group

In this section, we study the symmetric group on n letters, S_n. Recall that S_n is the group of permutations (1-1 and onto maps) on $\{1, 2, \ldots, n\}$ under composition. Its order is $n!$.

A permutation γ of $\{1, 2, \ldots, n\}$ which permutes a sequence of distinct elements $i_1, i_2, \ldots, i_r$, $r > 1$, cyclically in the sense that

$$\gamma(i_1) = i_2, \gamma(i_2) = i_3, \ldots, \gamma(i_{r-1}) = i_r, \text{ and } \gamma(i_r) = i_1$$

and fixed (that is, leaves unchanged) the other numbers in $\{1, 2, \ldots, n\}$ is called a **cycle** or an r**-cycle**. We denote this as

$$\gamma = (i_1 i_2 \ldots i_r).$$

It is clear that we can equally well write

$$\gamma = (i_2 i_3 \ldots i_r i_1) = (i_3 i_4 \ldots i_r i_1 i_2), \text{ etc.}$$

Two cycles γ and γ' are said to be **disjoint** if their symbols contain no common letters. In this case, it is clear that any number moved by one of these transformations is fixed by the other, i.e., $\forall i, \gamma(i) \neq i \Rightarrow \gamma'(i) = i$. Hence, if i is any number such that $\gamma(i) \neq i$, then $\gamma\gamma'(i) = \gamma(i)$, and since also $\gamma^2(i) \neq \gamma'(i)$, $\gamma'\gamma(i)$. Similarly, if $\gamma'(i) \neq i$, then $\gamma'\gamma(i) = \gamma'(i) = \gamma\gamma'(i)$. Also if $\gamma(i) = i = \gamma'(i)$, then $\gamma\gamma'(i) = \gamma'\gamma(i)$. Thus $\gamma\gamma' = \gamma'\gamma$, that is, we have proved (1) of the following theorem.

Theorem 1.5.1. [Order of a Cycle]
1. *Any two disjoint cycles commute.*
2. *If $\gamma = (i_1 i_2 \dots i_r)$ is an r-cycle, then the order of γ is r.*
3. *If $\alpha = (i_1 i_2 \dots i_{r_1})(j_1 j_2 \dots j_{r_2}) \dots (k_1 k_2 \dots k_{r_s})$ is a product of disjoint cycles, then the order of α is the least common multiple of $r_1, r_2, \dots, r_s$.*

Proof. For (2), clearly, $\gamma^r = (1)$. Let $1 \leq s < r$. Then $\gamma^s(i_1) = i_{s+1} \neq i_1$, so $\gamma^s \neq (1)$. (3) follows from (2) and the fact that $|ab| = \operatorname{lcm}(|a|, |b|)$ for all $a, b \in G$ such that $ab = ba$ and $\forall m \in \mathbb{N}, a^m b^m = e \Rightarrow a^m = b^m = e$ (see Exercises 1.4). □

It is convenient to extend the definition of cycles and the cycle notation to 1-cycles where we adopt the convention that for any i, (i) is the identity mapping. With this convention, we can see that:

Theorem 1.5.2. [Decomposition of a Permutation] *Every permutation is a product of disjoint cycles. Moreover, the product is unique up to rearranging its cycles and cyclically permuting the numbers within each cycle.*

Proof. Let $\sigma \in S_n$. If $\sigma = (1)$, we are done. Assume that $\sigma \neq 1$. Let $G = \langle \sigma \rangle$ act on $\{1, 2, \dots, n$ naturally as in Examples 1.3.1 (1). Let $B_1, B_2, \dots, B_r$ be distinct orbits of $\{1, 2, \dots, n\}$ under this action. For each $j \in \{1, 2, \dots, r\}$, we define the cycle μ_i by

$$\mu_i(x) = \begin{cases} \sigma(x), & \text{if } x \in B_i; \\ x, & \text{if } x \in \{1, 2, \dots, n\} \setminus B_i. \end{cases}$$

Since B_i, $i = 1, 2, \dots, r$, are disjoint, μ_i are disjoint cycles, and clearly, $\sigma = \mu_1 \mu_2 \dots \mu_r$. □

Remark. The above two theorems tell us how to find the order of an element in S_n.

Next, we shall discuss the cycle structure and the conjugacy class of a permutation.

Lemma 1.5.3. *If $\alpha \in S_n$ is a permutation, then*

$$\alpha(i_1 i_2 \dots i_r)\alpha^{-1} = (\alpha(i_1)\alpha(i_2) \dots \alpha(i_r)).$$

Proof. For $x \in \{1, 2, \dots, n\}$,

$$\begin{aligned} \alpha(i_1 i_2 \dots i_r)(x) &= \begin{cases} \alpha(i_{m+1 \mod r}), & \text{if } x = i_{m \mod r}; \\ \alpha(x), & \text{if } x \notin \{i_1, i_2, \dots, i_r\} \end{cases} \\ &= \begin{cases} \alpha(i_{m+1 \mod r}), & \text{if } \alpha(x) = \alpha(i_{m \mod r}); \\ \alpha(x), & \text{if } \alpha(x) \notin \{\alpha(i_1), \alpha(i_2), \dots, \alpha(i_r)\} \end{cases} \\ &= (\alpha(i_1)\alpha(i_2) \dots \alpha(i_r))(\alpha(x)). \end{aligned}$$

Hence, $\alpha(i_1 i_2 \dots i_r) = (\alpha(i_1)\alpha(i_2) \dots \alpha(i_r))\alpha$. □

If $\sigma \in S_n$ is the product of disjoint cycles of lengths $r_1, r_2, \ldots, r_s$ with $r_1 \leq r_2 \leq \ldots \leq r_s$ (including its 1-cycles) then the integers $r_1, r_2, \ldots, r_s$ are called the **cycle structure** of σ.

A **partition** of a positive integer n is any nondecreasing sequence of positive integers whose sum is n. For example, 5 has seven partitions, namely, $1+1+1+1+1$, $1+1+1+2$, $1+2+2$, $1+1+3$, $1+4$, $2+3$ and 5. The following result shows that the number of conjugacy classes of S_n and the number of partitions of n are coincide.

Theorem 1.5.4. *Two elements of S_n are conjugate if and only if they have the same cycle structure. The number of conjugacy classes of S_n equals the number of partitions of n.*

Proof. Assume that σ and τ are conjugate. Then $\tau = \alpha\sigma\alpha^{-1}$ for some $\tau \in S_n$. Write

$$\sigma = (i_1 i_2 \ldots i_{r_1})(j_1 j_2 \ldots j_{r_2}) \ldots (k_1 k_2 \ldots k_{r_s})$$

as a product of disjoint cycles. Thus,

$$\begin{aligned}\tau = \alpha\sigma\alpha^{-1} &= \alpha(i_1 i_2 \ldots i_{r_1})\alpha^{-1}\alpha(j_1 j_2 \ldots j_{r_2})\alpha^{-1}\alpha(k_1 k_2 \ldots k_{r_s})\alpha^{-1} \\ &= (\alpha(i_1)\alpha(i_2)\ldots\alpha(i_{r_1}))(\alpha(j_1)\alpha(j_2)\ldots\alpha(j_{r_2}))(\alpha(k_1)\alpha(k_2)\ldots\alpha(k_{r_s})).\end{aligned}$$

Hence, σ and τ have the same cycle structure.

Conversely, suppose that σ and τ have the same cycle structure written as a product of s disjoint cycles (including 1-cycles) as

$$\sigma = (a_1 a_2 \ldots a_{r_1})(a_{r_1+1} a_{r_1+2} \ldots a_{r_1+r_2}) \ldots (a_{r_1+r_2+\cdots+r_{s-1}+1} \ldots a_{n-1} a_n)$$

and

$$\tau = (b_1 b_2 \ldots b_{r_1})(b_{r_1+1} b_{r_1+2} \ldots b_{r_1+r_2}) \ldots (b_{r_1+r_2+\cdots+r_{s-1}+1} \ldots b_{n-1} b_n).$$

Define $\alpha \in S_n$ by $\alpha(a_i) = b_i$ for all $i \in \{1, 2, \ldots, n\}$. Then $\alpha\sigma\alpha^{-1} = \tau$. □

Example 1.5.1. The number of conjugacy classes of S_5 is 7 and $|\{\alpha(12)(345)\alpha^{-1} : \alpha \in S_5\}| = \binom{5}{2}(3-1)! = 20$.

Example 1.5.2. The Klein group $V_4 = \{(1), (12)(34), (13)(24), (14)(23)\}$ is normal in S_4 because V_4 contains of all products of disjoint 2-cycles and so $\forall \alpha \in S_4, \alpha V_4 \alpha^{-1} = V_4$ by Theorem 1.5.4. Moreover, since the group $\{(1), (12)(34)\}$ is of index two in V_4, it is normal in V_4 by Theorem 1.3.3. However, $\{(1), (12)(34)\}$ is not normal in S_4. Thus, *normality of subgroups is not transitive*.

Corollary 1.5.5. [Center of S_n] *If $n \geq 3$, then the center of S_n is trivial, i.e., $Z(S_n) = \{(1)\}$.*

Proof. We wish to prove that $\forall \alpha \in S_n[\forall \beta \in S_n, \beta\alpha\beta^{-1} = \alpha \Rightarrow \alpha = (1)]$. By Theorem 1.5.4, $\forall \alpha \in S_n[\alpha \neq (1) \Rightarrow |\{\beta\alpha\beta^{-1} : \beta \in S_n\}| > 1]$. By Corollary 1.3.8,$\forall \alpha \in S_n[|\{\beta\alpha\beta^{-1} : \beta \in S_n\}| > 1 \Rightarrow \alpha \notin Z(S_n)]$. Hence, let $\alpha \in Z(S_n)$. Then $|\{\beta\alpha\beta^{-1} : \alpha \in S_n\}| = 1$, so $\alpha = (1)$. □

To define an important subgroup of S_n, namely the *alternating group*, we shall need some results on 2-cycles. A cycle of the form (ab), where $a \neq b$, is called a **transposition**. It is easy to verify that

$$(i_1 i_2 \ldots i_r) = (i_1 i_r) \ldots (i_1 i_3)(i_1 i_2),$$

a product of $r-1$ transpositions. It follows that any $\alpha \in S_n$ is a product of transpositions. Also, a transposition (ab) has order two in S_n.

Theorem 1.5.6. 1. *(1) is always a product of even number of transpositions.*
2. *If $\alpha \in S_n$ is written as a product of transpositions, then either the number of transpositions in any product is always odd or always even.*

Proof. (1) Assume that we have two transposition decompositions

$$(1) = (x_1y_1)(x_2y_2)\dots(x_ky_k) = (1x_1)(1y_1)(1x_1)(1x_2)(1y_2)(1x_2)\dots(1x_k)(1y_k)(1x_k)$$

with $x_i < y_i$ for all $i \in \{1, 2, \dots, k\}$. Consider any $(1u), u > 1$, in the right hand side. Since the opposite side is (1), $(1u)$ must occur twice (or even number of times) in the right hand side. Note that ($1 \mapsto u$ and $u \mapsto 1$) will give $u \mapsto u$. Thus each transposition in the right hand side occurs even numbers of times, which implies that the right hand side should have even number of transpositions. Hence, k is even.

(2) Assume

$$\alpha = (x_1y_1)(x_2y_2)\dots(x_ky_k) = (w_1z_1)(w_2z_2)\dots(w_lz_l)$$

for some $x_i \neq y_i$, $w_j \neq z_j$ and $k, l \in \mathbb{N}$. Since $|(w_iz_i)| = 2$ for all i,

$$\begin{aligned}(x_1y_1)(x_2y_2)\dots(x_ky_k)(w_lz_l)^{-1}(w_{l-1}z_{l-1})^{-1}\dots(w_1z_1)^{-1} &= (1)\\ (x_1y_1)(x_2y_2)\dots(x_ky_k)(w_lz_l)(w_{l-1}z_{l-1})\dots(w_1z_1) &= (1),\end{aligned}$$

so $k + l$ is even. Hence, k and l have the same parity. □

The previous theorem leads to the definition of parity of a permutation. We call the permutation α **even** or **odd** according as α factors as a product of an even or an odd number of transpositions.

Remarks. Let $\alpha, \beta \in S_n$.

1. $\alpha\beta$ is even $\Leftrightarrow$ α and β have the same parity.
2. Since $\alpha\alpha^{-1} = (1)$ which is even, α and α^{-1} have the same parity.

Theorem 1.5.7. *Let $n > 1$. The set A_n of all even permutations forms a normal subgroup of S_n of index two. It is called the* **alternating group of degree** *n and $|A_n| = n!/2$.*

Proof. By Theorem 1.5.6, (1) is even. It is clear that the product of even permutations is even. Since a transposition has order two, the inverse of an even permutation is even. Hence, A_n is a subgroup of S_n. Since $n > 1$, let (ab) be a transposition in S_n. Clearly, (ab) is an odd permutation. We will show that $S_n = A_n \cup (ab)A_n$. Let $\alpha \in S_n$. If α is even, then $\alpha \in A_n$. On the other hand, assume that α is odd. Then $(ab)\alpha$ is even, so $(ab)\alpha \in A_n$, i.e., $\alpha \in (ab)A_n$. Thus, $[S_n : A_n] = 2$. In addition, since α and α^{-1} have the same parity, $\alpha A_n \alpha^{-1} \subseteq A_n$. Hence, A_n is normal in S_n. □

The above proof also shows that if $n > 1$, then $S_n = A_n \cup (ab)A_n$ and the number of even permutations and odd permutations are the same.

Corollary 1.5.8. *Let a group G act on a finite set X, and assume that some element $h \in G$ induces an odd permutation on X. Then there exists a normal subgroup N of G with $[G : N] = 2$ and $h \notin N$.*

Proof. Consider the diagram

$$G \xrightarrow{\theta} S(X) \xrightarrow{\pi} S(X)/A(X)\ ,$$

where $A(X)$ is the alternating group of even permutations on X, $\theta : g \to \phi_g$ and π is the canonical map. Since ϕ_h is an odd permutation, $\pi \circ \theta$ is onto. Choose $N = \ker \pi \circ \theta$. Then $N \triangleleft G$ and $G/N \cong S(X)/A(X)$. Thus, $[G : N] = [S(X) : A(X)] = 2$. Since $(\pi \circ \theta)(h) = \phi_h A(X) \neq A(X)$, we have $h \notin N$. □

Finally, we talk about the simplicity of A_n. A group is **simple** if it has no nontrivial normal subgroup. That is, all normal subgroups of G are $\{e\}$ and G. For example, $\mathbb{Z}_p$ is simple for all primes p.

Corollary 1.5.9. *Let $|G| = 2m$, where m is odd. Then G has a normal subgroup of order m. In particular, if $m > 1$, then G is not simple.*

Proof. Since $|G|$ is even, let g be an element of order two in G. Let G act on G by left multiplication and consider $G \xrightarrow{\theta} S(G)$. Since the action is faithful, θ is 1-1, so $|\theta(g)| = 2$. Thus, $\theta(g) = \phi_g$ is an odd permutation. By the previous corollary, there exists a normal subgroup N of G such that $[G : N] = 2$. Hence, G is not simple. □

Example 1.5.3. Since the Klein group $V_4 = \{(1), (12)(34), (13)(24), (14)(23)\}$ is normal in A_4, it follows that A_4 is not simple.

In general, we have the next theorem.

Theorem 1.5.10. *A_n is simple for all $n \neq 4$.*

Proof. Clearly, A_2 and A_3 are simple. For $n \geq 5$, we give a step-by-step guideline in Project 5. □

Corollary 1.5.11. *If $n \neq 4$, then the only normal subgroups of S_n are $\{(1)\}$, A_n and S_n.*

Exercises 1.5.

1. Prove that A_4 is the only subgroup of S_4 of order 12.
2. Prove that A_4 has no subgroup of order six.
3. Determine all normal subgroups of S_4. (*Hint*. Use conjugacy classes.)
4. (a) Find the largest positive integer n such that S_{10} has a permutation of order n.
   (b) The **exponent** of a finite group G is the smallest positive integer n such that $g^n = 1$ for all $g \in G$. Find the exponent of S_{30}, the symmetric group on 30 letters.
5. Show that if H is any subgroup of S_n, $n \geq 2$, then either all permutations in H are even or exactly half are even.
6. Let G be a group of order 360 having a maximal subgroup isomorphic to A_5. Prove that $G \cong A_6$.

Project 5 (Simplicity of A_n)**.** Prove that A_n is simple for $n \geq 5$, following the steps and hints given.

(a) Show A_n contains every 3-cycle if $n \geq 3$.

(b) Show A_n is generated by the 3-cycles for $n \geq 3$. [*Hint*. Note that $(a, c)(a, b) = (a, b, c)$ and $(a, b)(c, d) = (a, c, b)(a, c, d)$.]

(c) Let r and s be fixed elements of $\{1, 2, \ldots, n\}$ for $n \geq 3$. Show that A_n is generated by the n "special" 3-cycles of the form (r, s, i) for $1 \leq i \leq n$. [*Hint*. Show every 3-cycle is the product of "special" 3-cycles by computing

$$(r, s, i)^2, (r, s, j)(r, s, i)^2, (r, s, j)^2(r, s, i) \quad \text{and} \quad (r, s, i)^2(r, s, k)(r, s, j)^2(r, s, i).$$

Observe that these products give all possible types of 3-cycles.]

(d) Let N be a normal subgroup of A_n for $n \geq 3$. Show that if N contains a 3-cycle, then $N = A_n$. [*Hint*. Show that $(r, s, i) \in N$ implies that $(r, s, j) \in N$ for $j = 1, 2, \ldots, n$ by computing

$$((r, s)(i, j))(r, s, i)^2((r, s)(i, j))^{-1}.]$$

(e) Let N be a nontrivial normal subgroup of A_n for $n \geq 5$. Show that one of the following cases must hold, and conclude in each case that $N = A_n$.
Case 1. N contains a 3-cycle.
Case 2. N contains a product of disjoint cycles, at least one of which has length greater than 3. [*Hint*. Suppose N contains the disjoint product $\sigma = \mu(a_1, a_2, \ldots, a_r)$. Show that $\sigma^{-1}(a_1, a_2, a_3)\sigma(a_1, a_2, a_3)^{-1}$

is in N, and compute it.]
Case 3. N contains a disjoint product of the form $\sigma = \mu(a_4, a_5, a_6)(a_1, a_2, a_3)$. [*Hint*. Show that $\sigma^{-1}(a_1, a_2, a_4)\sigma(a_1, a_2, a_4)^{-1}$ is in N, and compute it.]
Case 4. N contains a disjoint product of the form $\sigma = \mu(a_1, a_2, a_3)$ where μ is a product of disjoint 2-cycles. [*Hint*. Show $\sigma^2 \in N$ and compute it.]
Case 5. N contains a disjoint product σ of the form $\sigma = \mu(a_3, a_4)(a_1, a_2)$, where μ is a product of an even number of disjoint $2-$cycles.
[*Hint*. Show that $\sigma^{-1}(a_1, a_2, a_3)\sigma(a_1, a_2, a_3)^{-1}$ is in N, and compute it to deduce that $\alpha = (a_2, a_4)(a_1, a_3)$ is in N. Using $n \geq 5$ for the first time, find $i \in \{1, 2, \ldots, n\}$, where $i \neq a_1, a_2, a_3, a_4$. Let $\beta = (a_1, a_3, i)$. Show that $\beta^{-1}\alpha\beta\alpha \in N$, and compute it.]

Project 6 (Wilson)**.** Let p be a prime. Taking $G = S_p$ in the proof of Cauchy's theorem (Theorem 1.3.12), we see that the set $\{\sigma \in S_p : \sigma^p = (1)\}$ is of cardinality a multiple of p. Count the number of elements in this set and deduce that $(p-1)! \equiv -1 \pmod{p}$. This provides another proof of Wilson's theorem.

1.6 Sylow Theorems

We know the order of a subgroup of a finite group G must divide $|G|$. If $|G|$ is cyclic (even only abelian), then there exist subgroups of every order dividing $|G|$. A natural question is: If k divides $|G|$ is there always a subgroup of G of order k? A little experimenting shows that this is not so. For example, the alternating group A_4, whose order is 12, contain no subgroup of order six. Moreover, A_n for $n \geq 5$ is simple, that is, contains no normal subgroup $\neq 1$, A_n. Since any subgroup of index two is normal, it follows that A_n, $n \geq 5$, contains no subgroup of order $n!/4$.

1.6.1 Sylow p-subgroups

The main positive result of the type we are discussing was discovered by Sylow `/see lov/`. Its proof provides us another application of action of a group on a set. Unless specified otherwise, p denotes a prime.

A group G is said to be a p**-group** if $|a|$ is a power of p for all $a \in G$.

Example 1.6.1. The group $\mathbb{Z}_{p^n}$ is a p-group. If X is a set, then $(P(X), \triangle)$ is a 2-group.

Since we mainly study finite groups, the following corollary will be useful. Lagrange theorem and Cauchy theorem imply each direction, respectively.

Corollary 1.6.1. *Let G be a finite group. Then G is a p-group $\Leftrightarrow$ $|G|$ is a power of p.*

Remark. Let P and Q be subgroups of G. If P is a p-group and Q is a q-group, where p and q are distinct primes, then $P \cap Q = \{e\}$.

Theorem 1.6.2. *Let G be a finite p-group and $|G| > 1$. Then the following statements hold.*
1. *$|Z(G)| > 1$.*
2. *If $|G| = p^2$, then G is abelian.*

Proof. By Corollary 1.6.1, $|G| = p^l$ for some $l \in \mathbb{N}$. Recall from Corollary 1.3.8 that

$$|G| = |Z(G)| + \sum_{i=1}^{s} |\{gx_ig^{-1} : g \in G\}| = |Z(G)| + \sum_{i=1}^{s} [G : C_G(x_i)],$$

where $x_1, \ldots, x_s$ represent the conjugacy classes of G which contains more than one element. Since $[G : C_G(x_i)] = |G|/|C_G(x_i)| > 1$ for all i and $|G| = p^l$, p divides $|\{gx_ig^{-1} : g \in G\}|$ for all $i \in \{1, 2, \ldots, s\}$. Hence, $p \mid |Z(G)|$, so $|Z(G)| > 1$. This proves (1). For the second part, assume that $|G| = p^2$. We know that $Z(G)$ is a normal subgroup of G and $|Z(G)| > 1$. By Lagrange Theorem, $|Z(G)| = p$ or $|Z(G)| = p^2$. If $|Z(G)| = p^2$, we have $Z(G) = G$ and so G is abelian. Suppose that $|Z(G)| = p$. Then $G/Z(G)$ is of order p and so a cyclic group. This implies that G is abelian. □

Because all subgroups of a p-group have p-power index, the length of an orbit under an action by a p-group is a multiple of p unless the point is a fixed point, when its orbit has length one. This leads to an important congruence modulo p when a p-group is acting.

Lemma 1.6.3. [Fixed Point Congruence] *Let G be a finite p-group.*
If G acts on a finite set X and $X_0 = \{x \in X : gx = x \text{ for all } g \in G\}$, then $|X_0| \equiv |X| \mod p$.
Here, X_0 is called the **set of fixed points**.

Proof. We observe first that $X_0 = \{x \in X : |G \cdot x| = 1\}$. Let $x_1, \ldots, x_s$ represent the orbits of X which contains more than one element. Then

$$|X| = |X_0| + \sum_{i=1}^{s} |G \cdot x_i|.$$

By Orbit-Stabilizer Theorem, for each $i \in \{1, 2, \ldots, s\}$, $1 < |G \cdot x_i| = [G : \mathrm{Stab}_G x_i]$ which is divisible by p. Hence, $|X_0| \equiv |X| \mod p$ as desired. □

Lemma 1.6.4. *Let G be a finite group and $H, P \leq G$. If H is a p-group, then*
1. *$|\{xP : x \in G \text{ and } H \subseteq xPx^{-1}\}| \equiv [G : P] \mod p$,*
2. *$[N_G(H) : H] \equiv [G : H] \mod p$, and*
3. *if $p \mid [G : H]$, then $p \mid [N_G(H) : H]$ and $N_G(H) \neq H$.*

Proof. Let $X = \{xP : x \in G\}$ and let H act on X by $h \cdot xP = hxP$ for all $x \in G$ and $h \in H$. Clearly, $|X| = [G : P]$ and

$$\begin{aligned} X_0 &= \{xP : x \in G \text{ and } hxP = xP \text{ for all } h \in H\} \\ &= \{xP : x \in G \text{ and } x^{-1}hx \in P \text{ for all } h \in H\} \\ &= \{xP : x \in G \text{ and } x^{-1}Hx \subseteq P\} = \{xP : x \in G \text{ and } H \subseteq xPx^{-1}\}, \end{aligned}$$

so $|\{xP : x \in G \text{ and } H \subseteq xPx^{-1}\}| \equiv [G : P] \mod p$ by Lemma 1.6.3. Furthermore, if $P = H$, we have

$$X_0 = \{xH : x \in G \text{ and } H \subseteq xHx^{-1}\}.$$

Since $\forall x \in G, |xHx^{-1}| = |H|$ and H is finite, we have

$$|X_0| = |\{xH : x \in G \text{ and } x^{-1}Hx = H\}| = |\{xH : x \in N_G(H)\}| = [N_G(H) : H],$$

so $[N_G(H) : H] \equiv [G : H] \mod p$. The final result clearly follows from (2). □

We now discuss three theorems due to Sylow. The first theorem shows the existance of a maximal p-subgroup of a finite group G.

Theorem 1.6.5. [First Sylow Theorem] *Let G be a group of order $p^n m$ where $n \geq 1$ and p does not divide m. Then the following statements hold.*
1. *G contains a subgroup of order p^i for all $1 \leq i \leq n$.*
2. *For each i, where $1 \leq i < n$, every subgroup H of G of order p^i is a normal subgroup of a subgroup of order p^{i+1}.*

Proof. Since p divides $|G|$, by Cauchy theorem, G has a subgroup H_1 of order p. Assume that $k \in \{1, 2, \ldots, n-1\}$ and G has a subgroup H_k of order p^k. Then the index $[G : H_k] = p^{n-k}m$ and $n - k \geq 1$. By Lemma 1.6.4, p divides $[N_G(H_k) : H_k] = |N_G(H_k)/H_k|$. Again, by Cauchy theorem, $N_G(H_k)/H_k$ has a subgroup $\mathcal{H}$ of order p. By the Third Isomorphism Theorem, $\mathcal{H} = H_{k+1}/H_k$ for some subgroup H_{k+1} of $N_G(H_k)$ containing H_k. Moreover, $H_k \triangleleft H_{k+1}$ and $|H_{k+1}| = |\mathcal{H}||H_k| = p\,p^k = p^{k+1}$. Hence, there are subgroups $H_1, H_2, \ldots, H_n$ of G such that $|H_i| = p^i$ for $i = 1, 2, \ldots, n$ and $H_1 \triangleleft H_2 \triangleleft \cdots \triangleleft H_n$. □

A maximal p-subgroup of a group G is called a **Sylow p-subgroup of G**. By Zorn's lemma, we have the following statements.

1. A Sylow p-subgroup of a group G always exists and it may be trivial.
2. Every p-subgroup of a group G is contained in a Sylow p-subgroup of G.

By Corollary 1.6.1 and Theorem 1.6.5, if G is a finite group and p is a prime such that p divides $|G|$, then G has a Sylow p-subgroup of order $p^n \| |G|$. (Here, $p^n \| |G|$ means n is the highest power of p dividing $|G|$.) That is, $[G : P]$ is not divisible by p. Moreover, we have:

Corollary 1.6.6. *Let G be a group of order $p^n m$ where $n \geq 1$ and p does not divide m.*

1. *G has a Sylow p-subgroup of order p^n.*
2. *For $H < G$, H is a Sylow p-subgroup of $G \Leftrightarrow |H| = p^n$.*
3. *Every conjugate of a Sylow p-subgroup of G is a Sylow p-subgroup of G.*
4. *If P is the only one Sylow p-subgroup of G, then P is normal in G.*

Proof. (1) and (2) follow from the definition and the above discussion. Since a conjugate of a subgroup of G is of the same order as the subgroup, (2) implies (3). Finally, (4) follows from (3). □

The second and third Sylow theorems determine all Sylow p-subgroups and possible numbers of Sylow p-subgroups, respectively. Also, they give the converse of the above results.

Theorem 1.6.7. [Second Sylow Theorem] *Let G be a finite group.*

1. *If P is a Sylow p-subgroup of G and H is a p-subgroup of G, then $H \subseteq xPx^{-1}$ for some $x \in G$.*
2. *Any two Sylow p-subgroups of G are conjugate.*

Proof. By Lemma 1.6.4 $|\{xP : x \in G \text{ and } H \subseteq xPx^{-1}\}| \equiv [G : P] \mod p$. Since P is a Sylow p-subgroup of G, $p \nmid [G : P]$, so $\{xP : x \in G \text{ and } H \subseteq xPx^{-1}\} \neq \emptyset$. Thus, there exists an $x \in G$ such that $H \subseteq xPx^{-1}$.

Next, we let P_1 and P_2 be Sylow p-subgroups of G. Then there exists an $x \in G$ such that $P_1 \subseteq xP_2x^{-1}$. But $|P_1| = |P_2| = |xP_2x^{-1}|$ and G is finite, $P_1 = xP_2x^{-1}$. □

Corollary 1.6.8. *Let G be a finite group and let P be a Sylow p-subgroup of G.*

1. *$\{xPx^{-1} : x \in G\}$ is the set of all Sylow p-subgroups of G.*
2. *The number of Sylow p-subgroups of G is $[G : N_G(P)]$ and it divides $[G : P]$ and $|G|$.*
3. *P is normal in $G \Leftrightarrow P$ is the only one Sylow p-subgroup of G.*

For a finite group G and a prime p divides $|G|$, we write $n_p(G)$ for the number of Sylow p-subgroups of G.

Theorem 1.6.9. [Third Sylow Theorem] *If G is a finite group and a prime p divides $|G|$, then $n_p(G) \equiv 1 \mod p$.*

Proof. Let P be a Sylow p-subgroup of G. Then the set $X = \{xPx^{-1} : x \in G\}$ consists of all Sylow p-subgroups of G. Let P act on X by conjugation, namely, $(g, xPx^{-1}) \mapsto gxPx^{-1}g^{-1}$ for all $g \in P$ and $x \in G$. Since $gPg^{-1} = P$ for all $g \in P$, $P \in X_0$. Let $Q \in X_0$. Then $gQg^{-1} = Q$ for all $g \in P$, so $P \subseteq N_G(Q)$. Since P and Q are Sylow p-subgroups of $N_G(Q)$ and Q is normal in $N_G(Q)$, $P = Q$ by the uniqueness of normal Sylow p-subgroup. This proves $X_0 = \{P\}$. By Lemma 1.6.3, we have $n_p(G) = |X| \equiv |X_0| = 1 \mod p$ as desired. □

1.6.2 Applications of Sylow Theorems

Here, we present some applications of Sylow theorems on a finite group. The proofs use basic properties of subgroups, quotient groups, cyclic groups and symmetric groups studied previously. We shall see many techniques in group theory in this subsection.

Theorem 1.6.10. *Let G be a finite group. If P is a Sylow p-subgroup of G, then*

$$N_G(N_G(P)) = N_G(P).$$

Proof. Since $P \triangleleft N_G(P)$, P is the only Sylow p-subgroup of $N_G(P)$. Let $x \in N_G(N_G(P))$. Then $xN_G(P)x^{-1} = N_G(P)$. Since $P \subseteq N_G(P)$, $xPx^{-1} \subseteq N_G(P)$. Thus, $xPx^{-1} = P$ since xPx^{-1} is a Sylow p-subgroup of G. Hence, $x \in N_G(P)$. □

Theorem 1.6.11. [Group of Order pq] *Let G be a group of order pq where p and q are primes and $p < q$. Then G is a cyclic group, or G has q Sylow p-subgroups and $p \mid (q-1)$.*

Proof. Since the number of Sylow p-subgroups divides $|G| = pq$, it is 1, p, q or pq. But this number is $\equiv 1 \mod p$, so it is 1 or q. If G has q Sylow p-subgroups, then we are done. Assume that G has only one Sylow p-subgroup, say P. Then P is normal in G. Consider the number of Sylow q-subgroups of G. It is again 1, p, q or pq, and $\equiv 1 \mod q$, so the only possibility is 1 since $p < q$. Thus, G also has a unique Sylow q-subgroup, say Q, and so Q is normal in G. Since the orders of P and Q are prime, both P and Q are cyclic. Let a and b be generators of P and Q, respectively. Note that $aba^{-1}b^{-1} \in P \cap Q = \{e\}$. Thus, $ab = ba$, so $|ab| = pq = |G|$. Hence, $G = \langle ab \rangle$. □

Remark. Theorem 1.6.11 demonstrates the power of the Sylow theorems in classifying the finite groups whose orders have small numbers of prime factor. Results along this lines of this theorem exist for groups of order p^2q, p^2q^2, p^3 and p^4, where $p < q$ are primes.

Example 1.6.2. There can be no simple groups of order 200 and of order 280.

Proof. Let H be a group of order 200. Let P be a Sylow 5-subgroup of H. Then $n_5(H)$ divides $[H : P] = 8$ and $n_5(H) \equiv 1 \pmod 5$, so $n_5(H) = 1$. Hence, P is normal in H.

Next, let G be a group of order 280. By Corollary 1.6.8 and Theorem 1.6.9, we have $n_2(G) = 1, 5, 7$ or 35, $n_5(G) = 1$ or 56 and $n_7(G) = 1$ or 8. If $n_5(G) = 1$ or $n_7(G) = 1$, we are done. Assume that $n_5(G) = 56$ and $n_7(G) = 8$. Then we have $56 \cdot 4 = 224$ elements of order 5, and $8 \cdot 6 = 48$ elements of order 7. Hence, G has a unique Sylow 2-subgroup. □

Example 1.6.3. Let G be a group of order 30. Then

1. Either a Sylow 3-subgroup or a Sylow 5-subgroup is normal in G.
2. G has a normal subgroup of order 15.
3. Both a Sylow 3-subgroup and a Sylow 5-subgroup are normal in G.

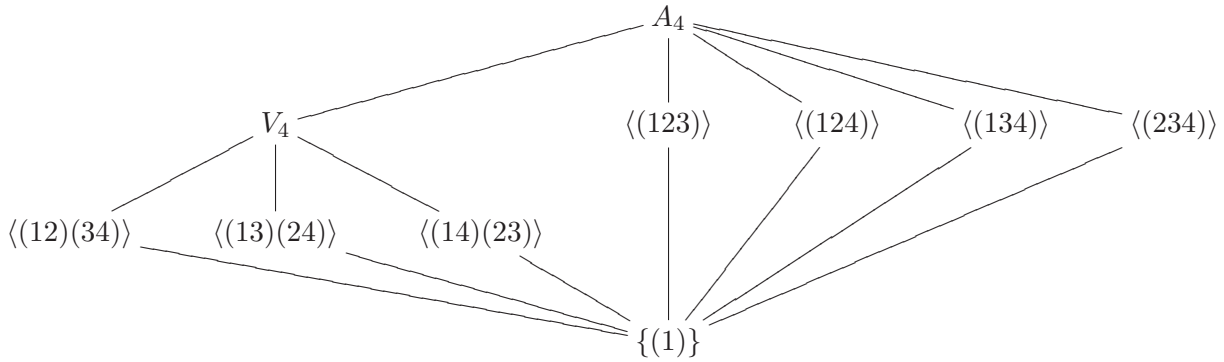
Proof. Assume that neither a Sylow 3-subgroup nor a Sylow 5-subgroup are normal in G. By Corollary 1.6.8, $n_3(G)$ and $n_5(G)$ are more than one and are factors of $|G|$. By Third Sylow Theorem, $n_3(G) \geq 10$ and $n_5(G) \geq 6$, so G contains at least 20 elements of order three and at least 24 elements of order five. This exceeds the number of elements of G, a contradiction. Thus, we have (1). Now, let P_3 and P_5 be a Sylow 3-subgroup and a Sylow 5-subgroup of G, respectively. By (1), we see that P_3 or P_5 is normal in G, so P_3P_5 is a subgroup of G. Since $P_3 \cap P_5 = \{e\}$, $|P_3P_5| = 15$, so the index $[G : P_3P_5]$ is two. Hence, P_3P_5 is normal in G. This proves (2).

Finally, we assume that P_3 is normal while P_5 is not. Thus, G has two elements of order three at least 24 elements of order five. By Theorem 1.6.11, P_3P_5 is cyclic, so G has $\phi(15) = 8$ elements of order 15. Hence, G contains more than 30 elements, a contradiction. On the other hand, we assume that P_5 is normal while P_3 is not. Thus, G has four elements of order five at least 20 elements of order three. Again, G also contains 8 elements of order 15. This leads to a contradiction, so P_3 and P_5 are normal in G as desired. □

Example 1.6.4. Every group G of order 12 that is not isomorphic with A_4 contains an element of order 6.

Proof. If A is a Sylow 3-subgroup, then $A = \langle a \rangle$ and $|a| = 3$. Let G act on $\{A, x_2A, x_3A, x_4A\}$ by $(g, xA) \mapsto gxA$. This action induces a homomorphism $\theta : G \to S_4$ whose kernel K is a subgroup of A. Then $K = \{e\}$ or $K = A$. If $K = \{e\}$, then G is isomorphic to a subgroup of S_4 of order 12, so $G \cong A_4$ which is excluded by hypothesis. Thus, $A = K$ is normal in G which implies that A is a unique Sylow 3-subgroup of G. Hence, a and a^2 are only two elements of order 3 in G. Since $[G : C_G(a)]$ is the number of conjugates of a which is 1 or 2, $|C_G(a)| = 12$ or 6, so there is a $b \in C_G(a)$ of order two. Since $ab = ba$, $|ab| = 6$. □

Example 1.6.5. Recall that $V_4 = \{(1), (12)(34), (13)(24), (14)(23)\}$ is a normal subgroup of A_4. Since $|A_4| = 12 = 2^2 \cdot 3$, V_4 is the unique Sylow 2-subgroup of A_4. Moreover V_4 has three subgroups of order two, namely $\langle (12)(34) \rangle$, $\langle (13)(24) \rangle$ and $\langle (14)(23) \rangle$. Next, we analyze the Sylow 3-subgroups of A_4. They are cyclic subgroups of order three generated by a 3-cycle. Note that there are eight 3-cycles in A_4, so we have four subgroups of order three, which are $\langle (123) \rangle$, $\langle (124) \rangle$, $\langle (134) \rangle$ and $\langle (234) \rangle$. By Exercises 1.5, A_4 has no subgroup of order six. Hence, the diagram below shows all subgroups of A_4.



We shall see more applications of Sylow theorems in Section 3.3. It turns out that any finite nilpotent group is the direct product of its Sylow p-subgroups.

Exercises 1.6.

1. If G is a finite p-group where p is a prime, N is normal in G and $N \neq \{e\}$, prove that $N \cap Z(G) \neq \{e\}$.
2. Prove that if $|G| = pn$ with $p > n$, p is a prime, and H is a subgroup of G of order p, then $H \triangleleft G$.
3. Let p be the smallest prime dividing the order of a finite group G. Show that any subgroup H of G of index p is normal.

4. Let G be a group of order p^n where p is a prime and $n \in \mathbb{N}$. Prove that there exist normal subgroups $N_1, \ldots, N_n$ of G such that $N_1 < N_2 < \cdots < N_n$ with $|N_i| = p^i$ for all $i \in \{1, 2, \ldots, n\}$.
5. Let G be a group, $M \triangleleft G$ and $N \triangleleft G$. Prove the following statements.
   (a) If $M \cap N = \{e\}$, then $xy = yx$ for all $x \in M$ and $y \in N$.
   (b) If M and N are finite cyclic subgroups of G and $\gcd(|M|, |N|) = 1$, then MN is a cyclic subgroup of G of order $|M||N|$.
6. Let P be a Sylow p-subgroup of a finite group G and N a normal subgroup of G. Show that:
   (a) $P \cap N$ is a Sylow p-subgroup of N,
   (b) PN/N is a Sylow p-subgroup of G/N.
7. Show that there are no simple groups of order 148 or of order 56.
8. How many elements of order 7 are there in a simple group of order 168?
9. Let G be a group of order 153. Prove that G is abelian.
10. Let G be a group of order 231. Show that $n_{11}(G) = 1$ and the Sylow 11-subgroup of G is contained in $Z(G)$.
11. Show that there is no non-abelian finite simple group of order less than 60. (*Hint*. We may focus on groups of the following orders: $24, 30, 40, 48, 54$ and 56.)
12. Let G be a group of order 385. Show that a Sylow 11-subgroup of G is normal and a Sylow 7-subgroup of G is contained in $Z(G)$.
13. Let p be a prime and P a Sylow p-subgroup of a finite group G. Suppose that, for all $g \in G$, if $P \neq gPg^{-1}$, then $P \cap gPg^{-1} = \{e\}$. Show that $n_p(G) \equiv 1 \mod |P|$.
14. Let G be a group of order 2013. Prove that G has a proper normal subgroup N such that G/N is cyclic.
15. (a) Let G be a finite group and N a normal subgroup of G. If N contains a Sylow p-subgroup of G, prove that the number of Sylow p-subgroups of N is the same as that of G (i.e., $n_p(N) = n_p(G)$).
    (b) Show that if G is a group of order 130, then G has a normal subgroup of order 5.
16. Let G be a finite group acting transitively on a finite set X. Let $x \in X$ and G_x the stabilizer of x. Let P be a Sylow p-subgroup of G_x. Show that the subgroup $N_G(P) = \{z \in G : zPz^{-1} = P\}$ of G acts transitively on $Y = \{y \in X : hy = y \text{ for all } h \in P\}$.

Project 7 (Simple groups of small order)**.** We have learned from the above exercise that the smallest non-abelian simple group is of order 60 by using Sylow's theorems to eliminate the groups of smaller order. Write a computer program that uses Sylow's theorems to eliminate all orders between 1 and say 1,000 (or more) for which group that cannot be simple. For any order that could have a simple group G, list $n_p(G)$ for all primes p dividing the order.

Project 8 (Lucas' congruence)**.** Let p be a prime and let $n \geq m$ be non-negative integers. Write $n = pn' + a_0$ and $m = pm' + b_0$ where $0 \leq a_0, b_0 \leq p - 1$. Decompose $\{1, 2, \ldots, n\}$ into a union of p blocks on n' consecutive integers, from 1 to pn', followed by a final block of length a_0. That is, let

$$A_i = \{in' + 1, in' + 2, \ldots, (i + 1)n'\}$$

for $0 \leq i \leq p - 1$, so

$$\{1, 2, \ldots, n\} = A_0 \cup A_1 \cup \cdots \cup A_{p-1} \cup \{pn' + 1, pn' + 2, \ldots, pn' + a_0\}.$$

For $1 \leq t \leq n'$, let σ_t be the p-cycle

$$\sigma_t = (t, n' + t, 2n' + t, \ldots, (p - 1)n' + t).$$

This cycle cyclically permutes the numbers in $A_0, A_1, \ldots, A_{p-1}$ that are $\equiv t \pmod{n'}$. The σ_t's for different t are disjoint, so they commute. Set $\sigma = \sigma_1 \sigma_2 \ldots \sigma_n$. Then σ has order p as a permutation of $\{1, 2, \ldots, n\}$ (fixing all numbers above pn'). Let X be the set of m-element subsets of $\{1, 2, \ldots, n\}$. Then $|X| = \binom{n}{m}$. Let the group $\langle \sigma \rangle$ act on X.

(a) Show that the number of fixed points of this action is $\binom{a_0}{b_0}\binom{n'}{m'}$. Deduce, by Lemma 1.6.3, that $\binom{n}{m} \equiv \binom{a_0}{b_0}\binom{n'}{m'} \pmod{p}$.

(b) Prove Lucas' congruence: if $n = a_0 + a_1 p + a_2 p^2 + \cdots + a_k p^k$ and $m = b_0 + b_1 p + b_2 p^2 + \cdots + b_k p^k$ with $0 \le a_i, b_i \le p-1$, then

$$\binom{n}{m} \equiv \binom{a_0}{b_0}\binom{a_1}{b_1}\cdots\binom{a_k}{b_k} \pmod{p}.$$

1.7 Finite Abelian Groups

The study of finite non-abelian groups is complicated as we have learned that the Sylow theorems give us some important information about them. This section gives us complete information about all finite abelian groups. We start with formal definitions of the direct product of groups.

Let A and B be groups. The **direct product of A and B** is defined as:

1. a set $A \times B = \{(a,b) : a \in A \text{ and } b \in B\}$ is the Cartesian product of A and B,
2. multiplication is coordinatewise, namely, $(a,b)(c,d) = (ac, bd)$.

More generally, if $\{A_i : i \in I\}$ is a family of groups, then $\prod_{i\in I} A_i$ is a group with coordinatewise multiplication. It is called the **direct product of the groups** A_i. The subgroup

$$\prod_{i\in I}^{w} A_i = \left\{(a_i) \in \prod_{i\in I} A_i : a_i = e \text{ for all but finitely many } i\right\}$$

of $\prod_{i\in I} A_i$ is called the **external weak direct product of the groups** A_i. Note that it is normal in $\prod_{i\in I} A_i$. In case A_i are all additive abelian groups, we may write $\sum_{i\in I} A_i$ for $\prod_{i\in I} A_i$ and $\sum_{i\in I}^{w} A_i$ for $\prod_{i\in I}^{w} A_i$.

Let G be a group. It is easy to show that:

1. If $N_1, \ldots, N_m$ are normal subgroups of G, then

$$N_1 N_2 \cdots N_m = \langle N_1 \cup N_2 \cup \cdots \cup N_m \rangle.$$

2. If $\{N_i : i \in I\}$ is a family of normal subgroups of G, then

$$\left\langle \bigcup_{i\in I} N_i \right\rangle = \bigcup_{\substack{i_1,\ldots,i_m \in I,\\ m\in\mathbb{N}}} N_{i_1} \cdots N_{i_m}.$$

Theorem 1.7.1. *Let $\{N_i : i \in I\}$ be a family of normal subgroups of G such that*
1. $G = \left\langle \bigcup_{i\in I} N_i \right\rangle$ and 2. $\forall k \in I, N_k \cap \left\langle \bigcup_{i\in I\setminus\{k\}} N_i \right\rangle = \{e\}$.
Then $G \cong \prod_{i\in I}^{w} N_i$.

The group G satisfies conditions of Theorem 1.7.1 is called the **internal weak direct product of** $\{N_i : i \in I\}$ and we write $G = \prod_{i\in I}^{w} N_i$. If G is additive abelian, then G is called the **internal direct sum of** $\{N_i : i \in I\}$ and we write $G = \bigoplus_{i\in I} N_i$.

Corollary 1.7.2. *Let $N_1, N_2, \ldots, N_m$ be normal subgroups of G. If $G = N_1 \ldots N_m$ and $N_k \cap (N_1 \ldots N_{k-1} N_{k+1} \ldots N_m) = \{e\}$ for all $k \in \{1, \ldots, m\}$, then $G \cong N_1 \times \cdots \times N_m$.*

Proof of Theorem 1.7.1. From (2), for each $i, j \in I$ with $i \neq j$, we have $N_i \cap N_j = \{e\}$, this implies that $xy = yx$ for all $x \in N_i$ and $y \in N_j$ because N_i and N_j are normal in G.

Define $\varphi : \prod_{i\in I}^{w} N_i \to G$ by

$$\varphi(\{a_i\}) = \prod_{i\in I} a_i$$

which is a finite product since $a_i = e$ for all but a finite number of $i \in I$ and it is a well defined homomorphism by the previous observation. To show that this map is onto, let $x \in G$. Since

G is generated by $\bigcup_{i\in I} N_i$, $G = \bigcup_{\substack{i_1,\ldots,i_m\in I,\\ m\in\mathbb{N}}} N_{i_1}\cdots N_{i_m}$, so there are distinct $k_1,\ldots,k_l \in I$ and $a_1 \in N_{k_1},\ldots,a_l \in N_{k_l}$ such that $x = a_{k_1}\ldots a_{k_l}$. Let $\{x_i\}$ in $\prod_{i\in I}^w N_i$ be defined by $x_i = a_i$ if $i \in \{k_1,\ldots,k_l\}$ and $x_i = e$ for other i. Then $\varphi(\{x_i\}) = \prod_{i\in I} x_i = a_{k_1}\ldots a_{k_l} = x$ as required. Finally, we show that φ is injective, let $\{a_i\} \in \prod_{i\in I}^w N_i$ be such that $\prod_{i\in I} a_i = e$. Then for each $k \in I$, $a_k^{-1} = \prod_{i\in I\setminus\{k\}}$ is in $N_k \cap \langle\bigcup_{i\in I\setminus\{k\}} N_i\rangle = \{e\}$. This implies that $a_k = e$ for all $k \in I$, and hence φ is an isomorphism. □

The proof of injectivitity above also implies the following theorems.

Theorem 1.7.3. *Let $N_1, N_2, \ldots, N_m$ be normal subgroups of G. Then the following statements are equivalent.*
(i) G is the internal weak direct product of $N_1,\ldots,N_m$.
(ii) $\forall x \in G, \exists! a_1 \in N_1,\ldots,a_m \in N_m, x = a_1\ldots a_m$.

Theorem 1.7.4. *Let $\{N_i : i \in I\}$ be a family of normal subgroups of a group G. Then the following statements are equivalent.*
(i) G is the internal weak direct product of $\{N_i : i \in I\}$.
(ii) $\forall x \in G\setminus\{e\}, \exists! i_1,\ldots,i_m \in I, \exists! a_{i_1} \in N_{i_1}\setminus\{e\},\ldots,a_{i_m} \in N_{i_m}\setminus\{e\}$, $x = a_{i_1}\ldots a_{i_m}$.

Corollary 1.7.5. [Internal Direct Product] *Let G be a group. Suppose that A and B are normal subgroups of G such that*
1. $A \cap B = \{e\}$ 2. $AB = G$ and 3. $\forall a \in A, b \in B, ab = ba$.
Then $G \cong A \times B$. In this case, we say that G is the **internal direct product of A and B.**

An application of the first isomorphism theorem with the natural map gives the next results.

Theorem 1.7.6. *Let $\{G_i : i \in I\}$ be a family of groups, and for $i \in I$, let N_i be normal in G_i. Then $\prod_{i\in I} N_i$ is normal in $\prod_{i\in I} G_i$ and $\prod_{i\in I} G_i/\prod_{i\in I} N_i \cong \prod_{i\in I}(G_i/N_i)$. Similarly, $\prod_{i\in I}^w N_i$ is normal in $\prod_{i\in I}^w G_i$ and $\prod_{i\in I}^w G_i/\prod_{i\in I}^w N_i \cong \prod_{i\in I}^w(G_i/N_i)$.*

Corollary 1.7.7. *Let $G_1,\ldots,G_m$ be abelian groups, and for $1 \le j \le m$, let H_j be a subgroup of G_j. Then $(G_1 \oplus \cdots \oplus G_n)/(H_1 \oplus \cdots \oplus H_n) \cong (G_1/H_1) \oplus \cdots \oplus (G_n/H_n)$.*

Next, we study the structure of a finite abelian group. Results on elements of finite order are presented in the next theorem and we recall the Chinese Remainder Theorem in a group theoretic language. Their proof are routine and left as exercises.

Theorem 1.7.8. *Let A be an abelian group and $n \in \mathbb{N}$. Then the following statements hold.*
1. *The mapping $\varphi_n : A \to A$ defined by $\varphi_n(a) = a^n$ is a group homomorphism.*
2. *$A^n = \{a^n : a \in A\} = \operatorname{im}\varphi_n$ is a subgroup of A.*
3. *$A(n) = \{a \in A : a^n = e\} = \ker\varphi_n$ is a subgroup of A.*
4. *$\tau(A) = \{a \in A : \exists k \in \mathbb{N}, a^k = e\} = \bigcup_{n\in\mathbb{N}} A(n)$ is a subgroup of A. It is called the* **torsion subgroup of A.**

Theorem 1.7.9. [Chinese Remainder Theorem] *Suppose $m_1,\ldots,m_k$ are pairwise relatively prime (i.e., if $i \ne j$, then $\gcd(m_i, m_j) = 1$), and let $n_1,\ldots,n_k$ be any integers. Then there exists a unique integer v modulo $m = m_1\ldots m_k$ such that*

$$v \equiv n_i \mod m_i$$

for all $i = 1,\ldots,k$.

Remark. The Chinese remainder theorem may be restated as: if $m_1, \ldots, m_k$ are pairwise relatively primes and $m = m_1 \ldots m_k$, then

$$\mathbb{Z}_m \cong \mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_k}.$$

Corollary 1.7.10. *If $m = p_1^{n_1} \ldots p_k^{n_k}$ where $n_1, \ldots, n_k \in \mathbb{N}$ and $p_1, \ldots, p_k$ are distinct primes, then*

$$\mathbb{Z}_m \cong \mathbb{Z}_{p_1^{n_1}} \oplus \cdots \oplus \mathbb{Z}_{p_k^{n_k}}.$$

We shall use the Chinese remainder theorem to prove our first lemma.

Lemma 1.7.11. *Let $g \in G$ be an element such that $g^m = 1$ where $m = m_1 \ldots m_k$ and m_i are pairwise relatively prime. Then there exist unique elements $g_1, \ldots, g_k$ of G satisfying the following conditions:*
(a) $g_i^{m_i} = e$ for all $i \in \{1, \ldots, k\}$
(b) $g_1, \ldots, g_k$ commute pairwise and
(c) $g = g_1 \ldots g_k$.

Proof. First we show existence, then uniqueness. The g_i are in fact powers of g.
Existence: By the Chinese Remainder Theorem, choose $v_1, \ldots, v_k$ satisfying

$$v_i \equiv 1 \mod m_i \text{ and } v_i \equiv 0 \mod m/m_i.$$

For each i, let $v_i = \lambda_i(m/m_i)$ for some $\lambda_i \in \mathbb{Z}$ and set $g_i = g^{v_i}$. Then we have
(i) $g_i^{m_i} = g^{v_i m_i} = g^{\lambda_i (m/m_i) m_i} = g^{\lambda_i m}$ and
(ii) $g_1, \ldots, g_k$ are powers of g and hence commute pairwise.
(iii) Note that $v_1 + \cdots + v_k - 1 \equiv 0 \mod m_i$ for $i = 1, 2, \ldots, k$, that is, $m_i | (v_1 + \cdots + v_k - 1)$. Since $m_1, \ldots, m_k$ are pairwise relatively prime, $v_1 + \cdots + v_k \equiv 1 \mod m_1 \ldots m_k$, so

$$g_1 \ldots g_k = g^{v_1} \ldots g^{v_k} = g^{v_1 + \cdots + v_k} = g.$$

Uniqueness: Suppose $g = g_1 \ldots g_k$ where $g = g_1 \ldots g_k$ and $g_1, \ldots, g_k$ satisfy (i), (ii) and (iii). Then for each i,

$$g^{v_i} = g_1^{v_i} \ldots g_k^{v_i} = g_i,$$

that is, $g_i = g^{v_i}$ is the only possibility. □

Example 1.7.1. Consider $g \in G$ with $g^{60} = e$. Then $m = 3 \cdot 4 \cdot 5$, so $v_1 = 45, v_2 = 40$ and $v_3 = 36$. Thus, $g = g^{45} g^{40} g^{36} = g_1 g_2 g_3$.

In case g has order $m = p_1^{a_1} \ldots p_k^{a_k}$ where $m_i = p_i^{a_i}$ and $p_1, \ldots, p_k$ are distinct primes, g_i is called the **p_i-primary part of** g. We have the first step of our decomposition.

Theorem 1.7.12. *Let A be a finite abelian group of order $m = m_1 \ldots m_k$ where $m_1, \ldots, m_k$ are pairwise relatively prime. For each $i \in \{1, 2, \ldots, k\}$, let $A_i = \{g \in A : g^{m_i} = e\}$. Then $A_1 \times \cdots \times A_k \cong A$. Moreover, $|A_i| = m_i$ for all i.*

Proof. Define $\phi : A_1 \times \cdots \times A_k \to A$ by $\phi(g_1, \ldots, g_k) = g_1 \ldots g_k$. Clearly, ϕ is a group homomorphism. By Lemma 1.7.11, ϕ is 1-1 and onto. Finally, $m_1 \ldots m_k = |A| = |A_1 \times \cdots \times A_k| = |A_1| \ldots |A_k|$. Let

$$\begin{aligned} m_1 &= p_{11}^{u_{11}} \ldots p_{1r_1}^{u_{1r_1}} \\ m_2 &= p_{21}^{u_{21}} \ldots p_{2r_2}^{u_{2r_2}} \\ &\vdots \\ m_k &= p_{k1}^{u_{k1}} \ldots p_{kr_k}^{u_{kr_k}}. \end{aligned}$$

Here, the p_{ij} are distinct primes and $u_{ij} \geq 1$. Since every element of A_i satisfies $g^{m_i} = e$, $|A_i|$ involves only those primes occuring in m_i by Cauchy theorem. This forces $|A_i| = m_i$ for all i. □

If $m = p_1^{a_1} \dots p_k^{a_k}$, then A_i in Theorem 1.7.12 is just the *Sylow p_i-subgroup of A*. It now suffices to studied each factor A_i which is a Sylow p_i-subgroup of A. To investigate them, we shall need the following definition.

The positive integer n is an **exponent** for a group G if for each $g \in G$, $g^n = 1$. In this case, G is said to have finite exponent and the least such n is called *the* **exponent of** G. For example, 12 is *an* exponent of $\mathbb{Z}_6$ but 6 is the exponent of $\mathbb{Z}_6$. We denote the exponent of a group G (if exists) by $\exp G$. Note that $\mathbb{Z}$ has no exponent and $\exp G$ divides $|G|$. The exponent of a finite abelian p-group gives information on its structure as follows.

Theorem 1.7.13. *Let A be an abelian group with $|A| = p^u$ where p is a prime. Suppose A has the exponent p, (that is, $a^p = e$ for all $a \in A$). Then*

$$A \cong \underbrace{\mathbb{Z}_p \times \cdots \times \mathbb{Z}_p}_{u \text{ copies}} = (\mathbb{Z}_p)^u.$$

Proof. Note that if $a \in A \setminus \{e\}$, $\langle a \rangle \cong \mathbb{Z}_p$. For any subset $\{a_1, \dots, a_k\} \subseteq A \setminus \{e\}$, we can define a group homomorphism $\theta : \langle a_1 \rangle \times \cdots \times \langle a_k \rangle \to A$ by

$$\theta(a_1^{i_1}, \dots, a_k^{i_k}) = a_1^{i_1} \dots a_k^{i_k}.$$

We shall say that $a_1, \dots, a_k$ are "linearly independent" if θ is 1-1. This is equivalent to saying: If $\theta(a_1^{i_1}, \dots, a_k^{i_k}) = a_1^{i_1} \dots a_k^{i_k} = e$, then $i_1 \equiv \dots \equiv i_k \equiv 0 \mod p$.

Now there exists subsets of $A \setminus \{e\}$ for which θ is 1-1, e.g., the empty set, a singleton set. Choose a subset $\{a_1, \dots, a_k\}$ for which θ is 1-1 and k is as large as possible. We claim that in this case

$$\langle a_1 \rangle \times \cdots \times \langle a_k \rangle \xrightarrow{\theta} A$$

is onto, and hence is an isomorphism.

To see that θ is onto, let $b \in A$. If $b = e$, clearly $b \in \operatorname{im} \theta$. If $b \neq e$, consider

$$\langle a_1 \rangle \times \cdots \times \langle a_k \rangle \times \langle b \rangle \xrightarrow{\bar{\theta}} A.$$

By the maximal choice of $\{a_1, \dots, a_k\}$, $\bar{\theta}$ is not 1-1. Thus,

$$e = \bar{\theta}(a_1^{i_1}, \dots, a_k^{i_k}, b^j) = a_1^{i_1} \dots a_k^{i_k} b^j$$

and $j \not\equiv 0 \mod p$ since θ is not 1-1. Hence, there is a λ such that $j\lambda \equiv 1 \mod p$, so

$$e = a_1^{\lambda i_1} \dots a_k^{\lambda i_k} b^{j\lambda} = a_1^{\lambda i_1} \dots a_k^{\lambda i_k} b$$

which implies $b = a_1^{-\lambda i_1} \dots a_k^{-\lambda i_k} \in \operatorname{im} \theta$. Therefore, θ is onto as claimed, and we have an isomorphism

$$\langle a_1 \rangle \times \cdots \times \langle a_k \rangle \xrightarrow[\cong]{\theta} A\ .$$

Thus, $p^k = |\langle a_1 \rangle| \dots |\langle a_k \rangle| = |\langle a_1 \rangle \times \cdots \times \langle a_k \rangle| = |A| = p^u$, so $k = u$ and the theorem is proved. □

Remark. If we write A in Theorem 1.7.13 additively, we see that it is just a vector space over the field $\mathbb{Z}_p$. Since A is finite, it is a finite dimensional vector space over $\mathbb{Z}_p$. All we were doing in Theorem 1.7.13 is finding a basis for A as a vector space over $\mathbb{Z}_p$.

Theorem 1.7.14. [Burnside Basis Theorem for Abelian p-groups] *Suppose A is an abelian group of exponent p^k where p is a prime. Let $A^p = \{a^p : a \in A\}$. If H is a subgroup of A and $HA^p = A$, then $H = A$. Equivalently, if the cosets $A^pa_1, \ldots, A^pa_k$ of A/A^p generate A/A^p, then $a_1, \ldots, a_k$ generate A.*

Proof. Observe that $HA^p = A$ implies $H^pA^{p^2} = A^p$, so

$$A = HA^p = H(H^pA^{p^2}) = HA^{p^2}.$$

Also, $HA^{p^2} = A$ implies $H^pA^{p^3} = A^p$, so

$$A = HA^p = H(H^pA^{p^3}) = HA^{p^3}.$$

Continue inductively, we have $A = HA^{p^r}$ for all r. But $A^{p^k} = \{e\}$, so $A = HA^{p^k} = H$. This completes the proof. □

Theorem 1.7.15. *Let A be a finite abelian group of exponent p where p is a prime, and let H be a subgroup of A. Then there exists a subgroup K of A such that $H \cap K = \{e\}$ and $HK = A$. In other words, A is the internal direct product of H and K.*

Proof. Let K be a subgroup of A satisfying $H \cap K = \{e\}$ and among all subgroups K of A satisfying $H \cap K = \{e\}$, K is as large as possible. We claim that $HK = A$ which proves the theorem.

For, suppose conversely that $a \in A$ and $a \notin HK$. Then $H \cap \langle K, a\rangle \neq \{e\}$ by the maximal choice of K, so there is a nontrivial element

$$e \neq h = ka^i \in H \cap \langle K, a\rangle \text{ where } h \in H \in \text{ and } k \in K.$$

If $p|i$, $a^i = e$ and $h = k \in H \cap K = \{e\}$, a contradiction. If $p \nmid i$, there is a λ with $a^{i\lambda} = a$ ($i\lambda \equiv 1 \mod p$) and then $a = a^{i\lambda} = (hk^{-1})^\lambda \in HK$, a contradiction. Hence, $HK = A$ as required. □

Remark. As with Theorem 1.7.13, the above theorem can be regarded as a statement about vector spaces over $\mathbb{Z}_p$ as follows: If V is a finite dimensional vector space over $\mathbb{Z}_p$ and U is a subspace, then there is a subspace W such that $V = U \oplus W$.

We are now ready to prove the structure theorem for a finite abelian p-group.

Theorem 1.7.16. *Let A be a finite abelian p-group. Then A is (isomorphic to) a direct product of cyclic groups.*

Proof. We use induction on $|A|$. If $|A| = 1$, the result is clear. Now suppose $|A| = p^u > 1$. We assume inductively that any p-group where order is less than p^u is a direct product of cyclic groups. Consider the group

$$A^p = \{a^p : a \in A\}.$$

Claim $A \neq A^p$. For, suppose $A = A^p$. Then

$$A = A^p = A^{p^2} = \cdots = A^{p^u} = \{e\}$$

since $|A| = p^u$. As $|A| > 1$, we must have $A \neq A^p$.

Thus, $A > A^p$, so $|A| > |A^p|$ and by the inductive hypothesis, A^p is the (internal) direct product of cyclic subgroup. But every element of A^p has the form a^p. Hence, so do the generators of these cyclic factors, so there exist $a_1, \ldots, a_k \in A$ such that

$$A^p = \langle a_1^p\rangle \times \cdots \times \langle a_k^p\rangle.$$

More precisely, the map $(a_1^{pi_1}, \ldots, a_k^{pi_k}) \overset{\theta_0}{\longmapsto} a_1^{pi_1} \ldots a_k^{pi_k}$ is an isomorphism.

Now let $H = \langle a_1, \ldots, a_k \rangle$ be the subgroup of A generated by $a_1, \ldots, a_k$. We claim that it is the (internal) direct product of the groups $\langle a_1 \rangle, \ldots, \langle a_k \rangle$, or that

$$\begin{aligned} \langle a_1 \rangle \times \cdots \times \langle a_k \rangle &\overset{\theta}{\to} H \\ (a_1^{i_1}, \ldots, a_k^{i_k}) &\mapsto a_1^{i_1} \ldots a_k^{i_k} \end{aligned}$$

is an isomorphism. Since $H = \langle a_1, \ldots, a_k \rangle$, θ is onto. To see that θ is 1-1, suppose

$$e = \theta(a_1^{i_1}, \ldots, a_k^{i_k}) = a_1^{i_1} \ldots a_k^{i_k}. \tag{1.7.1}$$

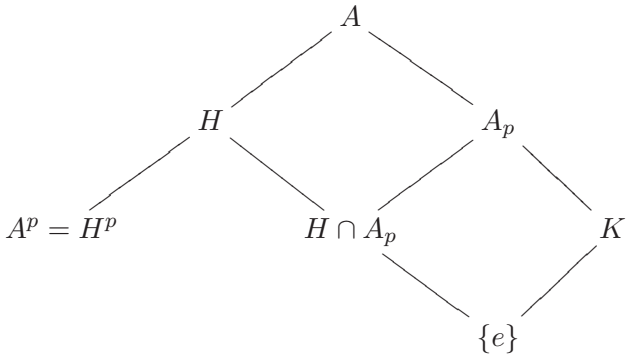
Then $e = (a_1^{i_1} \ldots a_k^{i_k})^p = a_1^{pi_1} \ldots a_k^{pi_k}$, so

$$a_1^{pi_1} = \cdots = a_k^{pi_k} = e$$

since the map θ_0 above is 1-1. Consider the integers $i_1, \ldots, i_k$. If $p \nmid i_t$ for some of these integers, then $a_t^{pi_t} = e$ implies $a_t^p = e$. For, $(a_t^{pi_t})^{\lambda_t} = (a_t^{\lambda_t i_t})^p = a_t^p$, where $\lambda_t i_t \equiv 1 \mod |a_t|$. Thus, $p \mid i_1, \ldots, i_k$, so $i_1 = pj_1, \ldots, i_k = pj_k$. But then (1.7.1) becomes

$$e = \theta(a_1^{pj_1}, \ldots, a_k^{pj_k}) = a_1^{pj_1} \ldots a_k^{pj_k},$$

so $a_1^{pj_1} = \cdots = a_k^{pj_k} = e$ since θ_0 is 1-1. Hence, θ is 1-1.



Next, let $A_p = \{a \in A : a^p = e\}$. Then A_p is a finite group of exponent p and contains $H \cap A_p$ as a subgroup. Therefore, $H \cap A_p$ has a component in A_p by Theorem 1.7.15. More precisely, there is a subgroup K of A_p such that

(a) $(H \cap A_p) \cap K = \{e\}$ and (b) $(H \cap A_p)K = A_p$.

Note that since K is a group of exponent p, K is a direct product of copies of $\mathbb{Z}_p$ by Theorem 1.7.13. Finally, we claim that

(I) $H \cap K = \{e\}$ and (II) $HK = A$.

They implies that A is a direct product of H and K which are both direct products of cyclic groups.

(I) Suppose $H \cap K \neq \{e\}$. Thus, there is some $x \in H \cap K$ with $x \neq e$ and $x^p = e$. But then $x \in A_p$, so $(H \cap A_p) \cap K = (H \cap K) \cap A_p \neq \{e\}$, contradicting (i) above.

(II) Suppose $a \in A$. Then $a^p \in A^p = \langle a_1^p \rangle \times \cdots \times \langle a_k^p \rangle$, so $a^p = a_1^{pi_1} \ldots a_k^{pi_k} = (a_1^{i_1} \ldots a_k^{i_k})^p = b^p$ where $b = a_1^{i_1} \ldots a_k^{i_k} \in H = \langle a_1, \ldots, a_k \rangle$. Thus, $b^{-1}a \in A_p = (H \cap A_p)K \subseteq HK$ by (ii) above. Hence, $a = b(b^{-1}a) \in HK$ and $A = HK$ as required.

This completes the proof. □

In addition, the above decomposition is unique. Hence, we are able to count the number of non-isomorphic abelian p-groups.

Theorem 1.7.17. *Suppose*

$$A = \underbrace{\mathbb{Z}_p \times \ldots \mathbb{Z}_p}_{u_1 \text{ copies}} \times \underbrace{\mathbb{Z}_{p^2} \times \cdots \times \mathbb{Z}_{p^2}}_{u_2 \text{ copies}} \times \cdots \times \underbrace{\mathbb{Z}_{p^m} \times \cdots \times \mathbb{Z}_{p^m}}_{u_m \text{ copies}}$$

is isomorphic to

$$B = \underbrace{\mathbb{Z}_p \times \ldots \mathbb{Z}_p}_{v_1 \text{ copies}} \times \underbrace{\mathbb{Z}_{p^2} \times \cdots \times \mathbb{Z}_{p^2}}_{v_2 \text{ copies}} \times \cdots \times \underbrace{\mathbb{Z}_{p^m} \times \cdots \times \mathbb{Z}_{p^m}}_{v_m \text{ copies}}$$

where $u_i, v_i \geq 1$. Then $u_i = v_i$ for all $i = 1, \ldots, m$. In other words, the orders and multiplicities of the factors in a decomposition of a finite abelian p-group uniquely determine the group up to isomorphism.

Proof. Since $A \cong B$, it follows that for any positive integer n,

$$\#\text{ of solutions of } x^n = e \text{ in } A \quad = \quad \#\text{ of solutions of } x^n = e \text{ in } B.$$

Consider the following table.

n	# of solutions of $x^n = e$ in A	# of solutions of $x^n = e$ in B
p	$p^{u_1+u_2+u_3+\cdots+u_m}$	$p^{v_1+v_2+v_3+\cdots+v_m}$
p^2	$p^{u_1+2u_2+2u_3+\cdots+2u_m}$	$p^{v_1+2v_2+2v_3+\cdots+2v_m}$
p^3	$p^{u_1+2u_2+3u_3+\cdots+3u_m}$	$p^{v_1+2v_2+3v_3+\cdots+3v_m}$
$\vdots$	$\vdots$	$\vdots$
p^{m-1}	$p^{u_1+2u_2+3u_3+\cdots+(m-1)u_{m-1}+(m-1)u_m}$	$p^{v_1+v_2+\cdots+(m-1)v_{m-1}+(m-1)v_m}$
p^m	$p^{u_1+2u_2+3u_3+\cdots+(m-1)u_{m-1}+mu_m}$	$p^{v_1+v_2+\cdots+(m-1)v_{m-1}+mv_m}$

Then we have

$$\begin{aligned}
u_1 + u_2 + u_3 + \cdots + u_m &= v_1 + v_2 + v_3 + \cdots + v_m \\
u_1 + 2u_2 + 2u_3 + \cdots + 2u_m &= v_1 + 2v_2 + 2v_3 + \cdots + 2v_m \\
u_1 + 2u_2 + 3u_3 + \cdots + 3u_m &= v_1 + 2v_2 + 3v_3 + \cdots + 3v_m \\
&\vdots \\
u_1 + 2u_2 + 3u_3 + \cdots + (m-1)u_{m-1} + (m-1)u_m &= \\
&\quad v_1 + 2v_2 + 3v_3 + \cdots + (m-1)v_{m-1} + (m-1)v_m \\
u_1 + 2u_2 + 3u_3 + \cdots + (m-1)u_{m-1} + mu_m &= \\
&\quad v_1 + 2v_2 + 3v_3 + \cdots + (m-1)v_{m-1} + mv_m.
\end{aligned}$$

It is easy to see that the above equations force $u_1 = v_1, u_2 = v_2, \ldots, u_m = v_m$ as required. □

Theorem 1.7.18. *Let p be any prime and n a positive integer. Then*

$$\{r_1 \leq r_2 \leq \cdots \leq r_k\} \longleftrightarrow \mathbb{Z}_{p^{r_1}} \times \mathbb{Z}_{p^{r_2}} \times \cdots \times \mathbb{Z}_{p^{r_k}}$$

defines a 1-1 correspondence between partitions of n and isomorphism classes of abelian groups of order p^n. In particular, the number of isomorphism classes of abelian groups of order p^n is the number of partitions of n.

Examples 1.7.2. 1. Abelian groups of order p^3.

partitions of 3	corresponding abelian groups
$\{1,1,1\}$	$\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$
$\{1,2\}$	$\mathbb{Z}_p \times \mathbb{Z}_{p^2}$
$\{3\}$	$\mathbb{Z}_{p^3}$

2. Abelian groups of order p^5.

partitions of 5	corresponding abelian groups
$\{1,1,1,1,1\}$	$\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$
$\{1,1,1,2\}$	$\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_{p^2}$
$\{1,2,2\}$	$\mathbb{Z}_p \times \mathbb{Z}_{p^2} \times \mathbb{Z}_{p^2}$
$\{1,1,3\}$	$\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_{p^3}$
$\{2,3\}$	$\mathbb{Z}_{p^2} \times \mathbb{Z}_{p^3}$
$\{1,4\}$	$\mathbb{Z}_p \times \mathbb{Z}_{p^4}$
$\{5\}$	$\mathbb{Z}_{p^5}$

Suppose A is a finite abelian group of order $p_1^{a_1} \dots p_k^{a_k}$ where $p_1, \dots, p_k$ are distinct primes. Let

$$A_i = \{g \in A : g^{p_i^{a_i}} = e\}.$$

By Theorem 1.7.12, $A \cong A_1 \times \cdots \times A_k$ where $|A_i| = p_i^{a_i}$. Since each A_i is a direct product of cyclic group by Theorem 1.7.17, this yields the following theorem.

Theorem 1.7.19. *A finite abelian group is (isomorphic to) a direct product of cyclic groups.*

Corollary 1.7.20. *If m is a square free integer, then every abelian group of order m is cyclic.*

Proof. Assume that an abelian group A is of order $m = p_1 \dots p_r$ where p_i are distinct primes. Then

$$A \cong \mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_r} \cong \mathbb{Z}_m$$

by Theorem 1.7.12 and the Chinese remainder theorem, respectively. □

Let $A \cong A_1 \times \cdots \times A_k$ as above. It is clear that A_i is the unique largest p_i-subgroup of A. Moreover, if B is finite abelian, and $B \cong B_1 \times \cdots \times B_k$ where B_i is a p_i-group, then

$$A \cong B \Leftrightarrow (A_1 \cong B_1 \wedge \ldots \wedge A_k \cong B_k).$$

Recall that if m and n are relatively prime then $\mathbb{Z}_m \times \mathbb{Z}_n \cong \mathbb{Z}_{mn}$. It follows that if $n = p_1^{u_1} \cdots p_k^{u_k}$ where $p_1, \dots, p_k$ are distinct primes, then

$$\mathbb{Z}_{p_1^{u_1}} \times \cdots \times \mathbb{Z}_{p_k^{u_k}} \cong \mathbb{Z}_n. \tag{1.7.2}$$

This gives rise to a second way of writing a finite abelian group A as a direct product of cyclic groups. Namely, let $p_1, \dots, p_k$ be the primes dividing $|A|$, and let

$$A = A_1 \times \cdots \times A_k,$$

where A_i is the p_i-primary part of A. Express each A_i as a direct product of cyclic factors and assume that t is the largest number of factors occurring in any A_i. Write

$$\begin{aligned}
A_1 &= \mathbb{Z}_{p_1^{v_{11}}} \times \mathbb{Z}_{p_1^{v_{12}}} \times \cdots \times \mathbb{Z}_{p_1^{v_{1t}}} \\
A_2 &= \mathbb{Z}_{p_2^{v_{21}}} \times \mathbb{Z}_{p_2^{v_{22}}} \times \cdots \times \mathbb{Z}_{p_2^{v_{2t}}} \\
&\vdots \\
A_k &= \mathbb{Z}_{p_k^{v_{k1}}} \times \mathbb{Z}_{p_k^{v_{k2}}} \times \cdots \times \mathbb{Z}_{p_k^{v_{kt}}}
\end{aligned}$$

where

$$0 \leq v_{i1} \leq v_{i2} \leq \cdots \leq v_{it} \tag{1.7.3}$$

and we have allowed (for notational convenience) some v_{ij} to be zero. Let

$$\begin{aligned} n_1 &= p_1^{v_{11}} p_2^{v_{21}} \cdots p_k^{v_{k1}} \\ n_2 &= p_1^{v_{12}} p_2^{v_{22}} \cdots p_k^{v_{k2}} \\ &\vdots \\ n_i &= p_1^{v_{1i}} p_2^{v_{2i}} \cdots p_k^{v_{ki}} \\ &\vdots \\ n_t &= p_1^{v_{1t}} p_2^{v_{2t}} \cdots p_k^{v_{kt}}. \end{aligned}$$

Condition (1.7.3) guarantees that $n_1 \mid n_2, n_2 \mid n_3, \ldots, n_{t-1} \mid n_t$. Then, using (1.7.2) gives

$$\begin{aligned} A &= A_1 \times \cdots \times A_k \\ &= (\mathbb{Z}_{p_1^{v_{11}}} \times \cdots \times \mathbb{Z}_{p_1^{v_{1t}}}) \times (\mathbb{Z}_{p_1^{v_{12}}} \times \cdots \times \mathbb{Z}_{p_k^{v_{k2}}}) \times \cdots \times (\mathbb{Z}_{p_1^{v_{1t}}} \times \cdots \times \mathbb{Z}_{p_k^{v_{kt}}}) \\ &\cong \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_t}. \end{aligned}$$

The integers n_j are completely determined by the decomposition of A into a direct of cyclic p_i-groups. Conversely, given that

$$A = \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_t}.$$

where $n_1 \mid n_2, n_2 \mid n_3, \ldots, n_{t-1} \mid n_t$, the decomposition of A into a direct product of cyclic p_i-groups is completely determined. Thus, we have:

Theorem 1.7.21. [Structure Theorem for Finite Abelian Groups] *Let A be a finite abelian group. Then there exist integers $n_1, \ldots, n_t > 1$ such that $n_1 \mid n_2, n_2 \mid n_3, \ldots, n_{t-1} \mid n_t$ and*

$$A \cong \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_t},$$

where these integers are uniquely defined by A. More precisely, if $m_1, \ldots, m_s$ are positive integers greater than 1 such that $m_1 \mid m_2, m_2 \mid m_3, \ldots, m_{s-1} \mid m_s$, and

$$A \cong \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_t} \cong \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \cdots \times \mathbb{Z}_{m_s},$$

then $t = s$, and $n_1 = m_1, \ldots, n_t = m_t$.

Example 1.7.3. Find all non-isomorphic abelian groups of order:
1. 6 2. 12 3. 27 4. 500.

Solution. By Example 1.7.2 and Theorem 1.7.21. We have the following answers.
$6 = 2 \cdot 3$: $\mathbb{Z}_2 \times \mathbb{Z}_3 \cong \mathbb{Z}_6$,
$12 = 2^2 \cdot 3$: $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3 \cong \mathbb{Z}_2 \times \mathbb{Z}_6$ and $\mathbb{Z}_{2^2} \times \mathbb{Z}_3 \cong \mathbb{Z}_{12}$,
$27 = 3^3$: $\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_3$, $\mathbb{Z}_3 \times \mathbb{Z}_{3^2}$ and $\mathbb{Z}_{3^3}$,
$500 = 2^2 \cdot 5^3$: $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_5 \cong \mathbb{Z}_2 \times \mathbb{Z}_{10} \times \mathbb{Z}_{10}$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_5 \times \mathbb{Z}_{5^2} \cong \mathbb{Z}_{10} \times \mathbb{Z}_{50}$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_{5^3} \cong \mathbb{Z}_2 \times \mathbb{Z}_{250}$, $\mathbb{Z}_{2^2} \times \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_5 \cong \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_{20}$, $\mathbb{Z}_{2^2} \times \mathbb{Z}_5 \times \mathbb{Z}_{5^2} \cong \mathbb{Z}_5 \times \mathbb{Z}_{100}$ and $\mathbb{Z}_{2^2} \times \mathbb{Z}_{5^3} \cong \mathbb{Z}_{500}$. □

By Cauchy's theorem, if p is a divisor of the order of a group G, then G has a subgroup of order p. We also see that A_4 is a group of order 12 but it has no subgroup of order six. Hence, it may not hold that G will have a subgroup of order m when m is a divisor of $|G|$. However, if G is an abelian group, we have our final results.

Corollary 1.7.22. *Let A be a finite abelian group. If m divides the order of A, then A has a subgroup of order m.*

Proof. Write $A = \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_t}$ as in the above theorem. Then $|A| = n_1 n_2 \dots n_t$. Since m divides $|A|$, $m = l_1 l_2 \dots l_t$ with $l_i \mid n_i$ for all $i \in \{1, 2, \dots, t\}$. Then $(n_i/l_i)\mathbb{Z}_{n_i}$ is a subgroup of $\mathbb{Z}_{n_i}$ of order l_i for all i. Thus,

$$(n_1/l_1)\mathbb{Z}_{n_1} \times (n_2/l_2)\mathbb{Z}_{n_2} \times \cdots \times (n_t/l_t)\mathbb{Z}_{n_t}$$

is a subgroup of A of order $l_1 l_2 \dots l_t = m$ as desired. □

Corollary 1.7.23. *Let A be a finite abelian group. Then there exists $g \in A$ such that the order of g is the exponent of A.*

Proof. By Theorem 1.7.21, there exist positive integers $n_1, n_2, \dots, n_t \geq 1$ such that $n_1 \mid n_2 \mid \cdots \mid n_t$ and

$$A \cong \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_t}.$$

Thus, $\exp A = n_t$ and $(0, 0, \dots, 0, 1)$ in the rightmost group has order n_t. □

Remark. The above corollary is false if A is non-abelian. For example, the exponent of S_3 is 6, however S_3 contains no elements of order 6.

Exercises 1.7.

1. Suppose G_1 and G_2 are finite groups of relatively prime orders. Show that every subgroup of $G_1 \times G_2$ is of the form $H_1 \times H_2$ for some subgroups H_1 and H_2 of G_1 and G_2, respectively.
2. Let G_1 and G_2 be simple groups. Show that every nontrivial normal subgroup of $G = G_1 \times G_2$ is isomorphic to either G_1 or G_2.
3. Proof Theorem 1.7.8.
4. Find the order of torsion subgroup of $\mathbb{Z}_4 \times \mathbb{Z} \times \mathbb{Z}$ and of $\mathbb{Z}_{12} \times \mathbb{Z}_{12} \times \mathbb{Z}$.
5. Find the torsion subgroup of the multiplicative group $\mathbb{R}^*$.
6. Let G be an abelian group of order 72.
   (a) Can you say how many subgroups of order eight G has? Why?
   (b) Can you say how many subgroups of order four G has? Why?
7. Show that a finite abelian group is not cyclic if and only if it contains a subgroup isomorphic to $\mathbb{Z}_p \times \mathbb{Z}_p$ for some prime p.
8. Find the exponent of the following groups: (a) $\mathbb{Z}_6 \times \mathbb{Z}_9$ (b) $\mathbb{Z}_5 \times S_4$.
9. Find all non-isomorphic abelian groups of order (a) 35 (b) 48 (c) 360.
10. List (up to isomorphism) all abelian groups of order 108 and express each in the following two ways:
    (a) As a direct sum of cyclic groups of prime power order.
    (b) As a direct sum of cyclic groups of order $d_1, d_2, \dots, d_k$ where $d_i \mid d_{i+1}$ for $i = 1, 2, \dots, k-1$.
11. List all groups of order 99 and of order 1225 up to isomorphism. (*Hint*. Show that they must be abelian.)
12. Let G be a finite group. Prove the following statements.
    (a) The exponent of G divides its order $|G|$.
    (b) If H is a subgroup of G, then the exponent of H divides the exponent of G.
    (c) If G is cyclic, then $\exp G = |G|$.
13. Let $G = S_3 \times \mathbb{Z}_4$.
    (a) Find $\exp G$. (b) Determine *all* Sylow 2-subgroups and Sylow 3-subgroups of G.
14. Let G be a group of order $2156 = 2^2 \cdot 7^2 \cdot 11$.
    (a) If G is abelian, list *all* G up to isomorphism. (b) Prove that G cannot be simple.
15. List all finite groups G which have the property: $\forall g, h \in G$, g is a power of h or h is a power of g.

Project 9 (Characters of a group). Let $\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle in the complex plane. It is a subgroup of $\mathbb{C}^\times$. A **character** of a finite group G is a homomorphism from G to $\mathbb{S}^1$. The character sending G to 1 is called the **trivial character**. Let G be a finite group. Prove the following statements.

(a) All characters of G form an abelian group under pointwise multiplication, called the **dual group of** G and denoted by $\widehat{G}$.
(b) Prove that if G is a finite cyclic group, then G and $\widehat{G}$ are isomorphic.
(c) If G is a finite abelian group, then G is isomorphic to its dual group $\widehat{G}$. (*Hint*. Use (b) and Theorem 1.7.19.)

This concludes the basic theory of groups. More advanced group theory will be studied in Chapter 3.

2 | Rings and Fields

Rings and fields are the most common algebraic structures for students. They have learned addition together with multiplication since elementary schools. The abstract treatments using groups are presented in the first section. Ideals and factorizations are discussed in details. Finally, we talk about polynomials over a ring and which will be used in a construction of field extensions.

2.1 Basic Concepts

2.1.1 Rings

A **ring** is a triple $(R, +, \cdot)$ where

(R1) $(R, +)$ is an abelian group,

(R2) $(R, \cdot)$ is a semigroup, and

(R3) $+$ and $\cdot$ satisfy the distributive laws, namely,

$\forall a, b, c \in R, a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ and $\forall a, b, c \in R, (b + c) \cdot a = (b \cdot a) + (c \cdot a)$.

The binary operations $+$ and $\cdot$ are called the **addition** and the **multiplication** of the ring $(R, +, \cdot)$, respectively, 0, the identity of $(R, +)$ is called the **zero** of R, for $a, b \in R$, ab (juxtaposition) may denote $a \cdot b$ and $(R, +, \cdot)$ may be denoted by R. For $a \in R$, $-a$ is called the **additive inverse** of a in R.

A ring R is said to be **commutative** if $\forall a, b \in R, ab = ba$. If 1 is the identity of $(R, \cdot)$, then 1 is called the **identity** or **unity** of the ring R. If R contains an identity, i.e., $(R, \cdot)$ is a multiplicative monoid, then R is called a **ring with identity**.

Unless the contrary is explicitly stated "ring" will mean "ring with identity".

A subset S of a ring R is a **subring** if S is a subgroup of the additive group and also a *submonoid* of the multiplicative monoid of R. Clearly the intersection of any set of subrings of R is a subring. Hence, if A is a subset of R, we may define the **subring generated by** A to be the intersection of all subrings of R which contain A.

Examples 2.1.1 (Examples of rings)**.**

1. $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$ are commutative rings under usual addition and multiplication.
2. For $n \in \mathbb{N}$, $(\mathbb{Z}_n, +, \cdot)$ is a commutative ring.
3. Recall that for a nonempty set X and $P(X)$ the power set of X, we define $A \triangle B = (A \setminus B) \cup (B \setminus A)$ for all subsets A and B of X. Then $(P(X), \triangle, \cap)$ is a commutative ring with identity X.
4. If A is an abelian group, then $\mathrm{End}(A)$, the set of all homomorphisms on A, is a ring with the addition and multiplication are given by

$$(f + g)(a) = f(a) + g(a) \quad \text{and} \quad (f \cdot g)(a) = f(g(a))$$

for all $f, g \in \mathrm{End}(A)$ and $a \in A$. Its identity is the identity map.

5. If X is a topological space, then $C(X)$, the set of all continuous functions from X to $\mathbb{R}$, is a commutative ring with pairwise operations. That is,

$$(f+g)(x) = f(x) + g(x) \quad \text{and} \quad (f \cdot g)(x) = f(x)g(x)$$

for all $f, g \in C(X)$ and $x \in X$.
6. Let d be a square free integer. The set $\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} : a, b \in \mathbb{Z}\}$ is a subring of $\mathbb{C}$.

A number of elementary properties of rings are consequences of the fact that a ring is an abelian group relative to addition and a monoid relative to multiplication. For example, we have $-(a+b) = -a - b := (-a) + (-b)$ and if na is defined for $n \in \mathbb{Z}$ as before, then the rules for multiples (or powers) in an abelian group,

$$n(a+b) = na + nb, \quad (n+m)a = na + ma, \quad \text{and} \quad (nm)a = n(ma)$$

hold. There are also a number of simple consequences of the distributive laws which we now note. In the first place, induction on m and n give the generalization

$$\left(\sum_{i=1}^{m} a_i\right)\left(\sum_{j=1}^{n} b_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j.$$

We note next that

$$a0 = 0a = 0$$

for all $a \in R$; for we have $a0 = a(0+0) = a0 + a0$. Addition of $-a0$ gives $a0 = 0$. Similarly, $0a = 0$. We have the equation

$$0 = 0b = (a + (-a))b = ab + (-a)b,$$

which shows that $(-a)b = -ab$. Similarly, $a(-b) = -ab$; consequently

$$(-a)(-b) = -a(-b) = -(-ab) = ab.$$

If a and b commute, that is, $ab = ba$, then $a^m b^n = b^n a^m$. Also, by induction we can prove the **binomial theorem**

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + b^n$$

for all $n \in \mathbb{N}$.

Remark. In the case $1 = 0$ in a ring R, we have that

$$ab = (a \cdot 1)b = a(1 \cdot b) = a(0 \cdot b) = a \cdot 0 = 0 \quad \text{for all } a, b \in R.$$

A ring with this property is called a **zero ring**.

If R and S are rings, so is $R \times S$, with coordinatewise operations:

$$(r, s) + (r', s') = (r + r', s + s') \quad \text{and} \quad (r, s)(r', s') = (rr', ss').$$

Note that $(1_R, 1_S)$ is the identity of $R \times S$. More generally, if R_α is a family of rings, then $\prod_\alpha R_\alpha$ is a ring with coordinatewise operations.

Let R be a ring. An element $x \in R$ is said to be **invertible** or a **unit** if there is a $y \in R$ such that $xy = yx = 1$. In this case, y is called the **inverse of** x.

Remark. If x is invertible, its inverse is unique. The invertible elements of R form a group under multiplication, called the **group of units of** R and denoted by $\mathcal{U}(R)$ or $R^\times$.

A ring D is a **division ring** or **skew field** if every nonzero element of D is invertible. A commutative division ring is called a **field**.

Examples 2.1.2. 1. $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields. $\mathbb{Z}$ is not a field.
2. By Example 1.2.1 (5), we have $\forall n \in \mathbb{N}$, $\mathbb{Z}_n$ is a field $\Leftrightarrow$ n is a prime.

Example 2.1.3. Let $n \in \mathbb{N}$, R a ring and $M_n(R)$ the set of all $n \times n$ matrices over R. Then $(M_n(R), +, \cdot)$ is a ring under the usual addition and multiplication of matrices with unity I_n, the identity matrix. If $n > 1$, then $M_n(R)$ is not commutative. The group of invertible elements of $M_n(R)$ is called the **general linear group** and denoted by $\mathrm{GL}_n(R)$. For the case R is commutative, we can derive the determinant criterion for a matrix A to be invertible. We have the following results.

Theorem 2.1.1. *Let R be a commutative ring and $A \in M_n(R)$. Then*

$$A(\operatorname{adj} A) = (\det A)I_n = (\operatorname{adj} A)A.$$

In particular, A is invertible if and only if its determinant is invertible in R.

A noteworthy special case of the theorem is the next corollary.

Corollary 2.1.2. *If F is a field, $A \in M_n(F)$ is invertible if and only if* $\det A \neq 0$.

Some rings do not have the property that the product of two nonzero elements is always nonzero. If so, it leads to the cancellation property in the rings. Let R be a ring and $0 \neq a \in R$. a is called a **left [right] zero divisor** if $\exists b \in R \setminus \{0\}, ab = 0$ $[ba = 0]$, and it is called a **zero divisor** if it is both a left and a right zero divisor. R is **entire** if it possesses no zero divisors. A commutative entire ring is called an **integral domain**.

Examples 2.1.4. 1. $\mathbb{Z}$ is an integral domain which is not a field.
2. Every field is an integral domain.
3. $C([0,1])$ is not an integral domain.

Remark. Let R be a ring. Then we have
R is entire $\Leftrightarrow \forall a, b \in R\,[ab = 0 \Rightarrow (a = 0 \text{ or } b = 0)] \Leftrightarrow (R \setminus \{0\}, \cdot)$ is a cancellative semigroup.

Theorem 1.2.2 implies an important result on finiteness of an integral domain.

Theorem 2.1.3. *Every finite integral domain is a field.*

Proof. Let D be a finite integral domain. Then $(D \setminus \{0\}, \cdot)$ is a *finite* cancellative semigroup. By Theorem 1.2.2, $(D \setminus \{0\}, \cdot)$ is a group. Hence, if $a \in D$ and $a \neq 0$, then a has an inverse under $\cdot$. Since D is commutative, D is a field. □

2.1.2 Quaternions

In 1843, W. R. Hamilton constructed the first example of a division ring in which the commutative law of multiplication does not hold. This was an extension of the field of complex numbers, whose elements were quadruples of real numbers $(\alpha, \beta, \gamma, \delta)$ for which the usual addition and a multiplication were defined so that $1 = (1, 0, 0, 0)$ is the unit and $i = (0, 1, 0, 0), j = (0, 0, 1, 0)$, and $k = (0, 0, 0, 1)$ satisfy $i^2 = j^2 = k^2 = -1 = ijk$. Hamilton called his quadruples, **quaternions**. Previously, he had defined complex numbers as pairs of real numbers (α, β) with the product $(\alpha, \beta)(\gamma, \delta) = (\alpha\gamma - \beta\delta, \alpha\delta + \beta\gamma)$. Hamilton's discovery of quaternions led to a good deal of experimentation with other such "hypercomplex" number systems and eventually to a structure theory whose goal was to classify such systems. A good deal of important algebra thus evolved from the discovery of quaternions.

We shall not follow Hamilton's way of introducing quaternions. Instead we shall define this system as a certain subring of the ring $M_2(\mathbb{C})$ of 2×2 matrices with complex number entries. This will have the advantage of reducing the calculations to a single simple verification.

We consider the subset $\mathbb{H}$ of the ring $M_2(\mathbb{C})$ of complex 2×2 matrices that have the form

$$x = \begin{bmatrix} a & b \\ -\bar{b} & \bar{a} \end{bmatrix} = \begin{bmatrix} \alpha_0 + \alpha_1\sqrt{-1} & \alpha_2 + \alpha_3\sqrt{-1} \\ -\alpha_2 + \alpha_3\sqrt{-1} & \alpha_0 - \alpha_1\sqrt{-1} \end{bmatrix}, \quad \alpha_i \text{ real.} \tag{2.1.1}$$

We claim that $\mathbb{H}$ is a subring of $M_2(\mathbb{C})$. Since $\overline{a_1 - a_2} = \bar{a}_1 - \bar{a}_2$ for complex numbers, it is clear that $\mathbb{H}$ is closed under subtraction; hence $\mathbb{H}$ is a subgroup of the additive group of $M_2(\mathbb{C})$. We obtain the unit matrix by taking $a = 1, b = 0$ in (2.1.1). Hence, $1 \in \mathbb{H}$. Since

$$\begin{bmatrix} a & b \\ -\bar{b} & \bar{a} \end{bmatrix} \begin{bmatrix} c & d \\ -\bar{d} & \bar{c} \end{bmatrix} = \begin{bmatrix} ac - b\bar{d} & ad + b\bar{c} \\ -\bar{b}c - \bar{a}\bar{d} & -\bar{b}d + \bar{a}\bar{c} \end{bmatrix}$$

and $\overline{a_1 a_2} = \bar{a}_1 \bar{a}_2$, the right-hand side has the form

$$\begin{bmatrix} u & v \\ -\bar{v} & \bar{u} \end{bmatrix}$$

where $u = ac - b\bar{d}, v = ad + b\bar{c}$. Therefore, $\mathbb{H}$ is closed under multiplication and so $\mathbb{H}$ is a subring of $M_2(\mathbb{C})$.

We now show that $\mathbb{H}$ is a division ring. We note first that

$$\Delta := \det \begin{bmatrix} \alpha_0 + \alpha_1\sqrt{-1} & \alpha_2 + \alpha_3\sqrt{-1} \\ -\alpha_2 + \alpha_3\sqrt{-1} & \alpha_0 - \alpha_1\sqrt{-1} \end{bmatrix} = \alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2.$$

Since the α_i are real numbers, this is real, and is 0 only if every $\alpha_i = 0$, that is, if the matrix is 0. Hence, every non-zero element of $\mathbb{H}$ has an inverse in $M_2(\mathbb{C})$. Moreover, we have, by the definition of the adjoint, that

$$\text{adj} \begin{bmatrix} a & b \\ -\bar{b} & \bar{a} \end{bmatrix} = \begin{bmatrix} \bar{a} & -b \\ \bar{b} & a \end{bmatrix}.$$

Since $\bar{\bar{a}} = a$, this is obtained from the x in (2.1.1) by replacing a by $\bar{a}$ and b by $-b$, and so it is contained in $\mathbb{H}$. Thus, if the matrix x is $\neq 0$ then its inverse is

$$\begin{bmatrix} \bar{a}\Delta^{-1} & -b\Delta^{-1} \\ \bar{b}\Delta^{-1} & a\Delta^{-1} \end{bmatrix}$$

and this is contained in $\mathbb{H}$. Hence, $\mathbb{H}$ is a division ring.

The ring $\mathbb{H}$ contains in its center the field $\mathbb{R}$ of real numbers identified with the set of diagonal matrices $\begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$, $\alpha \in \mathbb{R}$. $\mathbb{H}$ also contains the matrices

$$i = \begin{bmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{bmatrix}, j = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, k = \begin{bmatrix} 0 & \sqrt{-1} \\ \sqrt{-1} & 0 \end{bmatrix}.$$

We verify that

$$x = \alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 k \tag{2.1.2}$$

and if $\alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 k = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k, \beta_i \in \mathbb{R}$, then

$$\begin{bmatrix} \alpha_0 + \alpha_1\sqrt{-1} & \alpha_2 + \alpha_3\sqrt{-1} \\ -\alpha_2 + \alpha_3\sqrt{-1} & \alpha_0 - \alpha_1\sqrt{-1} \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1\sqrt{-1} & \beta_2 + \beta_3\sqrt{-1} \\ -\beta_2 + \beta_3\sqrt{-1} & \beta_0 - \beta_1\sqrt{-1} \end{bmatrix}$$

so $\alpha_l = \beta_l, 0 \leq l \leq 3$. Thus, any x in $\mathbb{H}$ can be written in one and only one way in the from (2.1.2). The product of two elements in $\mathbb{H}$

$$(\alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 k)(\beta_0 + \beta_1 i + \beta_2 j + \beta_3 k)$$

is determined by the product and sum in $\mathbb{R}$, the distributive laws and the multiplication table

$$i^2 = j^2 = k^2 = -1$$

$$ij = -ji = k, jk = -kj = i, ki = -ik = j.$$

Incidentally, because these show that $\mathbb{H}$ is not commutative we have constructed an *infinite* division ring that is not a field. The ring $\mathbb{H}$ is called the **division ring of real quaternions**. Recall that $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ is called the quaternion group (see Porject 1).

Algebra is not as rich in division ring as it is in fields. For example, there are no finite division rings. This is the content of a famous theorem of Wedderburn. Its proof can be found in Section 5.7.

Theorem 2.1.4. [Wedderburn, 1909] *A finite division ring is a field.*

2.1.3 Characteristic

Let R be a ring. If there is a smallest positive integer n such that $na = 0$ for all $a \in R$, then R is said to have **characteristic** n. If no such n exists, R is said to have **characteristic zero**. We denote the characteristic of R by $\operatorname{char} R$. The characteristic of a ring gives some information on its additive group structure.

Remark. It is easy to see that $\operatorname{char} R = n$ if and only if n is the smallest positive integer such that $n1_R = 0$.

Example 2.1.5. The rings $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ and $\mathbb{H}$ are of characteristic zero, $\operatorname{char} \mathbb{Z}_n = n$ and $\operatorname{char}(\mathbb{Z}_m \times \mathbb{Z}_n) = \operatorname{lcm}(m, n)$.

Theorem 2.1.5. [Characteristic of an Integral Domain] *If R is an integral domain, then R is of characteristic zero or a prime p. In particular, every field is of characteristic zero or a prime p.*

Proof. Let R be an integral domain of characteristic $n > 0$. Assume that $n = ab$ for some $a, b \in \mathbb{N}$. It follows that $0 = n1_R = (ab)1_R = (a1_R)(b1_R)$. Since R has no zero divisor, $a1_R = 0$ or $b1_R = 0$. Then $a = n$ or $b = n$. Hence, n is a prime. □

Theorem 2.1.6. *Let R be a ring of characteristic a prime p and $a, b \in R$. If a and b commute, then*

$$(a + b)^p = a^p + b^p \quad \textit{and} \quad (a + b)^{p^k} = a^{p^k} + b^{p^k} \quad \textit{for all } k \in \mathbb{N}.$$

Proof. Note that if $1 \leq r \leq p-1$, then the binomial coefficient $\binom{p}{r}$ is a multiple of p, so it is 0 in R. Hence,

$$(a+b)^p = a^p + \binom{p}{1}a^{p-1}b + \cdots + \binom{p}{p-1}ab^{p-1} + b^p = a^p + b^p.$$

A simple induction on k gives the second equation. The inductive step is

$$(a+b)^{p^k} = ((a+b)^{p^{k-1}})^p = (a^{p^{k-1}} - b^{p^{k-1}})^p = a^{p^k} + b^{p^k}$$

and the proof is complete. □

2.1.4 Ring Homomorphisms and Group Rings

Like in groups, a ring homomorphism is a function between two rings that preserves both addition and multiplication. Let R and S be rings. A map $\varphi : R \to S$ is called a **homomorphism** if

$$\varphi(a+b) = \varphi(a) + \varphi(b) \quad \text{and} \quad \varphi(ab) = \varphi(a)\varphi(b)$$

for all $a, b \in R$. The definitions of monomorphisms, epimorphisms, endomorphisms, isomorphisms and automorphisms of rings are given as in groups.

Remarks. Let $\varphi : R \to S$ be a ring homomorphism.

1. $\varphi : (R, +) \to (S, +)$ is an additive group homomorphism.
2. $\varphi(1_R)$ may not be the identity of S.
3. If $\varphi(1_R) = 0$, then $\varphi(x) = 0$ for all $x \in R$.
4. If R has an identity and φ is onto, then $\varphi(1_R)$ is the identity of S.

   Proof. Let $s \in S$. Since φ is onto, $\exists x \in R, \varphi(x) = s$. Then $s\varphi(1_R) = \varphi(x)\varphi(1_R) = \varphi(x1_R) = \varphi(x) = s$. Similarly, $\varphi(1_R)s = s$. Hence, $\varphi(1_R)$ is the identity of S. □

Examples 2.1.6.

1. If $\varphi : \mathbb{Z} \to \mathbb{Z}$ is a ring homomorphism, then φ is the zero or the identity map.
2. If $\varphi : \mathbb{Q} \to \mathbb{Q}$ is a ring homomorphism, then φ is the zero or the identity map.
3. If $\varphi : \mathbb{R} \to \mathbb{R}$ is a ring homomorphism, then φ is the zero or the identity map.
4. If p is a prime and $\varphi : \mathbb{Z}_p \to \mathbb{Z}_p$ is a ring homomorphism, then φ is the zero or the identity map.

Proof. (1) Observe that $\varphi(1) = \varphi(1 \cdot 1) = \varphi(1)\varphi(1)$, so $\varphi(1) = 0$ or $\varphi(1) = 1$. Moreover, $\varphi(n) = n\varphi(1)$ for all $n \in \mathbb{Z}$. Thus, $\varphi(n) = 0$ for all $n \in \mathbb{Z}$ or $\varphi(n) = n$ for all $n \in \mathbb{Z}$ as desired.
(2) Similar to $\mathbb{Z}$, $\varphi(1) = 0$ or $\varphi(1) = 1$ and $\varphi(n) = n\varphi(1)$ for all $n \in \mathbb{Z}$. For $m \in \mathbb{Z}$ and $n \in \mathbb{N}$, we have $\varphi(m(1/n)) = m\varphi(1/n)$ and $\varphi(1) = \varphi(n(1/n)) = n\varphi(n)$. If $\varphi(1) = 0$, then $\varphi(1/n) = 0$ for all $n \in \mathbb{N}$, so φ is the zero map. On the other hand, if $\varphi(1) = 1$, then $\varphi(1/n) = 1/n$ for all $n \in \mathbb{N}$ which implies $\varphi(m/n) = m/n$ for all $m \in \mathbb{Z}$ and $n \in \mathbb{N}$.
(3) Assume that $\varphi(x)$ is not the zero map. We can show that $\varphi(x) = x$ for all $x \in \mathbb{Q}$. Moreover, for $x \in \mathbb{R}^+$, $\varphi(x) = \varphi((\sqrt{x})^2) = (\varphi(\sqrt{x}))^2 > 0$. This implies $\forall a, b \in \mathbb{R}, a < b \Rightarrow \varphi(a) < \varphi(b)$. Now, let $x \in \mathbb{R}$. Suppose that $\varphi(x) \neq x$. Then $\varphi(x) < x$ or $x < \varphi(x)$. By the density theorem, $\exists q_1, q_2 \in \mathbb{Q}$ such that $\varphi(x) < q_1 < x$ or $x < q_2 < \varphi(x)$. Thus, $\varphi(x) < q_1 < \varphi(x)$ or $\varphi(x) < q_2 < \varphi(x)$ yields a contradiction. Hence, $\varphi(x) = x$ for all $x \in \mathbb{R}$.
(4) is proved in the next section. □

Let $G = \{g_i : i \in I\}$ be any multiplicative group, and let R be any commutative ring. Let RG be the set of all formal sums

$$\sum_{i \in I} a_i g_i$$

for $a_i \in R$ and $g_i \in G$, where *all but finite number of the* a_i *are* 0. Define the sum of two elements of RG by

$$\left(\sum_{i\in I} a_i g_i\right) + \left(\sum_{i\in I} b_i g_i\right) = \sum_{i\in I}(a_i + b_i)g_i.$$

Observe that $(a_i + b_i) = 0$ except for a finite number of indices i, so the above sum is again in RG. It is immediate that $(RG, +)$ is an abelian group with additive identity $\sum_{i\in I} 0g_i$.

Multiplication of two elements of RG is defined by the use of the multiplications in G and R as follows:

$$\left(\sum_{i\in I} a_i g_i\right)\left(\sum_{i\in I} b_i g_i\right) = \sum_{i\in I}\left(\sum_{g_j g_k = g_i} a_j b_k\right) g_i.$$

Naively, we formally distribute the sum $\sum_{i\in I} a_i g_i$ over the sum $\sum_{i\in I} b_i g_i$ and rename a term $a_j g_j b_k g_k$ by $a_j b_k g_k$ where $g_j g_k = g_i$ in G. Since a_i and b_i are 0 for all but a finite number of i, the sum $\sum_{g_j g_k = g_i} a_j b_k$ contains only a finite number of nonzero summands $a_j b_k \in R$ and may thus be viewed as an element of R. Again at most a finite number of such sums $\sum_{g_j g_k = g_i} a_j b_k$ are nonzero. Thus, multiplication is closed on RG. We can proceed to show that

Theorem 2.1.7. [Group Ring] *If G is a multiplicative group and R is a commutative ring, then $(RG, +, \cdot)$ is a ring with unity $1_R\, e$.*

If we rename the element $\sum_{i\in I} a_i g_i$ of RG, where $a_i = 0$ for $i \neq j$ and $a_j = 1$, by g_j, we see that $(RG, \cdot)$ can be considered to contain G naturally. Thus, if G is not abelian, RG is not commutative. Clearly, $\operatorname{char} RG = \operatorname{char} R$, for any group G. The ring RG defined above is the **group ring of** G **over** R. If F is a field, then FG is the **group algebra of** G **over** F.

Exercises 2.1.

1. Define an addition and a multiplication on $\mathbb{Z}$ by

$$a \oplus b = a + b - 1 \quad \text{and} \quad a \odot b = ab - (a + b) + 2 \quad \text{for all } a, b \in \mathbb{Z}.$$

   Prove that $(\mathbb{Z}, \oplus, \odot)$ is an integral domain.
2. Let S be the set of complex numbers of the form $m + n\sqrt{-3}$ where either $m, n \in \mathbb{Z}$ or both m and n are halves of odd integers. Show that S is a subring of $\mathbb{C}$.
3. Show that if $1 - ab$ is invertible in a ring, then so is $1 - ba$.
4. Let a and b be elements of a ring such that a, b and $ab - 1$ are units. Show that $a - b^{-1}$ and $(a - b^{-1})^{-1} - a^{-1}$ are units and the following identity holds:

$$((a - b^{-1})^{-1} - a^{-1})^{-1} = aba - a.$$

5. A ring R is called a **Boolean ring** is $x^2 = x$ for all $x \in R$. Prove that every Boolean ring is commutative.
6. (a) Show that $\varphi : \mathbb{Z}_{12} \to \mathbb{Z}_{30}$ given by $\varphi([a]_{12}) = [10a]_{30}$ is a ring homomorphism.
   (b) Show that $\varphi : \mathbb{Z}_{12} \to \mathbb{Z}_{30}$ given by $\varphi([a]_{12}) = [5a]_{30}$ is a additive group homomorphism. Is it a ring homomorphism?
7. Consider $(S, +, \cdot)$, where S is a set and $+$ and $\cdot$ are binary operations on S which satisfy the distributive laws such that $(S, +)$ and $(S \setminus \{0\}, \cdot)$ are groups. Show that $(S, +, \cdot)$ is a division ring.
8. Let R be a ring. Define $C(R) = \{x \in R : \forall y \in R, xy = yx\}$, called the **center of** R.
   (a) Prove that $C(R)$ is a commutative subring of R.
   (b) Determine the centers of $\mathbb{H}$ and $M_n(F)$ where F is a field.
   (c) If R is a division ring, show that $C(R)$ is a field.
9. If p is a prime and $\varphi : \mathbb{Z}_p \to \mathbb{Z}_p$ is a ring homomorphism, show that φ is the zero or the identity map.
10. Show that if F is a field, $A \in M_n(F)$ is a zero divisor in this ring if and only if A is not invertible. Does this hold for arbitrary commutative ring R? Explain.

11. Let m and n be non-zero integers and let R be the subset of $M_2(\mathbb{C})$ consisting of the matrices of the form
$$\begin{bmatrix} a+b\sqrt{m} & c+d\sqrt{m} \\ n(c+d\sqrt{m}) & a-b\sqrt{m} \end{bmatrix}$$
where $a, b, c, d \in \mathbb{Q}$. Show that R is a subring of $M_2(\mathbb{C})$ and that R is a division ring if and only if the only rational numbers x, y, z, t satisfying the equation $x^2 - my^2 - nz^2 + mnt^2 = 0$ are $x = y = z = t = 0$. Give a choice of m, n that R is a division ring and a choice of m, n that R is not a division ring.
12. Let R be a ring which may not contain the unity 1. Define two binary operations on $R \times \mathbb{Z}$ by
$$(r,k) + (s,m) = (r+s, k+m) \quad \text{and} \quad (r,k) \cdot (s,m) = (rs + ks + mr, km).$$
Prove that $(R \times \mathbb{Z}, +, \cdot)$ is a ring with unity $(0,1)$ and of characteristic zero. Ditto the set $R \times \mathbb{Z}_n$ and prove that it is a ring of characteristic n.
13. A ring R is **simple** if R and $\{0\}$ are the only ideals in R. Show that the characteristic of a simple ring is either 0 or a prime p.
14. If R is a finite integral domain, prove that $|R|$ is a prime power.

2.2 Ideals, Quotient Rings and the Field of Fractions

Ideals play an important role in ring theory. They are used to construct quotient rings like normal subgroups.

Let R be a ring. A subset I of R is called a **left [right] ideal of** R if

1. I is a subgroup of $(R, +)$ and 2. $\forall r \in R \forall a \in I, ra \in I$ $[ar \in I]$.

It is called a **two-sided ideal** or an **ideal** of R if I is both a left and a right ideal.

E.g., $\{0\}$ and R are two-sided ideals of R.

Theorem 2.2.1. *Let $\varphi : R \to S$ be a homomorphism of rings. Then the* **kernel of** φ *given by*
$$\ker \varphi = \{x \in R : \varphi(x) = 0_S\}$$
is an ideal of R.

Proof. It is immediate that $\ker \varphi$ is a subgroup of $(R, +)$. If $a \in R$ and $x \in \ker \varphi$, then $\varphi(ax) = \varphi(a)\varphi(x) = \varphi(a)0 = 0$ and $\varphi(xa) = \varphi(x)\varphi(a) = 0\varphi(a) = 0$. Hence, ax and xa are in $\ker \varphi$. □

Remark. Similar to a group homomorphism, for a ring homomorphism, we have φ is one-to-one if and only if $\ker \varphi = \{0\}$.

For subsets X and Y of R, let XY denote the set of all finite sums in the form
$$\sum_{i=1}^{n} x_i y_i, \quad \text{where } x_i \in X,\ y_i \in Y \text{ and } n \in \mathbb{N}.$$

For $a \in R$, we have $Ra = \{ra : r \in R\}$ and $aR = \{ar : r \in R\}$.

Examples 2.2.1. 1. All distinct ideals of $\mathbb{Z}_n$ are $d\mathbb{Z}_n$, where $d = 0$ or ($d \in \mathbb{N}$ and $d \mid n$).
2. All distinct ideals of $\mathbb{Z}$ are $m\mathbb{Z}$, where $m \in \mathbb{N} \cup \{0\}$.

Remarks. Let R be a ring.

1. If I is an ideal of R, then $IR = I = RI$.
2. If a left [right, two-sided] ideal I of R contains a unit, then $I = R$.
3. If R is a division ring, then $\{0\}$ and R are the only left [right, two-sided] ideals of R.

4. An *arbitrary* intersection of left [right, two-sided] ideals of R is a left [right, two-sided] ideal of R.
5. If S is a subring of R and I is an ideal of R, then $S + I$ is a subring of R, I is an ideal of $S + I$ and $S \cap I$ is an ideal of S.
6. If I and J are left [right, two-sided] ideals of R, then $I + J$ and IJ are left [right, two-sided] ideals of R.

Example 2.2.2. Let R be a ring and $S = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} : a, b, c \in R \right\}$ a subring of $M_2(R)$. Then $\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} : a \in R \right\}$ is a left ideal of S and $\left\{ \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} : a \in R \right\}$ is a right ideal of S. They are not ideals of S.

Let R be a ring. If $X \subset R$, then the **ideal of R generated by X** is the intersection of all ideals containing X and it is denoted by (X), so (X) is the smallest ideal of R containing X. For $a_1, \ldots, a_n \in R$, let $(a_1, \ldots, a_n)$ denote $(\{a_1, \ldots, a_n\})$. An ideal I of R is called a **principal ideal** if $I = (a)$ for some $a \in R$. Observe that for $a \in R$,

$$(a) = \left\{ \sum_{i=1}^{m} r_i a s_i : r_i, s_i \in R \text{ and } m \in \mathbb{N} \right\}.$$

If R is a commutative ring, then $\forall a \in R, (a) = aR = Ra$.

A ring R is a **principal ideal ring** if every ideal of R is principal. A **principal ideal domain (PID)** is a principal ideal ring which is an integral domain. Hence, an integral domain R is a PID if $\{Ra : a \in R\}$ is the set of all ideals of R.

Examples 2.2.3.
1. $\mathbb{Z}_n$ is a principal ideal ring.
2. $\mathbb{Z}$ is a PID.
3. Every field has only two ideals, namely (0) and $(1) = F$, so it is a PID.

Remark. Let F be a field, R a ring and $\varphi : F \to R$ a ring homomorphism. Then $\ker \varphi$ is either $\{0\}$ or F which implies φ is 1-1 or is the zero map, respectively. Hence, every nonzero ring homomorphism of fields must be 1-1. In particular, one can readily verify that the only ring endomorphisms of $\mathbb{Z}_p$ are the zero map and the identity map. This finishes the proof of Example 2.1.6.

Theorem 2.2.2. *Let R be a commutative ring whose only ideals are $\{0\}$ and R itself. Then R is a field.*

Proof. Let $a \in R \setminus \{0\}$. Then $(a) = R$, so $1 \in (a)$. Since R is commutative, there is a $b \in R$ such that $ab = 1 = ba$. □

Suppose I is an ideal of R. Then I is a subgroup of R, considered as an abelian group, and so we can form the abelian group R/I. The elements of R/I are cosets

$$r + I = \{r + a : a \in I\}.$$

The addition in R/I is given by $(r + I) + (s + I) = (r + s) + I$.

Now let us define a multiplication on R/I, namely

$$(r + I)(s + I) = rs + I.$$

Note that if $r + I = r' + I$ and $s + I = s' + I$, then $r - r'$ and $s - s'$ are in I, so

$$rs - r's' = (r - r')s + r'(s - s') \in I.$$

Thus, the above multiplication is well defined, it is easy to see that R/I is a ring. Hence, we have the next theorem.

Theorem 2.2.3. [Quotient Ring] *Let R be a ring and I an ideal of R. Then the operators*

$$(r+I)+(s+I)=(r+s)+I \quad \textit{and} \quad (r+I)(s+I)=rs+I$$

make R/I into a ring with unity $1+I$, called the **factor** *or* **quotient ring of R by I.** *The map $\varphi : R \to R/I$ defined by $\varphi(r) = r + I$ is an onto ring homomorphism which has kernel I. It is called the* **canonical projection of R onto R/I.**

There also are three isomorphism theorems for rings. Their proofs are similar to isomorphism theorems for groups. Hence, we shall just sketch them.

Theorem 2.2.4. [First Isomorphism Theorem] *If $\varphi : R \to S$ is an onto ring homomorphism, then*

$$R/\ker\varphi \cong \operatorname{im} S.$$

Proof. Define $\bar{\varphi} : R/\ker\varphi \to S$ by $\bar{\varphi}(r+\ker\varphi)\varphi(r)$ for all $r \in R$. Clearly, $\bar{\varphi}$ is onto and it is easy to check that $\bar{\varphi}$ is a ring homomorphism. Moreover, for $r, s \in R$, we have

$$\varphi(r)=\varphi(s) \Leftrightarrow \varphi(r-s)=0 \Leftrightarrow r-s \in \ker\varphi \Leftrightarrow r+\ker\varphi = s+\ker\varphi.$$

Hence, $\bar{\varphi}$ is an isomorphism. □

Theorem 2.2.5. [Second Isomorphism Theorem] *If S is a subring and I is an ideal of R, then*

$$S/(S\cap I) \cong (S+I)/I.$$

Proof. Define $\varphi : S \to (S+I)/I$ by $\varphi(s) = s + I$ for all $s \in S$. It is easy to verify that φ is a ring homomorphism with kernel $S \cap I$ and the theorem follows from the first isomorphism theorem. □

Theorem 2.2.6. [Third Isomorphism Theorem] *If I and J are ideals of a ring R such that $I \subseteq J$, then J/I is an ideal of R/I and*

$$(R/I)/(J/I) \cong R/J.$$

Proof. Define $\varphi : R/I \to R/J$ by $\varphi(r+I) = r + J$ for all $r \in R$. It can be verified that φ is a ring homomorphism with kernel J/I and the theorem follows from the first isomorphism theorem. □

Remark. As for groups, the third isomorphism theorem gives a 1-1 correspondence between the set of ideals of R containing I and the set of ideals of R/I.

We end this section by embedding an integral domain into a field. We say that a ring R can be **embedded** in a ring R' if there exists a monomorphism (i.e., an injective homomorphism) of R into R'.

Example 2.2.4. A ring R can be embedded in the ring $M_n(R)$ by the diagonal map $a \mapsto aI_n$.

Theorem 2.2.7. *Any ring R without identity can be embedded in a ring R' with identity. Moreover, R' can be chosen to be either of characteristic zero or of same characteristic as R.*

Proof. Consider the rings $R \times \mathbb{Z}$ and $R \times \mathbb{Z}_n$ defined in Exercises 2.1. They are rings with unity $(0,1)$ and $(0,\bar{1})$, and of characteristic 0 and n, respectively. If $\operatorname{char} R = 0$, we define $\varphi : R \to R\times\mathbb{Z}$ by $\varphi(x) = (x, 0)$ and if $\operatorname{char} R = n$, we define $\varphi : R \to R \times \mathbb{Z}_n$ by $\varphi(x) = (x, \bar{0})$. It is easy to show that both functions are monomorphisms. This finishes the proof. □

We now wish to show that every integral domain can be embedded in a field, called its *field of fractions* such that every element of the field is a fraction a/b where a and b lie in the integral domain and $b \neq 0$. There is only one problem to overcome: we might wish to define the field to be the set of all "fraction" a/b, with $b \neq 0$. But this is not quite right because two different fractions may be the same number. E.g., $1/2 = 2/4 = 3/6$. We overcome this problem by defining an equivalence relation on certain pairs of elements in the integral domain. The results are presented in the next theorem. Its proof is routine and omitted.

Theorem 2.2.8. [Field of Fractions] *Suppose D is an integral domain, and let S be the set of pairs*

$$\{(r, s) : r, s \in D \text{ and } s \neq 0\}.$$

1. *$(r, s) \sim (r', s') \Leftrightarrow rs' = r's$ defines an equivalence relation on S.*
2. *Let $[r, s]$ denote the equivalence class of (r, s) and let $Q(D)$ denote the set of all equivalence classes. Then*

$$[r, s] + [r', s'] = [rs' + r's, ss'] \quad \text{and} \quad [r, s][r', s'] = [rr', ss']$$

   are well defined binary operations on $Q(D)$.
3. *The set $Q(D)$ is a field with these operations and D is embedded in $Q(D)$ by the monomorphism $r \mapsto [r, 1]$. The field $Q(D)$ is called the* **field of fractions** *or* **quotient field of** *D. The equivalence class $[r, s]$ is denoted by r/s.*

Remark. If R is an entire ring which is not commutative, the construction $Q(R)$ above does not exist in general.

Example 2.2.5. Let D be an integral domain and $a, b \in D$. If $a^m = b^m$ and $a^n = b^n$, for m and n relatively prime positive integers, prove that $a = b$.

Proof. If $a = 0$, then $b = 0$ since D has no zero divisor. Assume that $a \neq 0$. Then $b \neq 0$. Let F be the field of fractions of D. Since $(m, n) = 1$, $\exists x, y \in \mathbb{Z}, mx + ny = 1$. Thus, in F, we have

$$a = a^1 = a^{mx+ny} = (a^m)^x(a^n)^y = (b^m)^x(b^n)^y = b^{mx+ny} = b^1 = b,$$

so $a = b$ in D. □

Exercises 2.2.

1. An element a of a ring R is **nilpotent** if $a^n = 0$ for some $n \in \mathbb{N}$. Show that the set of all nilpotent elements N in a commutative ring R is an ideal, called the **nilradical of** R. Moreover, prove that R/N has no nonzero nilpotent.
2. Show that a ring R has no nonzero nilpotent element if and only if 0 is the only solution of $x^2 = 0$ in R.
3. Let $\varphi : R \to S$ be a homomorphisms of rings. Prove the following statements.
   (a) If I is an ideal of R and φ is onto, then $\varphi(I)$ is an ideal of S.
   (b) If J is an ideal of S, then $\varphi^{-1}(J)$ is an ideal of R containing $\ker \varphi$.
4. Let R be a commutative ring and I an ideal of R. Show that

$$\sqrt{I} = \{x \in R : \exists n \in \mathbb{N}, x^n \in I\}$$

   is an ideal of R which contains I, called the **radical of** I. In addition, prove that
   (a) $\sqrt{\sqrt{I}} = \sqrt{I}$ and (b) if $\sqrt{I} = R$, then $I = R$.
5. Let R and S be rings and $\varphi : R \to S$ be such that
   (i) $\forall r, s \in R, \varphi(r + s) = \varphi(r) + \varphi(s)$ and
   (ii) $\forall r, s \in R\, [\varphi(rs) = \varphi(r)\varphi(s) \vee \varphi(rs) = \varphi(s)\varphi(r)]$.
   Prove that $\forall r, s \in R, \varphi(rs) = \varphi(r)\varphi(s)$ or $\forall r, s \in R, \varphi(rs) = \varphi(s)\varphi(r)$.

6. [Chinese Remainder Theorem] If I and J are ideals of a ring R such that $I + J = R$, prove that $R/(I \cap J) \cong R/I \times R/J$.
7. Let R be a division ring. Prove that any nonzero ring homomorphism $\varphi : R \to R$ is 1-1.
8. Let I be an ideal of a ring R and let $M_n(I)$ be the set of $n \times n$ matrices with entries in I. Prove that
   (a) $M_n(I)$ is an ideal of $M_n(R)$ and $M_n(R)/M_n(I) \cong M_n(R/I)$, and
   (b) every ideal of $M_n(R)$ is of the form $M_n(I)$ for some ideal I of R. In particular, if R is a division ring, then the ring $M_n(R)$ has only two ideals.

Project 10. Let $n \in \mathbb{N}$ and $n \geq 2$. Define $\mathbb{Z}_n[i] = \{a + ib : a, b \in \mathbb{Z}_n\}$ where $i^2 \equiv -1 \pmod{n}$.
(a) Prove that $\mathbb{Z}_n[i]$ is a ring containing $\mathbb{Z}_n$ as a subring.
(b) Determine all units, zero divisors and nilpotent elements in $\mathbb{Z}_n$.
(c) Determine all units, zero divisors and nilpotent elements in $\mathbb{Z}_n[i]$.

2.3 Maximal Ideals and Prime Ideals

We have learned that a ring R has two trivial ideals, namely $\{0\}$ and R itself. In this section, we shall discover properties of maximal ideals and prime ideals. These are two kinds of important ideals in commutative algebra and algebraic geometry.

An ideal M of R is **maximal** if $M \neq R$ and for every ideal J of R,

$$M \subseteq J \subseteq R \Rightarrow J = M \text{ or } J = R.$$

Example 2.3.1. In the ring $\mathbb{Z}$, for $n \in \mathbb{N}$, $n\mathbb{Z}$ is maximal if and only if n is a prime.

Proof. Let n be a prime and let J be an ideal of $\mathbb{Z}$ such that $n\mathbb{Z} \subseteq J \subseteq \mathbb{Z}$. Then $J = d\mathbb{Z}$ for some $d \in \mathbb{N}$ and $d \mid n$, so $d = 1$ or $d = n$. Hence, $J = n\mathbb{Z}$ or $J = \mathbb{Z}$. On the other hand, assume that $n = ab$ for some $1 < a, b < n$. Then $n\mathbb{Z} \subseteq a\mathbb{Z} \subseteq \mathbb{Z}$, $a\mathbb{Z} \neq n\mathbb{Z}$ and $a\mathbb{Z} \neq \mathbb{Z}$, so $n\mathbb{Z}$ is not maximal. □

Remarks. 1. Every ideal $I \neq R$ is contained in some maximal ideal M.

Proof. Let $\mathscr{I} = \{J : J \neq R \text{ and } J \text{ is an ideal of } R \text{ containing } I\}$. Let $\mathscr{C} = \{J_\alpha\}_{\alpha \in \Lambda}$ be a chain in $\mathscr{I}$. Then $\cup\mathscr{C}$ is an ideal of R. If $\cup\mathscr{C} = R$, then $1 \in J_\alpha$ for some $\alpha \in \Lambda$, so $J_\alpha = R$, a contradiction. Hence, $\cup\mathscr{C}$ is an upper bound of $\mathscr{C}$ in $\mathscr{I}$. By, Zorn's lemma, we have $\mathscr{I}$ has a maximal element which turns out to be our desired maximal ideal containing I. □

2. If M is a maximal ideal and I is an ideal of R such that $I \nsubseteq M$, then $M + I = R$.

Proof. Let $x \in I, \notin M$. Consider the ideal $J = M + Rx$ which is larger than M. Since M is maximal, $J = R$. Thus, $R = M + Rx \subseteq M + I$. □

3. If M_1 and M_2 are distinct maximal ideals, then $M_1 + M_2 = R$. In addition, if R is commutative, then $M_1 M_2 = M_1 \cap M_2$.
4. If R is commutative, then $Ru = R$ if and only if u is a unit.

Theorem 2.3.1. *Let R be a commutative ring and M an ideal of R.*
Then M is a maximal ideal of R if and only if R/M is a field.

Proof. Clearly, R/M is a commutative ring with unity $1 + M$. Assume that M is a maximal ideal. Let $a \notin M$. Then $M + Ra = R$, so $\exists b \in R, 1 = m + ba$. Thus, $1 + M = ba + M = (b + M)(a + M)$, and hence R/M is a field. Conversely, suppose that R/M is a field. Let $M \subseteq J \subseteq R$ and $J \neq M$. Then $\exists a \in J \setminus M$. Since R/M is a field and $a \notin M$, $\exists b \in R, 1 + M = (a + M)(b + M) = ab + M$, so $1 - ab \in M \subseteq J$. Since $a \in J$, $ab \in J$ which implies $1 \in J$. Hence, $J = R$. □

An ideal P of R is **prime** of R if $P \neq R$ and for any ideals A, B of R,

$$AB \subseteq P \;\Rightarrow\; A \subseteq P \text{ or } B \subseteq P.$$

Theorem 2.3.2. *Let P be an ideal of R such that $P \neq R$.*
1. *If $\forall a, b \in R, ab \in P \Rightarrow a \in P$ or $b \in P$, then P is prime.*
2. *If R is commutative and P is prime, then $\forall a, b \in R, ab \in P \Rightarrow a \in P$ or $b \in P$.*

Proof. (1) Assume that $AB \subseteq P$ and $A \nsubseteq P$. Choose $a \in A, \notin P$. Let $b \in B$. Thus, $ab \in AB \subseteq P$, so $a \in P$ or $b \in P$. But $a \notin P$, hence $B \subseteq P$.
(2) Let $a, b \in R$ be such that $ab \in P$. Since R is commutative, $Rab = RaRb \subseteq P$, so $Ra \subseteq P$ or $Rb \subseteq P$. Hence, $a \in P$ or $b \in P$. □

Theorem 2.3.3. *Let R be a commutative ring and P an ideal of R.*
Then P is a prime ideal of R if and only if R/P is an integral domain.

Proof. This follows from Theorem 2.3.2 as follows. For an ideal P,

$$\begin{aligned} P \text{ is prime} &\Leftrightarrow \forall a, b \in R, ab \in P \Rightarrow a \in P \text{ or } b \in P \\ &\Leftrightarrow \forall a, b \in R, (a+P)(b+P) = 0+P \Rightarrow a+P = 0+P \text{ or } b+P = 0+P \\ &\Leftrightarrow R/P \text{ is an integral domain} \end{aligned}$$

as desired. □

Theorems 2.3.1 and 2.3.3 are the most useful for characterizing maximal ideals and prime ideals in commutative rings.

Corollary 2.3.4. *Let R be a commutative ring.*
1. *Every maximal ideal of R is prime.*
2. *If R is finite, then every prime ideal of R is maximal.*

Example 2.3.2. In $\mathbb{Z}$, $n\mathbb{Z}$ is prime if and only if $n = 0$ or n is a prime.

Remark. In the ring $\mathbb{Z}$, $\{0\}$ is a prime ideal which is not maximal.

The set of all prime ideals of a commutative ring R is denoted by $\operatorname{Spec} R$, called the **spectrum of** R. E.g., $\operatorname{Spec}\mathbb{Z} = \{p\mathbb{Z} : p \text{ is a prime}\} \cup \{\{0\}\}$. A **local ring** is a commutative ring which has a unique maximal ideal.

Examples 2.3.3.
1. $\mathbb{Z}$ has infinitely many maximal ideals of the form $p\mathbb{Z}$ where p is a prime, so it is not a local ring.
2. Every field is a local ring with maximal ideal $\{0\}$.
3. $\mathbb{Z}_{p^n}$ is a local ring with the maximal ideal $p\mathbb{Z}_{p^n}$ for all primes p and $n \in \mathbb{N}$.

Theorem 2.3.5. *Let R be a commutative ring.*
Then R is a local ring if and only if the nonunits of R form an ideal.

Proof. Assume R is a local ring with the maximal ideal M. Let $a \in R \setminus M$. If $aR \neq R$, then aR is contained in some maximal ideal, so $aR \subseteq M$ which yields a contradiction. Thus, $aR = R$, so a is a unit. Hence, M is the set of nonunits of R. Conversely, suppose that the nonunits of R form an ideal M of R. Clearly, M is maximal. Let M' be another maximal ideal of R. If $\exists a \in M' \setminus M$, then a is a unit, so $M' = R$, a contradiction. Thus, $M' \subseteq M$. Since M' is maximal, $M' = M$. □

Corollary 2.3.6. *In a finite local ring R, every element is either a unit or a nilpotent element.*

Example 2.3.4. Fix a prime p and let

$$Z_p = \{m/n : m, n \in \mathbb{Z} \text{ and } p \text{ does not divide } n\}.$$

Then Z_p is a subring of $\mathbb{Q}$ and is local. Its unique maximal ideal is $\{pk/n : k, n \in \mathbb{Z} \text{ and } p \nmid n\}$.

We shall show an important structure theorem for finite commutative rings in Section 4.6. It says that every finite commutative ring is a direct product of a finite number of local rings. (Corollary 4.6.7). Hence, a local ring turns out to be a core when we study a finite commutative ring. It has many applications coding theory and cryptography.

Exercises 2.3. 1. Let R be a ring and I an ideal of R. Prove that the map $J \mapsto J/I$ gives a 1-1 correspondence

$$\{\text{ideals of } R \text{ containing } I\} \longleftrightarrow \{\text{ideals of } R/I\}.$$

Moreover, this correspondence carries maximal ideals to maximal ideals.

2. Prove Corollary 2.3.6 and Example 2.3.4.
3. Find all ideals, all prime ideals and all maximal ideals of
   (a) $\mathbb{Z}_{12}$ (b) $\mathbb{Z}_2 \times \mathbb{Z}_4$ (c) $\mathbb{Q} \times \mathbb{Q}$ (d) $\mathbb{Q} \times \mathbb{Z}$ (e) $\mathbb{Z} \times \mathbb{Z}_4 \times \mathbb{Z}_5$.
4. Let R be a commutative ring. If every ideal proper of R is prime, show that R is a field.
5. Show that in a Boolean ring R, every prime ideal $P \neq R$ is maximal.
6. Let R be a commutative ring and $b \in R$ a nilpotent element. Prove that $u + b$ is a unit for all units u in R.

Project 11 (Chain ring)**.** A ring is called a **chain ring** if all its ideals form a chain under inclusion. For example, $\mathbb{Z}_{p^n}$, p a prime and $n \in \mathbb{N}$, is a chain ring. Also, every field is a chain ring. Let R be a finite commutative ring. Prove that R is a chain ring if and only if R is a local ring whose maximal ideal is principal.

A finite chain ring arises in algebraic number theory as quotient rings of rings of integers in number fields. It has many applications in coding theory because of the similarity with finite fields. Galois rings in Project 16 are examples for this situation.

2.4 Factorizations

From elementary number theory, we know that every positive integer can be decomposed uniquely into a product of prime numbers (Theorem 1.1.5). It is the unique factorization property of the ring $\mathbb{Z}$. In this section, we shall learn about factorizations in any other integral domains.

2.4.1 Irreducible Elements and Prime Elements

Let R be a commutative ring and suppose that $a, b \in R$. We say that a **divides** b and write $a \mid b$, if there is an $r \in R$ such that $ra = b$. This definition coincides the divisibility discussed previously in Section 1.1.

Remarks. Let R be a commutative ring and $a, b \in R$.

1. a divides $b \Leftrightarrow b \in Ra \Leftrightarrow Rb \subseteq Ra$.
2. a divides 0 ($R0 \subseteq Ra$).
3. $a \in R$, 1 divides a ($Ra \subseteq R \cdot 1 = R$).
4. a divides $1 \Leftrightarrow R = Ra \Leftrightarrow a$ is a unit.
5. 0 divides $a \Leftrightarrow Ra \subseteq R0 \Leftrightarrow a = 0$.

Let R be an integral domain and suppose $a, b \in R$. We say that a and b are **associates** if $a \mid b$ and $b \mid a$.

Theorem 2.4.1. *Let R be an integral domain, $a, b \in R$. The following statements are equivalent.*
(i) a and b are associates. *(ii) $Ra = Rb$.* *(iii) $a = ub$ for some unit $u \in R$.*

Proof. (i) $\Rightarrow$ (iii) If $a = 0$, then $b = 0$ and (3) is clear. Suppose then that $a \neq 0$. Since $a \mid b$ and $b \mid a$, we can write $a = ub$ and $b = va$. Thus, $a = ub = uva$, so $(uv - 1)a = 0$, so $uv = 1$. Hence, $a = ub$ and u is a unit of R.
(iii) $\Rightarrow$ (ii) If $a = ub$ where u is a unit, then $Ra = Rub = (Ru)b = Rb$.
(ii) $\Rightarrow$ (i) If $Ra = Rb$, then $a = rb$, $b = sa$, so $b \mid a$ and $a \mid b$. Hence, a and b are associates. □

Let R be an integral domain. We say that a nonzero nonunit element a in R is an **irreducible element** or **atom** if a cannot be expressed as a product $a = bc$ where b and c are nonunits. For example, in $\mathbb{Z}$, p and $-p$, p a prime number, are irreducible elements.

Theorem 2.4.2. *Let R be an integral domain and a a nonzero nonunit in R.*
1. *a is irreducible $\Leftrightarrow$ ($\forall b, c \in R, a = bc \Rightarrow$ b or c is a unit).*
2. *If Ra is maximal, then a is irreducible. The converse holds if R is a PID.*

Proof. (1) It follows directly from the definition.
(2) Assume that Ra is maximal. Let $b, c \in R$ be such that $a = bc$. Then $Ra \subseteq Rb \subseteq R$. Since Ra is maximal, $Ra = Rb$ or $Rb = R$. If $Rb = R$, then b is a unit. Let $Ra = Rb$. Then $a = bu$ for some unit $u \in R$, so $bc = bu$ which implies $c = u$ is a unit since R has no zero divisor.

Finally, we assume that R is a PID and $a \in R$ is irreducible. Let J be an ideal of R such that $Ra \subseteq J \subseteq R$. Since R is a PID, $J = Rb$ for some b in R, and so $a \in Rb$. Thus, $a = cb$ for some $c \in R$, so b or c is a unit. Hence, $Rb = R$ or $Ra = Rb$. □

A nonzero nonunit element p in R is a **prime element** if

$$\forall a, b \in R, p \mid ab \Rightarrow p \mid a \text{ or } p \mid b.$$

Note that a prime number is a prime element in $\mathbb{Z}$ by Corollary 1.1.4 (2).

Theorem 2.4.3. *Let R be an integral domain.*
(1) For a nonzero nonunit p in R, p is prime $\Leftrightarrow$ Rp is a prime ideal.
(2) Every prime element is irreducible. The converse holds if R is a PID.

Proof. (1) It follows directly from the definition and Theorem 2.3.2.
(2) Let p be a prime element. Assume that $p = ab$ for some $a, b \in R$. Then $Rab = Rp$, so $Ra \subseteq Rp$ or $Rb \subseteq Rp$. Since $Rp = Rab \subseteq (Ra \cap Rb)$, $Ra = Rp$ or $Rb = Rp$, so $au = p$ or $bv = p$ for some units u and v in R. Hence, $b = u$ or $a = v$ is a unit in R. Finally, suppose that R is a PID and p is irreducible. Then Rp is maximal, so it is a prime ideal. Hence, p is prime. □

2.4.2 Unique Factorization Domains

A **unique factorization domain (UFD)** is an integral domain R which satisfies:
1. Every nonzero nonunit of R is a product of atoms.
2. If a is a nonzero nonunit of R, then the expression of a as a product of atoms is unique in the following sense: "If $a = a_1 \dots a_r = b_1 \dots b_s$ where $a_1, \dots, a_r, b_1, \dots, b_s$ are atoms, then $r = s$ and there is a reordering $b_{i_1}, \dots, b_{i_r}$ of $b_1, \dots, b_s$ such that a_1 and b_{i_1} are associates, a_2 and b_{i_2} are associates, $\dots$, a_r and b_{i_r} are associates".

Examples 2.4.1. 1. The ring of rational integers $\mathbb{Z}$ is a UFD by the fundamental theorem of arithmetic (Theorem 1.1.5). Since $\mathcal{U}(\mathbb{Z}) = \{\pm 1\}$, the atoms of $\mathbb{Z}$ are $\pm p$ where p is a prime. Note that p and $-p$ are associates (e.g., $12 = 2 \cdot 2 \cdot 3 = (-2)(-3) \cdot 2$).

2. Let F be a field. Every element of F except 0 is a unit. Hence, every nonzero nonunit of F is uniquely a product of atoms (vacuously!). That is, F has no nonzero nonunits.

Theorem 2.4.4. *Let R be an integral domain. Then R is a UFD if and only if*
1. *every nonzero nonunit of R is a product of atoms and*
2. *every irreducible element is prime.*

Proof. Suppose R is a UFD. Then (1) holds, by the definition of a UFD. It remains to show that if x is irreducible, then x is prime. Suppose $x \mid bc$, and let $ax = bc$. Write a, b and c as products of atoms, so that

$$\underbrace{a_1 \dots a_k}_{a} x = \underbrace{b_1 \dots b_l}_{b} \underbrace{c_1 \dots c_m}_{c}.$$

Since these are two factorizations of $ax = bc$ into products of atoms and x is an atom, x must be an associate of some b_i or some c_j. Hence, $x \mid b$ or $x \mid c$. Thus, x is prime.

Conversely, suppose (1) and (2) are given. Then to show R is a UFD, it suffices to show that if

$$a_1 \dots a_r = b_1 \dots b_s$$

where the a_i and b_i are atoms, then $r = s$ and the b_i may be arranged so the a_i and b_i are associates for $i = 1, \dots, r$. The proof proceeds by induction on r.

When $r = 1$, $a_1 = b_1 \dots b_s$. Since a_1 is prime, a_1 divides b_i for some i. Assume that $a_1 \mid b_1$, and let $b_1 = ua_1$. Since b_1 is an atom, u must be a unit, so a_1 and b_1 are associates. Furthermore, $a_1 = b_1 \dots b_s = ua_1 b_2 \dots b_s$, so $1 = (ub_2) \dots b_s$. That is, $s = 1$ and $a_1 = b_1$. For the inductive step, write $a_1 \dots a_r = b_1 \dots b_s$. Since a_1 is prime, a_1 divides b_i for some i. As above, let $b_1 = ua_1$ where u is a unit and a_1 and b_1 are associates. Then $a_1 \dots a_r = b_1 \dots b_s = ua_1 b_2 \dots b_s$, so $a_2 \dots a_r = ub_2 \dots b_s$. Now the inductive hypothesis applies since we have $r - 1$ factors on the left. It follows that $r = s$ and after reordering the b_i, a_i and b_i are associates for $i = 2, \dots, r$. This completes the induction. □

To obtain more examples of a UFD and an integral domain which is not a UFD, we introduce: Let d be a square free integer. The set

$$\mathbb{Z}[\sqrt{d}] = \{x + y\sqrt{d} : x, y \in \mathbb{Z}\}$$

is a subring of $\mathbb{C}$. It is called the **ring of quadratic integers**. Note that if $x_1, x_2, y_1, y_2 \in \mathbb{Z}$ are such that $x_1 + y_1\sqrt{d} = x_2 + y_2\sqrt{d}$, then $x_1 = x_2$ and $y_1 = y_2$ because d is non-square. Define a function $N : \mathbb{Z}[\sqrt{d}] \to \mathbb{Z}$ by

$$N(x + y\sqrt{d}) = x^2 - dy^2 \quad \text{for all } x, y \in \mathbb{Z}.$$

It is called the **norm map on** $\mathbb{Z}[\sqrt{d}]$.

Theorem 2.4.5. 1. *If $\alpha \in \mathbb{Z}[\sqrt{d}]$ and $N(\alpha) = 0$, then $\alpha = 0 = 0 + 0\sqrt{d}$.*
2. *$\forall \alpha, \beta \in \mathbb{Z}[\sqrt{d}]$, $N(\alpha\beta) = N(\alpha)N(\beta)$ and $(\alpha \mid \beta \Rightarrow N(\alpha) \mid N(\beta))$.*
3. *$\forall \alpha \in \mathbb{Z}[\sqrt{d}]$, α is a unit $\Leftrightarrow N(\alpha) = \pm 1$.*
4. *If $\alpha \in \mathbb{Z}[\sqrt{d}]$ and $N(\alpha) = p$ is a prime number, then α is irreducible in $\mathbb{Z}[\sqrt{d}]$.*

Proof. Let $x, y \in \mathbb{Z}$ be such that $x^2 - dy^2 = 0$. Then $x^2 = dy^2$. If $y \neq 0$, then $d = x^2/y^2$, so $\sqrt{d} = |x/y| \in \mathbb{Q}$, which is a contradiction. Thus, we must have $y = 0$ which also forces $x = 0$. This proves (1). A direct calculation gives (2). For (3), let $\alpha \in \mathbb{Z}[\sqrt{d}]$. Suppose that α is a unit. Then $\alpha\beta = 1$ for some $\beta \in \mathbb{Z}[\sqrt{d}]$. Thus, $1 = N(1) = N(\alpha\beta) = N(\alpha)N(\beta)$, so $N(\alpha)$ divides 1 in $\mathbb{Z}$. This gives $N(1) = \pm 1$. Conversely, assume that $N(\alpha) = \pm 1$. Write $\alpha = x + y\sqrt{d}$. Then $\pm 1 = N(x+y\sqrt{d}) = x^2 - y^2 d = (x+y\sqrt{d})(x-y\sqrt{d})$ which implies that $x + y\sqrt{d}$ is a unit. Finally, (4) follows from (3). □

Example 2.4.2. The unit group of the ring $\mathbb{Z}[i]$ is $\{1, -1, i, -i\}$ where i denotes $\sqrt{-1}$.

Remark. The equation $x^2 - dy^2 = 1$ is called the **Pell's equation**. Every unit in $\mathbb{Z}[\sqrt{d}]$ is a solution of Pell's equation, or else of $x^2 - dy^2 = -1$, the **negative Pell's equation**. If $d < 0$, then $x^2 - dy^2 \geq 0$. In this case the negative Pell's equation has no solutions. In fact, Pell's equation only has very few solutions in this case, namely two, unless $d = -1$ when there are four solutions. If $d > 0$, there are infinitely many solutions to Pell's equation. The negative Pell's equation may or may not have solutions.

Example 2.4.3. Consider the ring $\mathbb{Z}[\sqrt{-5}]$.

1. $1 - \sqrt{-5}$, $1 + \sqrt{-5}$, 2 and 3 are irreducible elements.
2. $1 + \sqrt{-5}$ and 2 are not prime elements. Hence, $\mathbb{Z}[\sqrt{-5}]$ is not a UFD by Theorem 2.4.4.

Solution. (1) Assume that $1 - \sqrt{-5} = (a + b\sqrt{-5})(c + d\sqrt{-5})$ for some $a, b, c, d \in \mathbb{Z}$. By taking norms, we have

$$6 = (a^2 + 5b^2)(c^2 + 5d^2),$$

which implies that $a^2 + 5b^2 = 1, 2, 3$ or 6. Observe that $b = 0$ implies $a^2 = 1$, so $a + b\sqrt{5}$ is a unit. If $b \neq 0$, then $a^2 + 5b^2 \geq 5$, so $a^2 + 5b^2 = 6$. This forces that $c^2 + 5d^2 = 1$ and thus $c + d\sqrt{-5}$ is a unit. Hence, $1 - \sqrt{-5}$ is irreducible. Next, assume that $2 = (a + b\sqrt{-5})(c + d\sqrt{-5})$ for some $a, b, c, d \in \mathbb{Z}$. By taking norms, we have

$$4 = (a^2 + 5b^2)(c^2 + 5d^2),$$

which implies that $a^2 + 5b^2 = 1, 2$ or 4. If $a^2 + 5b^2 = 2$, then 2 is a square modulo 5 which is a contradiction. Thus, $a^2 + 5b^2 = 1$ or $c^2 + 5d^2 = 1$. Hence, $a + b\sqrt{-5}$ or $c + d\sqrt{-5}$ is a unit and so 2 is irreducible. Similarly, $1 + \sqrt{-5}$ and 3 are irreducible.
(2) Note that

$$(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6 = 2 \cdot 3.$$

Then $(1 + \sqrt{-5}) \mid 2 \cdot 3$. But, if $(1 + \sqrt{-5}) \mid 2$ or $(1 + \sqrt{-5}) \mid 3$, then $6 \mid 4$ or $6 \mid 9$, which are absurd. Thus, $1 + \sqrt{-5}$ is not a prime element. Similarly, $2 \mid (1 + \sqrt{-5})(1 - \sqrt{-5})$. If $2 \mid (1 + \sqrt{-5})$ or $2 \mid (1 - \sqrt{-5})$, then $4 \mid 6$, a contradiction. Hence, 2 is not a prime element. □

Next, we talk about common factors, gcd and lcm of elements in an integral domain (cf. Section 1.1). Let R be an integral domain and suppose $a, b \in R$. A **greatest common divisor of a and b**, $\gcd(a, b)$, is an element $d \in R$ which satisfies

1. $d \mid a$ and $d \mid b$ and 2. $\forall c \in R, (c \mid a \wedge c \mid b) \Rightarrow c \mid d$.

A **least common multiple of a and b**, $\operatorname{lcm}(a, b)$, is an element $m \in R$ which satisfies

1. $a \mid m$ and $b \mid m$ and 2. $\forall c \in R, (a \mid c \wedge b \mid c) \Rightarrow m \mid c$.

Remark. $+3$ and -3 are greatest common divisors of 12 and 15. 60 and -60 are least common multiples of 12 and 15. Thus, the gcd or lcm of two elements is not unique, (however we adopt the above notation anyway, e.g., $\gcd(12, 15) = 3$ and $\gcd(12, 15) = -3$ are both correct!). By their definitions, they are *unique up to associates* as recorded in the next theorem.

Theorem 2.4.6. *Let R be an integral domain and let $a, b \in R$.*
1. *If d and d' are gcd's of a and b, then d and d' are associates.*
2. *If m and m' are lcm's of a and b, then m and m' are associates.*

Let R be an integral domain and let $\mathcal{Q}(R)$ be the set of atoms of R. Define an equivalence relation on $\mathcal{Q}(R)$ by $a \sim b$ if a and b are associates. Then a set of representative atoms for R is a set $\mathcal{P} = \mathcal{P}(R)$ which contains exactly one atom from each equivalence class.

Example 2.4.4. $\mathcal{Q}(\mathbb{Z}) = \{\pm p \mid p \text{ is a prime}\}$ is the set of all atoms in $\mathbb{Z}$.
$\mathcal{P}(\mathbb{Z}) = \{p \mid p \text{ is a positive prime}\}$ is a set of representative atoms.
$\mathcal{P}(\mathbb{Z}) = \{+2, -3, +5, -7, \ldots\}$ is another set of representative atoms.

We obtain the next theorem directly from the definition of a UFD.

Theorem 2.4.7. *Let R be an integral domain and let $\mathcal{P}$ be a set of representative atoms for R. Then the following statements are equivalent.*
(i) R is a UFD.
(ii) Every nonzero element of R can be expressed uniquely (up to order of factors) as $a = ub_1^{i_1} \cdots b_k^{i_k}$, where u is a unit of R, $k \geq 0$, $i_1, \ldots, i_k > 0$ and $b_1, \ldots, b_k$ are distinct elements of $\mathcal{P}$.

Another important result from R being a UFD is the existence of $\gcd$ and lcm for any pair of nonzero elements. We also have the same relation for $\gcd$ and lcm as in elementary number theory.

Theorem 2.4.8. *Let R be a UFD and suppose $a, b \in R \setminus \{0\}$.*
1. *a and b have a gcd and an lcm.*
2. *Let $\mathcal{P}$ be a set of representative atoms for R. Then among the gcd's of a and b there is exactly one which is a product of elements of $\mathcal{P}$. The same is true for the lcm's of a and b.*
3. *If a and b are nonzero, $\gcd(a, b) = r$, and $\operatorname{lcm}(a, b) = s$, then ab and rs are associates. In other words,*
$$\operatorname{lcm}(a, b) = \frac{ab}{\gcd(a, b)}.$$

Proof. Let $\mathcal{P}$ be a set of representative atoms of R, and let $b_1, \ldots, b_k \in \mathcal{P}$ be all the atoms of $\mathcal{P}$ which occur in *either* a or b when they are factored as in Theorem 2.4.7. Write

$$a = ub_1^{i_1} \cdots b_k^{i_k} \quad \text{and} \quad b = vb_1^{j_1} \cdots b_k^{j_k}$$

where u and v are units and $i_s, j_s \geq 0$. Then we derive:
(a) $r = b_1^{\min(i_1, j_1)} \cdots b_r^{\min(i_r, j_r)}$ is a gcd for a and b.
(b) $s = b_1^{\max(i_1, j_1)} \cdots b_r^{\max(i_r, j_r)}$ is a lcm for a and b.
(c) r is the only gcd of a and b which is a product of elements of $\mathcal{P}$, and s is the only lcm of a and b which is a product of elements of $\mathcal{P}$.
(d) Since $i + j = \min(i, j) + \max(i, j)$ for any integers i and j,

$$ab = uvb_1^{i_1 + j_1} \cdots b_k^{i_k + j_k} = uvrs.$$

Hence, ab and rs are associates. □

Remark. Suppose R is an integral domain and $Ra + Rb = Rc$. Then $c = \gcd(a, b)$. The converse does not hold. E.g., $\mathbb{Q}[s, t]$, where s and t are indeterminates. Then $\gcd(s, t) = 1$ and $\mathbb{Q}[s, t] \neq \mathbb{Q}[s, t]s + \mathbb{Q}[s, t]t$.

Proof. Since $Rc \supseteq Ra$ and $Rc \supseteq Rb$, $c \mid a$ and $c \mid b$. Suppose $d \mid a$ and $d \mid b$. Then $Rd \supseteq Ra + Rb = Rc$, so $d \mid c$. Hence, $c = \gcd(a, b)$. □

Now, we shall prove that a PID is also a UFD. The following lemma is a key for R being a PID.

Lemma 2.4.9. [Ascending Chain Condition (ACC) for a PIR] *Let R be a principal ideal ring. If $I_1 \subseteq I_2 \subseteq \dots$ is a chain of ideals in R, then $\exists m \in \mathbb{N}, I_n = I_m$ for all $n \geq m$.*

Proof. Let $I = \bigcup_{n=1}^{\infty} I_n$. Then I is an ideal of R. Since R is a PIR, $\exists a \in R, (a) = I$. Then $a \in \bigcup_{n=1}^{\infty} I_n$, so $\exists m \in \mathbb{N}, a \in I_m$. Thus, $I = (a) \subseteq I_m \subseteq I$ which implies that $I_m = I$. Hence, $\forall n \geq m, I_n = I_m$. □

Lemma 2.4.10. *If R is a PID and a is a nonzero nonunit element in R, then there exists an atom $p \in R$ such that $p \mid a$.*

Proof. Since a is nonunit, $Ra \subsetneq R$. Then there exists a maximal ideal M of R such that $Ra \subseteq M$. Since R is a PID, $M = Rp$ for some atom p by Theorem 2.4.2. Since $Ra \subseteq Rp$, $p \mid a$. □

Theorem 2.4.11. *Every PID is a UFD.*

Proof. Let R be a PID. By Theorems 2.4.3 and 2.4.4, it suffices to show that every nonzero nonunit of R is a product of atoms. Let $a \in R$ be nonzero nonunit. By Lemma 2.4.10, there exists an atom p_1 dividing a. Write $a = p_1 b_1$ for some $b_1 \in R$. If b_1 is a unit, then a is an atom. If b_1 is nonunit, then there exists an atom p_2 dividing b_1, so we write $a = p_1 b_1 = p_1 p_2 b_2$. Continuing, we get a strictly ascending chain of ideals

$$(a) \subset (b_1) \subset (b_2) \subset \cdots .$$

Since R is a PID, this chain must terminate, by the ACC in Lemma 2.4.9, with some $b_r = p_r u_r$ where u_r is a unit and p_r is an atom. Hence, $a = p_1 p_2 \dots p_r u_r$, and so R is a UFD as desired. □

Finally, we study a generalization of the division algorithm which leads to a special kind of integral domains. An integral domain D is called a **Euclidean domain** if there exists a map

$$d : D \setminus \{0\} \to \mathbb{N} \cup \{0\},$$

called a **valuation map**, such that

1. $\forall a, b \in D \setminus \{0\}, d(a) \leq d(ab)$ and
2. $\forall a \in D, b \in D \setminus \{0\}, \exists q, r \in D, a = bq + r$ with $r = 0$ or $d(r) < d(b)$.

Examples 2.4.5.

1. Any field F is a Euclidean domain with valuation $d(a) = 1$ for all $a \neq 0$.
2. From the division algorithm for $\mathbb{Z}$ (Theorem 1.1.1), we have $\mathbb{Z}$ is a Euclidean domain if we define $d(a) = |a|$ for all $a \neq 0$.
3. The ring $\mathbb{Z}[i] = \{m + ni : m, n \in \mathbb{Z}\}$ is called the **ring of Gaussian integer**. This is a subring of $\mathbb{C}$, hence an integral domain. Its elements can be identified with the set of "lattice points", that is, points with integral coordinates in the complex plane. If $a = m + ni$, we put $d(a) = a\bar{a} = |a|^2 = m^2 + n^2$, the *norm map*. Then $d(a) \in \mathbb{N}$ and $d(ab) = d(a)d(b) \geq d(a)$ for all $a, b \in \mathbb{Z}[i] \setminus \{0\}$. To prove that d satisfies the condition of the definition of a Euclidean domain, we note that if $b \neq 0$, then $ab^{-1} = \mu + \nu i$, where μ and ν are rational numbers. Now we can find integers u and v such that $|u - \mu| \leq 1/2$, $|v - \nu| \leq 1/2$. Set $\varepsilon = \mu - u, \eta = \nu - v$, so that $|\varepsilon| \leq 1/2$ and $|\eta| \leq 1/2$. Then

$$a = b[(u + \varepsilon) + (v + \eta)i] = bq + r$$

where $q = u + vi$ is in $\mathbb{Z}[i]$. Since $r = a - bq$, $r \in \mathbb{Z}[i]$. Moreover if $r \neq 0$, then

$$d(r) = |r|^2 = |b|^2(\varepsilon^2 + \eta^2) \leq |b|^2(1/4 + 1/4) = d(b)/2.$$

Thus, $d(r) < d(b)$. Hence, $\mathbb{Z}[i]$ is a Euclidean domain.

Theorem 2.4.12. *A Euclidean domian is a PID, and hence is a UFD.*

Proof. Let I be an ideal in a Euclidean domain D. If $I = \{0\}$, we have $I = (0)$. Otherwise, let $b \neq 0$ be an element of I for which $d(b)$ is minimal for the nonzero elements of I. Let a be any element of I. Then $a = bq + r$ for some $q, r \in D$ with $r = 0$ or $d(r) < d(b)$. Since $r = a - bq \in I$ and $d(r) < d(b)$, we must have $r = 0$ by the choice of b in I. Hence, $a = bq$, so $I = (b)$. □

Example 2.4.6. Let $\theta = \frac{1}{2}(1 + \sqrt{-19})$ and $\mathbb{Z}[\theta] = \{a + b\theta : a, b \in \mathbb{Z}\}$. Assume that $u = a + b\theta$ is a unit in $\mathbb{Z}[\theta]$. Then $(a + b\theta)(c + d\theta) = 1$ for some $c, d \in \mathbb{Z}$. The squares of absolute value on both sides give

$$((2a + b)^2 + 19b^2)((2c + d)^2 + 19d^2) = 16$$

which implies $b = d = 0$ and so $ac = 1$. Hence, the unit group of $\mathbb{Z}[\theta] = \{\pm 1\}$. By a similar technique, we can show that 2 and 3 are irreducible in $\mathbb{Z}[\theta]$. Now, suppose that d is a valuation map on $\mathbb{Z}[\theta]$. Choose $m \in \mathbb{Z}[\theta]$ which is nonzero nonunit such that $d(m)$ is minimal. First, we divide 2 by m and get $q, r \in \mathbb{Z}[\theta]$ and

$$2 = mq + r \quad \text{with } d(r) < d(m) \text{ or } r = 0.$$

This means $r = 0, 1$ or -1. If $r = 0$, then $m \mid 2$ which forces $m = \pm 2$ since 2 is irreducible and m is not a unit. Similarly, if $r = -1$, then $m = \pm 3$. The case $r = 1$ cannot happen, for if it did, then $m \mid 1$, so m is a unit. Next, we divide θ by m in the same way, we obtain $q', r' \in \mathbb{Z}[\theta]$ and

$$\theta = mq' + r' \quad \text{with } d(r') < d(m) \text{ or } r' = 0.$$

Again, we have $r' = 0, 1$ or -1. Thus, one of θ, $\theta + 1$ or $\theta - 1$ is divisible by m. But $m = \pm 2$ or ± 3 and it is easy to see that none of these quotients is in R. This contradiction tell us that $\mathbb{Z}[\theta]$ is not a Euclidean domain.

Next, we shall show that $\mathbb{Z}[\theta]$ is a PID. Let I be a nonzero ideal of $\mathbb{Z}[\theta]$. Choose $b \in I$ so that $|b|$ is as small as possible. We aim to show that $I = \mathbb{Z}[\theta]b$. Suppose not. Then there is an element $a \in I \setminus \mathbb{Z}[\theta]b$. Note that $ap - bq \in I$ for all $p, q \in \mathbb{Z}[\theta]$, so if we can find p, q with $|ap - bq| < |b|$ (or equivalently $|(a/b)p - q| < 1$), then we shall be done. Since we may replace a by any element $a' = a - bq$, we can subtract any desired element of R from a/b. In particular, we can assume that the imaginary part y of $a/b = x + iy$ lies between $\pm\sqrt{19}/4$. Now, if the imaginary part of a/b lies strictly between $\pm\sqrt{3}/2$, then a/b lies at distance less than 1 from some rational integer and we are done. Thus, we may assume the imaginary part of a/b lies between $\sqrt{3}/2$ and $\sqrt{19}/4$ (or the negative of this, where the argument is similar). Hence, the imaginary part of $2(a/b) - (1 + \sqrt{-19})/2$ lies between $\sqrt{3} - \sqrt{19}/2$ and 0. But $\sqrt{19} < \sqrt{27} = 3\sqrt{3}$, so $\sqrt{3}/2 > \sqrt{19}/2 - \sqrt{3} > 0$. Therefore, the imaginary part of $2(a/b) - (1 + \sqrt{-19})/2$ is sufficient small that the complex number lies at a distance less than 1 for some rational integer. In both cases, we have found elements $p, q \in \mathbb{Z}[\theta]$ such that $|ap - bq| < |b|$ which is a contradiction. Hence, $\mathbb{Z}[\theta]$ is a PID.

Remark. In conclusion, recall that $\mathbb{Z}$ is an integral domain which is not a field and $\mathbb{Z}[\sqrt{-5}]$ is not a UFD. Besides, $\mathbb{Z}[\theta] = \{a + b\theta : a, b \in \mathbb{Z}\}$, where $\theta = (1 + \sqrt{-19})/2$, is a PID which is not a Euclidean domain as shown above. Finally, $\mathbb{Z}[x]$ (in the next section) is a UFD which is not a PID.

Exercises 2.4.

1. If p and q are prime elements in an integral domain R such that $p \mid q$, prove that p and q are associates.
2. Let R be a UFD and c a nonzero element in R. Prove that R/Rc contains a nonzero nilpotent element if and only if there is a prime element $p \in R$ with $p^2 \mid c$.
3. Let R be a UFD. If $a \in R$ is a nonzero nonunit element, prove that Ra is *the* product of a finite number of prime ideals.

4. If R is a PID and $\gcd(a,b)=1$, show that $Ra+Rb=R$, so $1=ax+by$ for some $x,y\in R$.
5. Let R be a PID and suppose that a,b and c are nonzero elements of R such that $Ra+Rb=Rc$. Show that there exist $u,v\in R$ such that $ua+vb=c$ and $Ru+Rv=R$.
6. Prove that $4+\sqrt{10}$ is irreducible but not prime in the ring $\{a+b\sqrt{10}:a,b\in\mathbb{Z}\}$. Deduce that $\mathbb{Z}[\sqrt{10}]$ is not a UFD.
7. Show that the ring $\mathbb{Z}[\sqrt{2}]$ has infinitely many units. (*Hint.* If u is a unit, so is u^n for all $n\in\mathbb{Z}$.)
8. (a) Let D be a Euclidean domain. Prove that u is a unit in D if and only if $d(u)=d(1)$.
   (b) Show that ±1 and $\pm i$ are units in $\mathbb{Z}[i]$ and prove that if $a+bi$ is not a unit in $\mathbb{Z}[i]$, then $a^2+b^2>1$.
9. Let R be a Euclidean ring and $a,b\in R$, $b\neq0$. Prove that there exist $q_0,q_1\ldots,q_n$ and $r_1,\ldots,r_n$ in R such that

$$\begin{aligned}
a&=q_0b+r_1, \quad d(r_1)<d(b),\\
b&=q_1r_1+r_2, \quad d(r_2)<d(r_1),\\
r_1&=q_2r_2+r_3, \quad d(r_3)<d(r_2),\\
&\cdots \quad \cdots \quad \cdots\\
r_{n-2}&=q_{n-1}r_{n-1}+r_n, \quad d(r_n)<d(r_{n-1}),\\
r_{n-1}&=q_nr_n
\end{aligned}$$

   and if a and b satisfy the above conditions, then r_n is a gcd of a and b. This algorithm is called the **Euclidean algorithm**. Find a gcd of $8+6i$ and $5-15i$ in $\mathbb{Z}[i]$ by using the Euclidean algorithm.
10. Let D be a UFD with field of fractions F and suppose $\alpha\in F$. Show that it is possible to write $\alpha=a/b$ with $a,b\in D$ and $\gcd(a,b)=1$.
11. Let R be a PID with field of fractions F, and let S be a ring with $R\subseteq S\subseteq F$.
    (a) If $\alpha\in S$, show that $\alpha=a/b$ with $a,b\in R$ and $1/b\in S$. (b) Prove that S is a PID.
12. Let $R=\{m/2^n:m,n\in\mathbb{Z}\text{ and }n\geq0\}$.
    (a) Prove that R is a subring of $\mathbb{Q}$ and determine all units of R.
    (b) Show that 3 is an irreducible element in R.
    (c) Prove that R is a PID.

Project 12 (Prime elements in the ring of Gaussian integers)**.** We have learned that all units in $\mathbb{Z}[i]$ are ±1, $\pm i$. In this project, we shall determine all prime elements in $\mathbb{Z}[i]$. Use the norm map, show that up to multiplication by units, the prime elements in $\mathbb{Z}[i]$ are of three types:

(a) p, where p is a prime in $\mathbb{Z}$ satisfying $p\equiv3\pmod4$,
(b) π or $\bar{\pi}$, where $q=\pi\bar{\pi}$ is a prime in $\mathbb{Z}$ satisfying $q\equiv1\pmod4$,
(c) $\alpha=1+i$.

Project 13 (Quadratic norm Euclidean domains)**.** Find all square free integers $d\equiv2,3\pmod4$ such that the norm map on $\mathbb{Z}[\sqrt{d}]$ satisfies the axiom of a Euclidean function. [*Answer*. They are -2, -1, 2, 3, 6, 7, 11, 19, 33.]

Moreover, for a square free integer $d\equiv1\pmod4$, let

$$\mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right]=\left\{\frac{x+y\sqrt{d}}{2}:x,y\in\mathbb{Z},a\equiv b\pmod2\right\}.$$

Define the **norm map** on $\mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right]$ by

$$N\left(\frac{x+y\sqrt{d}}{2}\right)=\frac{x^2-dy^2}{4}.$$

Find all square free integers $d\equiv1\pmod4$ such that the norm map on $\mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right]$ satisfies the axiom of a Euclidean function. [*Answer*. They are -11, -7, -3, 5, 13, 17, 21, 29, 37, 41, 57, 73. Note that $\mathbb{Z}\left[\frac{1+\sqrt{69}}{2}\right]$ is a Euclidean domain but not for norm.]

2.5 Polynomial Rings

One of familiar topic in elementary algebra is "polynomials". Algebraic equations usually involve factorization of polynomials. Here, we treat them in a more abstract way with the things that we have studied from the previous section. There will be many important results in this section.

2.5.1 Polynomials and Their Roots

Let R be a ring with identity 1 and let x be a symbol called an **indeterminate**, not representing any element in R. Let $R[x]$ denote the set of all symbols $a_0 + a_1x + \cdots + a_nx^n$ where $n \in \mathbb{N} \cup \{0\}, a_i \in R, x^0 = 1, x^1 = x$. For $i \in \mathbb{N}$, let x^i denote $1 \cdot x^i$. In the symbol $a_0 + a_1x + \cdots + a_nx^n$, we may drop a_ix^i if $a_i = 0$. Each element $a_0 + a_1x + \cdots + a_nx^n$ is called a **polynomial** and a_i is called the **coefficient** of x^i for $i \in \{1, \ldots, n\}$ and a_0 is called the **constant term**.

For $p(x) = a_0 + a_1x + \cdots + a_nx^n$ and $q(x) = b_0 + b_1x + \cdots + b_mx^m$ in $R[x]$, we can write $p(x) = a_0 + a_1x + \cdots + a_kx^k$ and $q(x) = b_0 + b_1x + \cdots + b_kx^k$ where $k \geq \max\{m, n\}, a_i = 0$ if $i > n$ and $b_j = 0$ if $j > m$ and we define

1. $p(x) = q(x) \Leftrightarrow a_i = b_i$ for all $i \in \{0, 1, \ldots, k\}$
2. $p(x) + q(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_k + b_k)x^k$ and
3. $p(x)q(x) = c_0 + c_1x + \cdots + c_{n+m}x^{n+m}$, where $c_l = \sum_{i=0}^{l} a_ib_{l-i} (= a_0b_l + a_1b_{l-1} + \cdots + a_lb_0)$ for all $l \in \{0, 1, \ldots, n + m\}$.

Hence, under the operation defined above $R[x]$ is a ring which has 1 as its identity and contains R as a subring (considered elements as constant polynomials). The ring $R[x]$ is called the **ring of polynomials over** R. If R is commutative, so is $R[x]$.

Set $R[x_1, x_2] = R[x_1][x_2]$ and $R[x_1, \ldots, x_n] = R[x_1, \ldots, x_{n-1}][x_n]$ if $n > 2$.

If $p(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$ and $a_n \neq 0$, then the **degree of** $p(x)$, denoted by $\deg p(x)$, is defined to be n. For technical reasons, we define the degree of the zero polynomial to be $-\infty$ and adopt the following conventions: $(-\infty) < n$ and $(-\infty) + n = -\infty = n + (-\infty)$ for every integer n; $(-\infty) + (-\infty) = -\infty$.

Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$. If $a_n \neq 0$, a_n is called the **leading coefficient** of $f(x)$ and $f(x)$ is a **monic polynomial** if $a_n = 1$. If R is commutative and $c \in R$, then

$$f(x) \mapsto f(c) := a_0 + a_1c + \cdots + a_nc^n$$

gives a homomorphism from $R[x]$ to R, called the **evaluation at** c. In addition, if $f(c) = 0$, then c is called a **root** of $f(x)$.

The following statements are clearly true.

1. Every unit in R is a unit in $R[x]$.
2. If $f(x) = a_0 + a_1x + \cdots + a_mx^m, g(x) = b_0 + b_1x + \cdots + b_nx^n \in R[x]$ and $a_mb_n \neq 0$, then $\deg(f(x)g(x)) = m + n$.

In particular, if R is an integral domain, we have:

Theorem 2.5.1. *Let R be an integral domain.*

1. *$R[x]$ is an integral domain.*
2. *$\forall f(x), g(x) \in R[x] \setminus \{0\}, \deg f(x)g(x) = \deg f(x) + \deg g(x)$.*
3. *The set of all units of $R[x]$ is the set of all units of R. In particular, $\mathcal{U}(\mathbb{Z}[x]) = \{\pm 1\}$ and $\mathcal{U}(F[x]) = F \setminus \{0\}$, where F is a field.*
4. *$\forall a \in R$, a is irreducible in $R \Leftrightarrow a$ is irreducible in $R[x]$.*
5. *$\forall a, b \in R$, b is a unit $\Rightarrow a + bx$ is irreducible in $R[x]$.*

Proof. (1) and (2) are clear from the above observation. Note that if $f(x)$ is a unit, then $f(x)g(x) = 1$ for some $g(x) \in R[x]$, so $\deg f(x) + \deg g(x) = \deg 1 = 0$ by (2). This forces that $\deg f(x) = 0 = \deg g(x)$ which implies that $f(x)$ lies in R and (3) follows. Next, let $a \in R$. If a is irreducible in $R[x]$, then a is clearly irreducible in R. On the other hand, if $a = f(x)g(x)$ for some

nonzero nonunits $f(x)$ and $g(x)$ in $R[x]$, we have $0 = \deg a = \deg f(x) + \deg g(x)$, so this again gives $\deg f(x) = \deg g(x) = 0$. This means that $f(x)$ and $g(x)$ indeed lie in R, and thus a is reducible in R. Finally, let $a, b \in R$ with b a unit. Then $\deg(a + bx) = 1$. If $a + bx = f(x)g(x)$ for some $f(x), g(x) \in R[x]$, then $1 = \deg f(x) + \deg g(x)$, so $f(x)$ or $g(x)$ must lie in R, say $f(x) = c$ a constant in R and $g(x) = u + vx$. Since $b = cv$ is a unit, c is a unit. Hence, $a + bx$ is irreducible. □

Theorem 2.5.2. [Division Algorithm] *Let R be a ring, $f(x), g(x) \in R[x]$ and $g(x) \neq 0$. Assume that the leading coefficient of $g(x)$ is a unit in R. Then $\exists$ unique $q(x), r(x) \in R[x]$ such that $f(x) = q(x)g(x) + r(x)$ where $r(x) = 0$ or $\deg r(x) < \deg g(x)$.*

Proof. If there exists an $h(x) \in R[x]$ such that $f(x) = h(x)g(x)$, let $q(x) = h(x)$ and $r(x) = 0$. Assume that $f(x) \neq h(x)g(x)$ for all $h(x) \in R[x]$. Let

$$S = \{\deg(f(x) - h(x)g(x)) : h(x) \in R[x]\} \subseteq \mathbb{N} \cup \{0\}.$$

Then $S \neq \emptyset$. By the Well-Ordering Principle, there exists a polynomial $q(x)$ in $R[x]$ such that $\deg(f(x) - q(x)g(x))$ has the least degree and we may write $r(x)$ for $f(x) - q(x)g(x)$. Then $r(x) \neq 0$. Assume that $\deg r(x) \geq \deg g(x)$. Write $r(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_n \neq 0$, and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ with b_m a unit. Since $\deg r(x) \geq \deg g(x)$, $n - m \geq 0$. Let $s(x) = r(x) - a_nb_m^{-1}x^{n-m}g(x)$. Thus, $\deg s(x) < n$ and

$$s(x) = f(x) - q(x)g(x) - a_nb_m^{-1}x^{n-m}g(x) = f(x) - (q(x) - a_nb_m^{-1}x^{n-m})g(x),$$

so $s(x) \in S$ and $\deg s(x) < \deg r(x)$, a contradiction.

To prove that $q(x)$ and $r(x)$ are unique, suppose that $q_2(x)$ and $r_2(x)$ are polynomials such that

$$f(x) = g(x)q_2(x) + r_2(x) \quad \text{where } r_2(x) = 0 \text{ or } \deg r_2(x) < \deg g(x).$$

Then $g(x)q(x) + r(x) = f(x) = g(x)q_2(x) + r_2(x)$. Subtracting yields

$$g(x)[q(x) - q_2(x)] = r_2(x) - r(x).$$

Since the leading coefficient of $g(x)$ is assumed to be a unit, we have

$$\deg(g(x)[q(x) - q_2(x)]) = \deg g(x) + \deg(q(x) - q_2(x)).$$

Since $\deg(r_2(x) - r(x)) < \deg g(x)$, this relation can hold only if $q(x) - q_2(x)$ is zero, i.e., $q(x) = q_2(x)$, and hence finally $r(x) = r_2(x)$. □

For a field R, the leading coefficient of a nonzero polynomial $g(x)$ in $R[x]$ is always a unit in R, so the division algorithm above gives:

Corollary 2.5.3. *If F is a field, then $F[x]$ is a Euclidean domain with valuation $d(p(x)) = \deg p(x)$ for all $p(x) \in F[x] \setminus \{0\}$. Moreover, $F[x]$ is a PID and a UFD.*

Theorem 2.5.4. [Remainder Theorem] *Let R be a ring and $f(x) \in R[x]$. Then for all $c \in R$, the remainder when $x - c$ divides $f(x)$ is $f(c)$.*

Proof. Let $c \in R$. By Theorem 2.5.2, there exist unique $q(x) \in R[x]$ and $r \in R$ such that $f(x) = q(x)(x - c) + r$. Then $f(c) = q(c)(c - c) + r = r$. □

Corollary 2.5.5. *Let R be a ring.*

1. *If $f(x) \in R[x]$, $c \in R$ and $f(c) = 0$, then $f(x) = q(x)(x - c)$ for some $q(x) \in R[x]$.*
2. *If R is commutative, $f(x) \in R[x]$ and $c \in R$, then $f(c) = 0 \Leftrightarrow (x - c) \mid f(x)$ in $R[x]$.*
3. *Let R be an integral domain, $f(x) \in R[x]$, $\deg f(x) = 2$ or 3 and the leading coefficient of $f(x)$ is a unit in R. Then $f(x)$ has a root in $R \Leftrightarrow f(x)$ is reducible in $R[x]$.*

Proof. (1) and (2) are clear. For (3), assume that c is a root of $f(x)$. Then $f(x) = q(x)(x-c)$ for some $q(x) \in R[x]$. Since $\deg f(x)$ is 2 or 3, $\deg q(x)$ is 1 or 2, so $f(x)$ is reducible. Conversely, suppose that $f(x) = g(x)h(x)$, where $g(x), h(x) \in R[x]$ of degree ≥ 1. Since $\deg f(x) = 2$ or 3, $\deg g(x) = 1$ or $\deg h(x) = 1$. Hence, $f(x)$ has a root in R. □

Examples 2.5.1.
1. $x^2 - 3$ is irreducible over $\mathbb{Q}$ but not over $\mathbb{R}$.
2. $x^2 + 1$ is irreducible over $\mathbb{R}$ but not over $\mathbb{C}$ since $x^2 + 1 = (x - i)(x + i)$.
3. $x^3 - x + 1$ is irreducible over $\mathbb{Z}_3$ but reducible over $\mathbb{R}$ by the intermediate value theorem. In general, every polynomial of odd degree over $\mathbb{R}$ has a root in $\mathbb{R}$ (Theorem 5.5.3).
4. $x^4 + 4$ has no roots in $\mathbb{R}$ but it can be factored as $(x^2 - 2x^2 + 2)(x^2 + 2x^2 + 2)$ in $\mathbb{R}[x]$.

Corollary 2.5.6. *Let F be a field.*
1. *If $f(x)$ is a polynomial over F of degree n, then $f(x)$ has at most n roots in F.*
2. *If $f(x)$ and $g(x)$ are polynomials over F of degree $\leq n$ such that $f(\alpha_1) = g(\alpha_1), \ldots, f(\alpha_{n+1}) = g(\alpha_{n+1})$ where $\alpha_1, \ldots, \alpha_{n+1}$ are distinct elements of F, then $f(x) = g(x)$.*
3. *If F is infinite and $f(x)$ and $g(x)$ are polynomials over F such that $f(\alpha) = g(\alpha)$ for all $\alpha \in F$, then $f(x) = g(x)$.*

Proof. We shall prove (1) by induction on $k = \deg f(x)$. It is clear when $f(x)$ is linear. Assume that $k > 1$ and any polynomials of degree k have at most k roots in F. Suppose that $f(x)$ is of degree $k + 1$. The statement is true when $f(x)$ has no root in F. Otherwise, let α be a root of $f(x)$ in F. Then $f(x) = (x - \alpha)q(x)$ for some polynomial $q(x) \in F[x]$ of degree k. By the inductive hypothesis, $q(x)$ has at most k roots. Hence, $f(x)$ has at most $k+1$ roots. The remaining statements follow from the first one. □

Remarks.
1. $f(x) = x^2 - 1$ has four roots in $\mathbb{Z}_{12}$, namely $1, -1, 5, -5$.
2. Corollary 2.5.6 says that two polynomials over an infinite field F which defined the same function on F are identical. This is NOT true if F is *finite*. Let $F = \mathbb{Z}_p$, $f(x) = x$ and $g(x) = x^p$. Then $f(\alpha) = g(\alpha)$ for all $\alpha \in \mathbb{Z}_p$ but $f(x) \neq g(x)$.

Theorem 2.5.7. *Let F be a field and $F[x]$ the polynomial ring over F. Then linear polynomials are the only atoms in $F[x]$ if and only if each polynomial $f(x) \in F[x]$ of positive degree has a root in F.*

Proof. Suppose that linear polynomials are the only atoms in $F[x]$. Let $f(x)$ be a polynomial of positive degree over F. Since $F[x]$ is a UFD, $f(x) = \alpha_1(x) \cdots \alpha_k(x)$, a product of atoms. Each $\alpha_i(x)$ is linear, so $\alpha_i(x) = b_i(x - c_i)$ ($b_i, c_i \in F$ with $b_i \neq 0$). Then $(x - c_i) \mid f(x)$, so $c_1, \ldots, c_k$ are roots of $f(x)$ in F. Conversely, assume that every $f(x) \in F[x]$ of positive degree has a root in F. Let $\alpha(x)$ be an atom in $F[x]$. We claim that $\alpha(x)$ is linear. For, let $b \in F$ be a root of $\alpha(x)$. Then $(x - b) \mid \alpha(x)$ so $\alpha(x) = (x - b)\beta(x)$ for some $\beta(x) \in F[x]$. Since $\alpha(x)$ is an atom, $\beta(x)$ must be a unit. That is, $\beta(x)$ is a constant lying in $F \setminus \{0\}$. Thus, $\alpha(x)$ is a linear polynomial. □

A field F is an **algebraically closed field** if every non-constant polynomial has a root in F.

Example 2.5.2. By the fundamental theorem of algebra (Theorem 5.5.6), the only atoms in $\mathbb{C}[x]$ are linear polynomials. Thus, $\mathbb{C}$ is an algebraically closed field.

Theorem 2.5.8. *Let R be an integral domain and $f(x) \in R[x]$ a nonzero polynomial. If $\alpha_1, \ldots, \alpha_k$ are distinct roots of $f(x)$, then $(x - \alpha_1) \ldots (x - \alpha_k)$ divides $f(x)$.*

Proof. We shall prove this result by induction of k. Corollary 2.5.5 (1) gives the basis step. Assume $k > 1$. By the inductive hypothesis $(x - \alpha_1) \dots (x - \alpha_{k-1})$ divides $f(x)$, so let $f(x) = (x - \alpha_1) \dots (x - \alpha_{k-1})g(x)$ for some $g(x) \in R[x]$. Then

$$0 = f(\alpha_k) = (\alpha_k - \alpha_1) \dots (\alpha_k - \alpha_{k-1})g(\alpha_k).$$

Thus, $g(\alpha_k) = 0$ since R is an integral domain, so $(x-\alpha_k) \mid g(x)$. It follows that $(x-\alpha_1) \dots (x-\alpha_k)$ divides $f(x)$. □

2.5.2 Factorizations in Polynomial Rings

When factor a polynomial, we first look for some common factors on its coefficients. For example, $2x^3+4 = 2(x^3+2)$. Taking the gcd of the coefficients out allows us to concentrate on polynomials with no common factor on their coefficients and leads to the following definitions.

Let R be a UFD and suppose that $f(x) = a_0 + a_1x + \cdots + a_nx^n$ is a nonzero polynomial in $R[x]$. The **content** of $f(x)$ is the gcd of $a_0, \dots, a_n$. We say that $f(x)$ is **primitive** if the content of $f(x)$ is unit in R, i.e., $a_0, \dots, a_n$ have no common factor except units.

Theorem 2.5.9. [Gauss' lemma] *Let R be a UFD and $f(x), g(x) \in R[x]$. If $f(x)$ and $g(x)$ are primitive, so is $f(x)g(x)$.*

Proof. Let

$$\begin{aligned} f(x) &= a_0 + a_1x + \cdots + a_mx^m \\ g(x) &= b_0 + b_1x + \cdots + b_nx^n \\ f(x)g(x) &= c_0 + c_1x + \cdots + c_{m+n}x^{m+n}. \end{aligned}$$

We shall suppose that $f(x)g(x)$ is not primitive and obtain a contradiction. Let $a \in R$ be an atom of R which divides all of $c_0, \dots, c_{m+n}$. Since R is a UFD, every atom is a prime, so Ra is a prime ideal. Then $(R/Ra)[x]$ is an integral domain. Since $R[x]/R[x]a \cong (R/Ra)[x]$, $R[x]a$ is a prime ideal. Let

$$^- : R[x] \to R[x]/R[x]a$$

be the canonical map. Since a divides $c_0, \dots, c_{m+n}$, $\bar{f}(x)\bar{g}(x) = 0$. But a does not divide all of $a_0, \dots, a_m$ or all of $b_0, \dots, b_n$, since f and g are primitive. Thus, $\bar{f}(x) \neq 0, \bar{g}(x) \neq 0$. This is a contradiction since $\bar{f}(x)\bar{g}(x) = 0$ and $\bar{f}(x), \bar{g}(x)$ lie in $R[x]/P$ which is an integral domain. Hence, $f(x)g(x)$ is primitive, as claimed. □

Theorem 2.5.10. *Let R be a UFD and $f(x), g(x)$ nonzero polynomials of $R[x]$. Then:*

1. *$f(x)$ is primitive $\Leftrightarrow$ the content of $f(x)$ is 1.*
2. *If a is the content of f, then $f(x) = af_1(x)$ where $f_1(x)$ is primitive.*
3. *If $f(x) = af_1(x)$ and $f_1(x)$ is primitive, then a is the content of $f(x)$.*
4. *If a and b are the contents of $f(x)$ and $g(x)$, respectively, then ab is the content of $f(x)g(x)$.*

Proof. (1), (2) and (3) are immediate from the definition of gcd. For the last statement, by (2), we write $f(x) = af_1(x)$ and $g(x) = bg_1(x)$ where $f_1(x)$ and $g_1(x)$ are primitive. By Gauss' lemma, $f_1(x)g_1(x)$ is primitive, and

$$f(x)g(x) = af_1(x)bg_1(x) = (ab)(f_1(x)g_1(x)).$$

Hence, ab is the content of $f(x)g(x)$, by (3). □

Theorem 2.5.11. *Let R be a UFD and let $F = Q(R) = \{r/s : r, s \in R, s \neq 0\}$ be its field of quotients. Suppose $f(x)$ is an irreducible polynomial in $R[x]$. Then $f(x)$, considered as a polynomial in $F[x]$, is irreducible in $F[x]$. In particular, if $f(x) \in \mathbb{Z}[x]$ is irreducible over $\mathbb{Z}$, it is irreducible over $\mathbb{Q}$.*

Proof. Suppose $f(x) = g(x)h(x)$ where $g(x)$ and $h(x)$ are polynomials of positive degree in $F[x]$. Let $g(x) = a_0/b_0 + (a_1/b_1)x + \cdots + (a_m/b_m)x^m$ and $h(x) = c_0/d_0 + (c_1/d_1)x + \cdots + (c_n/d_n)x^n$. Let b be a least common multiple of the b_i and d a least common multiple of the d_j so that

$$g_1(x) = bg(x) \quad \text{and} \quad h_1(x) = dh(x)$$

lie in $R[x]$. Then

$$bdf(x) = bg(x)dh(x) = g_1(x)h_1(x).$$

By Theorem 2.5.10, let $g_1(x) = ug_2(x)$ and $h_1(x) = vh_2(x)$ where u is the content of $g_1(x)$ and v is the content of $h_1(x)$, and $g_2(x)$ and $h_2(x)$ are primitive polynomials in $R[x]$. Thus,

$$bdf(x) = g_1(x)h_1(x) = uvg_2(x)h_2(x).$$

Since $g_2(x)$ and $h_2(x)$ are primitive, so is $g_2(x)h_2(x)$ and hence the equation above implies that $bd \mid uv$ in R. Canceling, we obtain

$$f(x) = wg_2(x)h_2(x) \quad \text{where } w = \frac{uv}{bd} \in R.$$

Therefore, $f(x)$ is reducible in $R[x]$, which proves the theorem. □

Let R be a UFD and $F = Q(R)$ its field of quotients. Suppose

$$h(x) = a_0/b_0 + (a_1/b_1)x + \cdots + (a_n/b_n)x^n \in F[x],$$

where $a_0/b_0, a_1/b_1, \ldots, a_n/b_n$ are in "lowest terms". That is, a_i and b_i have no common factor. Let $b = \operatorname{lcm}(b_0, \ldots, b_n)$. Then

$$bh(x) = a_0(b/b_0) + a_1(b/b_1)x + \cdots + a_n(b/b_n)x^n$$

is in $R[x]$. Let a be the content of $bh(x)$. It happens that $a = \gcd(a_0, \ldots, a_n)$, although knowing this is not essential. The main point is that

$$h_1(x) = (b/a)h(x)$$

is a primitive polynomial in $R[x]$. Moreover, the proof of Theorem 2.5.11 shows that if $f(x) \in R[x]$, then $h(x) \mid f(x)$ in $F[x] \Leftrightarrow h_1(x) \mid f(x)$ in $R[x]$. In particular, suppose $f(x) \in R[x]$, and $r/s \in F$ is a root of $f(x)$ where r and s are relatively prime. Then $h(x) = x - (r/s)$ divides $f(x)$ in $F[x]$, so $h_1(x) = sx - r$ divides $f(x)$ in $R[x]$. Thus, we have:

Theorem 2.5.12. *Let R be a UFD and F its field of quotients. Suppose $f(x) \in R[x]$ where $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $r/s \in F$ is a root of $f(x)$ where r and s are relatively prime. Then $s \mid a_n$ and $r \mid a_0$ if $r \neq 0$.*

Proof. The remarks above show that $(sx - r) \mid (a_0 + a_1x + \cdots + a_nx^n)$ in $R[x]$. It is easy to see that this implies our results. □

Remarks. 1. Suppose $f(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}[x]$, where $a_0 \neq 0$. Then there are only finitely many rationals which can possibly be roots of $f(x)$, namely the fractions r/s where $r \mid a_0$ and $s \mid a_n$.

2. Note that if $a_n = 1$ above, then $s = \pm 1$ and $r/s \in \mathbb{Z}$. In other words, if $a_n = 1$, then every rational root of $f(x)$ is an integer.

Another important criterion on irreduciblity in $\mathbb{Q}[x]$ is the next theorem.

Theorem 2.5.13. [Eisenstein's Criterion] *Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ lies in $\mathbb{Z}[x]$, and suppose that there is a prime number p such that*
1. $p \nmid a_n$, 2. $p \mid a_0, \ldots, a_{n-1}$, and 3. $p^2 \nmid a_0$.
Then $f(x)$ is irreducible in $\mathbb{Q}[x]$. Moreover, if $f(x)$ is primitive, then $f(x)$ is irreducible in $\mathbb{Z}[x]$.

Proof. We shall suppose that $f(x)$ is reducible in $\mathbb{Q}[x]$ and obtain a contradiction. By dividing $f(x)$ by its content, we may assume that $f(x)$ is primitive, this does not affect either the hypothesis or the reducibility of $f(x)$ in $\mathbb{Q}[x]$. By Theorem 2.5.11, $f(x)$ is reducible in $\mathbb{Z}[x]$, so let $f(x) = g(x)h(x)$ where $g(x) = b_0 + b_1x + \cdots + b_mx^m$ and $h(x) = c_0 + c_1x + \cdots + c_{n-m}x^{n-m}$ are in $\mathbb{Z}[x]$. Note that since $f(x)$ is primitive, neither $g(x)$ nor $h(x)$ is constant. That is, $m \geq 1$ and $n - m \geq 1$. Let $^- : \mathbb{Z}[x] \to \mathbb{Z}_p[x]$ be the canonical projection. Then $\bar{f}(x) = \bar{a}_nx^n$ where $\bar{a}_n \neq \bar{0}$ since $p \nmid a_n$, so $\bar{g}(x)\bar{h}(x) = \bar{f}(x) = \bar{a}_nx^n$. Since $\mathbb{Z}_p[x]$ is a UFD, this forms $\bar{g}(x) = \bar{b}_mx^m, \bar{h}(x) = \bar{c}_{n-m}x^{n-m}$, so that $\bar{b}_0 = \bar{c}_0 = \bar{0}$ (i.e., p divides b_0 and c_0). But then $p^2 \mid a_0$ since $a_0 = b_0c_0$, which contradicts part (3) of the hypotheses. Hence, $f(x)$ is irreducible in $\mathbb{Q}[x]$ as claimed. □

Example 2.5.3. $f(x) = 2x^5 - 6x^3 + 9x^2 - 15$ is irreducible in $\mathbb{Q}[x]$ and in $\mathbb{Z}[x]$.

Corollary 2.5.14. *The p***th cyclotomic polynomial**

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible in $\mathbb{Q}[x]$ for any prime p.

Proof. The polynomial

$$g(x) = \Phi_p(x + 1) = \frac{(x + 1)^p - 1}{(x + 1) - 1} = x^{p-1} + \binom{p}{1}x^{p-2} + \cdots + p$$

satisfies the Eisenstein criterion for the prime p and is thus irreducible in $\mathbb{Q}[x]$. But clearly if $\Phi_p(x) = h(x)r(x)$ were a nontrivial factorization of $\Phi_p(x)$ in $\mathbb{Z}[x]$, then

$$\Phi_p(x + 1) = g(x) = h(x + 1)r(x + 1)$$

would give a nontrivial factorization of $g(x)$ in $\mathbb{Z}[x]$. Thus, $\Phi_p(x)$ must also be irreducible in $\mathbb{Q}[x]$. □

We next wish to prove a famous theorem of Gauss: if R is a UFD, so is $R[x]$. Recall the criterion given in Theorem 2.4.4:

an integral domain is a UFD ⇔
1. every nonzero nonunit is a product of atoms and
2. every atom is prime.

Suppose R is a UFD. We first observe that $R[x]$ is an integral domain, so this presents no problem. We shall establish the criteria above for $R[x]$ (and these show that $R[x]$ is a UFD) by doing three things:
(a) We determine all atoms of $R[x]$ (Theorem 2.5.15).
(b) We show that they are primes (Theorem 2.5.16).

(c) We show that every nonzero nonunit of $R[x]$ is a product of atoms and conclude that $R[x]$ is a UFD (Theorem 2.5.17).

Theorem 2.5.15. *Let R be a UFD, F its field of quotients and $f(x) \in F[x]$. Then $f(x)$ is an atom of $R[x] \Leftrightarrow$ either*
(1) $f(x) \in R$ and $f(x)$ is an atom of R or
(2) $f(x)$ is a primitive polynomial of degree $n \geq 1$ and $f(x)$ is irreducible in $F[x]$.

Proof. Assume that $f(x)$ is an atom of $R[x]$. If $\deg f(x) = 0$, then $f(x) \in R$, and clearly $f(x)$ must be an atoms of R. Otherwise, suppose that $\deg f(x) = n \geq 1$, and let a be the content of $f(x)$. Then $f(x) = af_1(x)$ where $f_1(x)$ is primitive. Since $f(x)$ is irreducible in $R[x]$, a must be a unit in R, so $f(x)$ is primitive. Again, since $f(x)$ is irreducible in $R[x]$, it is also irreducible in $F[x]$ by Theorem 2.5.11.

Conversely, assume that (1) and (2) hold. If $f(x)$ is an atom of R, it is clearly an atom of $R[x]$ (Theorem 2.5.1). Suppose $f(x)$ is a primitive polynomial of degree $n \geq 1$ and $f(x)$ is irreducible in $F[x]$. We claim that $f(x)$ is an atom of $R[x]$. For, suppose not, and let

$$f(x) = g(x)h(x),$$

where $g(x)$ and $h(x)$ are nonunits of $R[x]$.
(a) If $g(x)$ or $h(x)$ lies in R, then $f(x)$ is not primitive, a contradiction.
(b) If $g(x)$ and $h(x)$ both have positive degree, then $f(x)$ is reducible in $F[x]$, again a contradiction.

Hence, if $f(x)$ is an atom of $R[x]$, it has the form (1) or (2), as required. □

Theorem 2.5.16. *Let R be a UFD and $f(x)$ an atom of $R[x]$. Then $R[x]f(x)$ is a prime ideal of $R[x]$. That is, $f(x)$ is a prime element.*

Proof. We consider separately the two types of atoms in $R[x]$ given in Theorem 2.5.15.
Case 1. $a \in R$ is an atom of R. Since R is a UFD, every atom is a prime, so Ra is a prime ideal. Then $(R/Ra)[x]$ is an integral domain. Since $R[x]/R[x]a \cong (R/Ra)[x]$, $R[x]a$ is a prime ideal, so a is prime.
Case 2. $f(x)$ is a primitive polynomial of degree $n \geq 1$ and $f(x)$ is irreducible in $F[x]$ where F is the quotient field of R. First we claim that $F[x]f(x) \cap R[x] = R[x]f(x)$. Clearly, $f(x) \in F[x]f(x) \cap R[x]$. Conversely, we suppose $g(x)f(x) \in R[x]$ with $g(x) = a_0/b_0 + (a_1/b_1)x + \cdots + (a_n/b_n)x^n \in F[x]$. We can find relatively prime $a, b \in R$ such that $(b/a)g(x) = g_1(x)$ where $g_1(x)$ is a primitive polynomial in $R[x]$. (In fact, $a = \gcd(a_0, a_1, \ldots, a_n)$ and $b = \operatorname{lcm}(b_0, b_1, \ldots, b_n)$ will do, provided each a_i and b_i are relatively prime.) Thus, $(b/a)g(x)f(x) = g_1(x)f(x) \in R[x]$. By Gauss' lemma, $g_1(x)f(x)$ is a primitive polynomial. In connection with the above equation, this forces b to be a unit of R, so $g(x) = (a/b)g_1(x) \in R[x]$. Hence, $g(x)f(x) \in R[x]f(x)$ which proves our claim.

By the second isomorphism theorem, we have

$$R[x]/R[x]f(x) = R[x]/(R[x] \cap F[x]f(x)) \cong (R[x] + F[x]f(x))/F[x]f(x).$$

Since $(R[x] + F[x]f(x))/F[x]f(x) \subseteq F[x]/F[x]f(x)$ which is a field because $f(x)$ is irreducible in $F[x]$, $R[x]+F[x]f(x))/F[x]f(x)$ is an integral domain. Thus, $R[x]/R[x]f(x)$ is an integral domain, so $R[x]f(x)$ is a prime ideal. Therefore, $f(x)$ is prime and this proves the theorem. □

Theorem 2.5.17. [Gauss] *If R is a UFD, so is $R[x]$. Hence, if R is a UFD, so is $R[x_1, \ldots, x_n]$ for all $n \in \mathbb{N}$.*

Proof. We know all the atoms of $R[x]$ by Theorem 2.5.15 and Theorem 2.5.16 tells us that each atom of $R[x]$ is prime. Hence, (by Theorem 2.4.4) to verify that $R[x]$ is a UFD, it remains to show that each nonzero nonunit $f(x) \in R[x]$ is a product of atoms.
Case 1. $\deg f(x) = 0$, i.e., $f(x) \in R$. Since R is a UFD and every atom of R is an atom of $R[x]$, we can express $f(x)$ as a product of atoms in R, and so in $R[x]$.
Case 2. $\deg f(x) = n \geq 1$. Let $f(x) = f_1(x) \dots f_k(x)$ where (a) each $f_i(x)$ has degree ≥ 1 and (b) k is as large as possible. Such a factorization exists because any factorization which satisfies (a) has at most n terms since $n = \deg f(x) = \deg f_1(x) + \cdots + \deg f_k(x) \geq k$. Now, let a_i be the content of $f_i(x)$, and let $f_i(x) = a_i g_i(x)$ where $g_i(x)$ is a primitive polynomial.

We claim that $g_i(x)$ is an atom in $R[x]$ because if $g_i(x) = r(x)s(x)$ where $r(x)$ and $s(x)$ are nonunits, then $r(x)$ and $s(x)$ cannot lie in R, since $g_i(x)$ is primitive. In addition, $r(x)$ and $s(x)$ cannot both have positive degree because then we could write $f(x)$ as a product of $k+1$ polynomials of positive degree, which violates (b). Thus, each $g_i(x)$ is an atom as desired. Hence,

$$\begin{aligned} f(x) &= f_1(x) \dots f_k(x) \\ &= a_1 g_1(x) \dots a_k g_k(x) \\ &= a_1 \dots a_k g_1(x) \dots g_k(x) \\ &= a g_1(x) \dots g_k(x), \text{ where } a \in R. \end{aligned}$$

By Case 1, a can be written as a product of atoms in $R[x]$ and therefore shows that $f(x)$ is a product of atoms in $R[x]$, which proves $R[x]$ is a UFD. □

Example 2.5.4. Since $\mathbb{Z}$ is a UFD, we have $\mathbb{Z}[x]$ is also a UFD. However, in the following exercises, we shall know that the ideal $(x, 2)$ of $\mathbb{Z}[x]$ is not principal, so $\mathbb{Z}[x]$ is an example of a UFD which is not a PID.

Exercises 2.5.

1. Let R be a ring.
   (a) Show that $M_n(R[x]) \cong M_n(R)[x]$ for all $n \in \mathbb{N}$, where x is an indeterminate in both cases.
   (b) If I is an ideal of R, prove that $I[x]$, the set of all polynomials with coefficients in I, is an ideal of $R[x]$ and $R[x]/I[x] \cong (R/I)[x]$.
2. Prove the following statements.
   (a) If R is an integral domain, then x is a prime element in $R[x]$.
   (b) In $\mathbb{Z}[x]$, (x) is a prime ideal but not a maximal ideal.
   (c) If F is a field, then (x) is a maximal ideal in $F[x]$.
3. Let F be a field. Show that $F[x]/(x^2) \cong \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} : a, b \in F \right\}$, a subring of $M_2(F)$.
4. Let F be a field. Find a ring isomorphism $F[x]/(x^2 - x) \to F \times F$.
5. Prove that the ideal $I = (x, 2)$ is not a principal ideal of $\mathbb{Z}[x]$. Hence $\mathbb{Z}[x]$ is not a PID. In addition, show that I is a maximal ideal in $\mathbb{Z}[x]$.
6. Construct a field of: (a) 125 elements (b) 81 elements.
7. Find all odd prime numbers p such that $x + 2$ is a factor of $x^4 + x^3 + x^2 - x + 1$ in $\mathbb{Z}_p[x]$.
8. Let $p(x) \in \mathbb{R}[x]$. Prove that if $p(a + bi) = 0$, then $p(a - bi) = 0$ for all $a, b \in \mathbb{R}$. Deduce by the fundamental theorem of algebra that there exist real numbers $c, r_1, \dots, r_k, a_1, b_1, \dots, a_m, b_m$ such that

   $$p(x) = c(x - r_1) \dots (x - r_k)(x^2 - (2a_1)x + (a_1^2 + b_1^2)) \dots (x^2 - (2a_m)x + (a_m^2 + b_m^2)).$$

   In addition, if $p(x) \in \mathbb{R}[x]$ is irreducible over $\mathbb{R}$, then $\deg p(x) = 1$ or 2, namely, $p(x) = bx + c$ or $p(x) = ax^2 + bx + c$ with $b^2 - 4ac < 0$.
9. Let p be an odd prime. Prove that $x^n - p$ is irreducible over $\mathbb{Z}[i][x]$.
10. If R is an intergral domain for which every ideal of $R[x]$ is principal, show that R must be a field.
11. Let D be an integral domain. If $\varphi : D[x] \to D[x]$ is an automorphism such that $\varphi(a) = a$ for all $a \in D$, prove that there exist $c, d \in D$ with c a unit in D such that $\varphi(x) = cx + d$. Here x stands for the indeterminate of $D[x]$.

12. Let R be a UFD and F its field of quotients. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ lies in $R[x]$, and suppose that there is an irreducible element $p \in R$ such that
(i) $p \nmid a_n$, (ii) $p \mid a_0, \ldots, a_{n-1}$, and (iii) $p^2 \nmid a_0$.
Prove that $f(x)$ is irreducible in $F[x]$. Moreover, if $f(x)$ is primitive, then $f(x)$ is irreducible in $R[x]$.

Project 14 (Units in a polynomial ring)**.** Let R be a commutative ring and $f(x) = a_0 + a_1x + \cdots + a_nx^n$ in $R[x]$. Prove that $f(x)$ is a unit in $R[x]$ if and only if a_0 is a unit in R and $a_1, \ldots, a_n$ are nilpotent elements in R. (This project generalizes the result in Theorem 2.5.1 (3) to any commutative ring.)

Project 15 (Generalized Eisenstein's criterion)**.** Let $P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ be a polynomial with integer coefficients. If there exist a prime number p and an integer $k \in \{0, 1, \ldots, n-1\}$ such that $p \mid a_0, a_1, \ldots, a_k, p \nmid a_{k+1}$ and $p^2 \nmid a_0$, then $P(x)$ has an irreducible factor in $\mathbb{Z}[x]$ of degree greater than k. Extend this result to a UFD similar to the last question of Exercises 2.5.

2.6 Field Extensions

Let F be a field and $f(x)$ a polynomial over F of degree $n \in \mathbb{N}$. Then the quotient ring

$$\begin{aligned} F[x]/(f(x)) &= \{g(x) + (f(x)) : g(x) \in F[x]\} \\ &= \{g(x) + (f(x)) : g(x) \in F[x] \text{ and } g(x) = 0 \text{ or } \deg g(x) < n\} \\ &= \{a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + (f(x)) : a_i \in F\} \end{aligned}$$

by the division algorithm. Thus, if F is a finite field, then $F[x]/(f(x))$ is a commutative ring of $|F|^n$ elements. In addition, if $f(x)$ is irreducible in $F[x]$, then $(f(x))$ is a maximal ideal, so $F[x]/(f(x))$ is a field. Note that F is isomorphic to $\{c + (f(x)) : c \in F\}$, so we may embed F into $F[x]/(f(x))$ by using the inclusion map.

Examples 2.6.1. 1. $\mathbb{R}[x]/(x^2+1)$ is a field isomorphic to $\mathbb{C}$ (with the map $f(x) \mapsto f(i)$).
2. $\mathbb{Z}_{11}[x]/(x^2+3)$ is a field of 121 elements.

Under the above idea, if a polynomial over a field does not possess a a root in its own field, we shall create a bigger field where we can find a root of it.

2.6.1 Algebraic and Transcendental Extensions

Before we extend a field, we first determine the smallest one possible according to its characteristic. Let F be a field. The intersection of all subfields of F is the smallest subfield of F, called the **prime field** of F.

Theorem 2.6.1. *Let F be a field with the prime subfield P and 1_F denote the identity of F.*
1. *If $\operatorname{char} F = p$, a prime, then $P = \{n \cdot 1_F : n = 0, 1, \ldots, p-1\} \cong \mathbb{Z}/p\mathbb{Z}$.*
2. *If $\operatorname{char} F = 0$, then $P = \{(m \cdot 1_F)(n \cdot 1_F)^{-1} : m, n \in \mathbb{Z}, n \neq 0\} \cong \mathbb{Q}$.*

Proof. Since P is a field, $1_F \in P$, so $\{n \cdot 1_F : n \in \mathbb{Z}\} \subseteq P$. Define $\varphi : \mathbb{Z} \to P$ by $\varphi(n) = n \cdot 1_F$ for all $n \in \mathbb{Z}$. Then φ is a ring homomorphism and $\operatorname{im}\varphi = \{n \cdot 1_F : n \in \mathbb{Z}\}$, so $\mathbb{Z}/\ker\varphi \cong \operatorname{im}\varphi$.
(1) Assume that $\operatorname{char} F = p$ is a prime. Then $\operatorname{im}\varphi = \{n \cdot 1_F : n = 0, 1, \ldots, p-1\}$ and p is the smallest positive integer such that $p \in \ker\varphi$, so $\ker\varphi = p\mathbb{Z}$. Hence, $\operatorname{im}\varphi \cong \mathbb{Z}/p\mathbb{Z}$ which is a field, so $P = \operatorname{im}\varphi \cong \mathbb{Z}/p\mathbb{Z}$.

(2) Assume that $\operatorname{char} F = 0$. Then φ is a monomorphism. Since $\{n \cdot 1_F : n \in \mathbb{Z}\} \subseteq P$ and P is a subfield of F, $\{(m \cdot 1_F)(n \cdot 1_F)^{-1} : m, n \in \mathbb{Z}, n \neq 0\} \subseteq P$. Define $\bar{\varphi} : \mathbb{Q} \to P$ by $\bar{\varphi}(m/n) = \varphi(m)\varphi(n)^{-1}$ for all $m, n \in \mathbb{Z}, n \neq 0$. Then $\bar{\varphi}$ is a monomorphism and $\bar{\varphi}|_{\mathbb{Z}} = \varphi$. Thus, $\mathbb{Q} \cong \operatorname{im} \bar{\varphi} = \{(m \cdot 1_F)(n \cdot 1_F)^{-1} : m, n \in \mathbb{Z}, n \neq 0\}$ which is a subfield of P, and hence they are equal. □

In this section, we require some background in vector spaces. A field K is said to be an **extension** of a field F if F is a subring of K.

Remark. By Theorem 2.6.1, any field can be considered as an extension field of the field $\mathbb{Q}$ or $\mathbb{Z}_p$ for some prime p.

Let K be an extension field of F. The **degree** of K over F, $[K : F]$, is the dimension of K as a vector space over F. More generally, if a field F is a subring of a ring R, then $[R : F]$ is the dimension of R as a vector space over F. For example, $[\mathbb{C} : \mathbb{R}] = 2$ and $[\mathbb{R} : \mathbb{Q}]$ is infinite (in fact $[\mathbb{R} : \mathbb{Q}] = |\mathbb{R}|$).

Theorem 2.6.2. *If $[L : K]$ and $[K : F]$ are finite, then $[L : F]$ is finite and*

$$[L : F] = [L : K][K : F].$$

In fact, $[L : F] = [L : K][K : F]$ whenever $F \subseteq K \subseteq L$.

Proof. With $F \subseteq K \subseteq L$, let $\{\beta_j\}_{j \in J}$ be a basis of K over F and $\{\alpha_i\}_{i \in I}$ a basis of L over K. Every element of L can be written uniquely as a finite linear combination of the elements of $\{\alpha_i\}_{i \in I}$ with coefficients in K, and every such coefficient can be written uniquely as a finite linear combination of the elements of $\{\beta_j\}_{j \in J}$ with coefficients in F. Hence, every element of L can be written uniquely as a finite linear combination of the elements of $\{\alpha_i \beta_j\}_{i \in I, j \in J}$ with coefficients in K: $\{\alpha_i \beta_j\}_{i \in I, j \in J}$ is a basis of L over F, and $[L : F] = |I \times J| = [L : K][K : F]$. □

Notation. Let K be an extension field of F.

1. If $t_1, \ldots, t_n$ are indeterminates over F, then $F(t_1, \ldots, t_n)$ denotes the field of quotients of the polynomial ring $F[t_1, \ldots, t_n]$.
2. If $u_1, \ldots, u_n \in K$ (or $S \subseteq K$), then $F[u_1, \ldots, u_n]$ (or $F[S]$) denotes the subring of K generated by F and $u_1, \ldots, u_n$ (or S), and $F(u_1, \ldots, u_n)$ (or $F(S)$) denotes its field of quotients.

Theorem 2.6.3. [Classification of Elements in an Extension Field] *Let K be a field extension of a field F and let $u \in K$. Then EITHER*

(a) $[F(u) : F] = \infty$ and $F[u] \cong F[t]$, so $F(u) \cong F(t)$ where t is an indeterminate OR

(b) $[F(u) : F]$ is finite and $F[u] = F(u)$.

Proof. Let t be an indeterminate and consider the ring homomorphism

$$F[t] \xrightarrow{\varphi} K$$

defined by $\varphi(t) = u$ (or $\varphi(f(t)) = f(u)$). Note that the kernel of φ is a prime ideal, since the image of φ has no zero divisors. There are two possibilities.
(1) $\ker \varphi = \{0\}$. Then we have (a).
(2) $\ker \varphi \neq \{0\}$. Then $\ker \varphi = F[t]g(t)$ where $g(t)$ is a monic prime (i.e., irreducible) polynomial. Since $F[t]$ is a PID, $F[t]g(t)$ is a maximal ideal. Thus,

$$F[u] \cong F[t]/F[t]g(t)$$

is a field, so $F[u] = F(u)$. □

Remarks. 1. If $g(t) = g_0 + g_1 t + \cdots + g_{n-1}t^{n-1} + t^n$, then $[F(u) : F] = n$ and $\{1, u, \ldots, u^{n-1}\}$ is a basis for $F(u)$ over F.

2. Consider $\mathbb{R} \subset \mathbb{C}$ and $g(t) = g_0 + g_1 t + t^2 \in \mathbb{R}[t]$. We distinguish three cases.
   (a) If $g_1^2 - 4g_0 > 0$, then $g(t) = (t - a)(t - b)$ where $a, b \in \mathbb{R}, a \neq b$ and $\mathbb{R}[t]/\mathbb{R}[t]g(t)$ is a ring without nonzero nilpotent elements.
   (b) If $g_1^2 - 4g_0 = 0$, then $g(t) = (t - a)^2$ and $\mathbb{R}[t]/\mathbb{R}[t]g(t)$ is a ring with nonzero nilpotent elements.
   (c) If $g_1^2 - 4g_0 < 0$, then $\mathbb{R}[t]/\mathbb{R}[t]g(t) \cong \mathbb{C}$.

3. If p is a prime, then $t^2 - p$ is irreducible over $\mathbb{Q}$ and the fields $\mathbb{Q}(\sqrt{p}) \cong \mathbb{Q}[t]/(t^2 - p)$ are all distinct.

   Proof. Let p and q be distinct primes. Assume that $\varphi : \mathbb{Q}[\sqrt{p}] \to \mathbb{Q}[\sqrt{q}]$ is an isomorphism. Then $\varphi(1) = 1$, and so $\varphi(r) = r$ for all r in $\mathbb{Q}$. Let $\varphi(\sqrt{p}) = a + b\sqrt{q}$ for some $a, b \in \mathbb{Q}$. Thus, $p = \varphi(p) = (\varphi(\sqrt{p}))^2 = (a + b\sqrt{q})^2 = (a + bq) + 2ab\sqrt{q}$. Since $\sqrt{q}$ is not rational, $ab = 0$. However, if $a = 0$, then $p = bq$ which implies $q \mid p$. If $b = 0$, then $\varphi(\sqrt{p}) = a = \varphi(a)$, so $\sqrt{p} = a$ is rational. Hence, both cases lead to a contradiction. Therefore, $\mathbb{Q}[\sqrt{p}]$ and $\mathbb{Q}[\sqrt{q}]$ are not isomorphic. □

An element in an extension field can be classified according to Theorem 2.6.3 as follows.

Let K be an extension field of a field F. An element $u \in K$ is **algebraic** over F in case there exists a nonzero polynomial $f(t) \in F[t]$ such that $f(u) = 0$ and **transcendental element** over F otherwise.

For example, every complex number is algebraic over $\mathbb{R}$; $\sqrt[3]{2}$ and $1 + \sqrt{5} \in \mathbb{R}$ are algebraic over $\mathbb{Q}$. It has been proved that e and $\pi \in \mathbb{R}$ are transcendental over $\mathbb{Q}$ (by the Lindemann-Weierstrass theorem). We show the existence of real numbers transcendental over $\mathbb{Q}$ in Corollary 5.2.3. Moreover, most of real numbers are in fact transcendental over $\mathbb{Q}$ (see Exercises 2.6). Theorem 2.6.3 yields characterizations of algebraic and transcendental elements.

Corollary 2.6.4. *Let K be an extension field of a field F and $u \in K$. The following conditions on u are equivalent:*
(i) u is transcendental over F (if $f(t) \in F[t]$ and $f(u) = 0$, then $f = 0$);
(ii) $F(u) \cong F(t)$;
(iii) $[F(u) : F]$ is infinite.

Corollary 2.6.5. *Let K be an extension field of a field F and $u \in K$. The following conditions on u are equivalent:*
(i) u is algebraic over F (there exists a polynomial $0 \neq f(t) \in F[t]$ such that $f(u) = 0$);
(ii) there exists a monic irreducible polynomial $g(t) \in F[t]$ such that $g(u) = 0$;
(iii) $[F(u) : F]$ is finite.
Moreover, in part (ii), we have $g(t)$ is unique; $f(u) = 0$ if and only if $g(t) \mid f(t)$; $F(u) \cong F[t]/(g(t))$; and $[F(u) : F] = \deg g(t)$.

When u is algebraic over F, the unique *monic irreducible* polynomial $g(t) \in F[t]$ in part (ii) is the **minimal polynomial of** u. The **degree of** u **over** F is $\deg g(t)$.

An extension field K of a field F is **algebraic** in case every element of K is algebraic over F. For example, $\mathbb{C}$ is an algebraic extension of $\mathbb{R}$, but $\mathbb{R}$ is not algebraic over $\mathbb{Q}$. Note that if $[K : F]$ is finite, then K is an algebraic extension because $[F(u) : F] \leq [K : F] < \infty$ for all $u \in K$. An extension field E of a field F is said to be a **simple extension** of F if $E = F(\alpha)$ for some $\alpha \in E$.

Example 2.6.2. Prove that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$ is a simple extension.

Solution. Let $K = \mathbb{Q}(\sqrt{2} + \sqrt{3})$. Since $\sqrt{2} + \sqrt{3} \in \mathbb{Q}(\sqrt{2}, \sqrt{3})$, we have $K \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3})$. For another inclusion, note that $(\sqrt{2} + \sqrt{3})^2 = 5 + 2\sqrt{2}\sqrt{3}$, so $\sqrt{2}\sqrt{3} \in K$. Thus,

$$\sqrt{2} = (\sqrt{2} + \sqrt{3})\sqrt{2}\sqrt{3} - 2(\sqrt{2} + \sqrt{3}) \quad \text{and} \quad \sqrt{3} = 3(\sqrt{2} + \sqrt{3}) - (\sqrt{2} + \sqrt{3})\sqrt{2}\sqrt{3}$$

are in K. Hence, $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = K$. □

Theorem 2.6.6. *If L is an algebraic extension of K and K is an algebraic extension of F, then L is algebraic extension over F.*

Proof. Let $u \in L$. Since L is algebraic over K, there exists $f(x) = a_0 + a_1x + \cdots + a_nx^n \in K[x]$ such that $f(u) = 0$. Since K is algebraic over F, $a_0, a_1, \ldots, a_n$ are algebraic over F, so $[F(a_0, a_1, \ldots, a_n) : F]$ is finite. For, let $E = F(a_0, a_1, \ldots, a_n)$. Then

$$[E : F] = [F(a_0) : F] \prod_{i=1}^{n} [F(a_0, a_1, \ldots, a_i) : F(a_0, a_1, \ldots, a_{i-1})],$$

a_0 is algebraic over F and a_i is algebraic over $F(a_0, \ldots, a_{i-1})$ for all $i \in \{1, \ldots, n\}$. Since $f(x) \in E[x]$, u is algebraic over E, so $[E(u) : E]$ is finite by Corollary 2.6.5. Thus,

$$[F(u) : F] \leq [E(u) : F] = [E(u) : E][E : F] < \infty.$$

Hence, u is algebraic over F. □

Corollary 2.6.7. *For $a, b \in K$, if a and b are algebraic over F of degree m and n, respectively, then $a \pm b$, ab and a/b (if $b \neq 0$) are all algebraic over F of degree $\leq mn$. Hence, the set of all algebraic elements of K over F is a subfield of K and is an algebraic extension over F.*

Proof. By Corollary 2.6.5, $[F(a) : F] = m$ and $[F(b) : F] = n$. Since b is algebraic over F, b is algebraic over $F(a)$, so $[F(a)(b) : F(a)] \leq n$. Thus, by Theorem 2.6.2, $[F(a, b) : F] = [F(a)(b) : F] = [F(a)(b) : F(a)][F(a) : F] \leq mn$. Since $a \pm b$, ab, ab^{-1} (if $b \neq 0$) are in $F(a, b)$ which is a finite extension, they are all algebraic over F of degree $\leq mn$. □

Example 2.6.3. Consider $\mathbb{Q} \subset \mathbb{C}$. Let $A = \{z \in \mathbb{C} : z \text{ is algebraic over } \mathbb{Q}\}$. By Corollary 2.6.7, A is algebraic over $\mathbb{Q}$. Assume that $[A : \mathbb{Q}] = n$ is finite. Let $f(x) = x^{n+1} - 3$. It is irreducible over $\mathbb{Q}$ by Eisenstien's criterion. Let $\alpha \in \mathbb{C}$ be such that $f(\alpha) = 0$. Then $\alpha \in A$ ans so $\mathbb{Q} \subset \mathbb{Q}[\alpha] \subset A$. But $[\mathbb{Q}[\alpha] : \mathbb{Q}] = n + 1 > [A : \mathbb{Q}]$, which is a contradiction. Hence, $[A : \mathbb{Q}]$ is infinite. This provides an example of infinite algebraic field extensions.

2.6.2 More on Roots of Polynomials

We conclude this chapter by working more on roots of polynomials. The theorem of Kronecker assures us that we may obtain an extension field of F in which the polynomial $p(x) \in F[x]$ has a root.

Theorem 2.6.8. *If F is a field and G is a finite subgroup of the multiplicative group of nonzero elements of F, then G is a cyclic group. In particular, the multiplicative group of all nonzero elements of a finite field is cyclic.*

Proof. If $G = \{1\}$, then G is cyclic. Assume that $G \neq \{1\}$. Since G is a finite abelian group,

$$G \cong \mathbb{Z}/(m_1) \oplus \cdots \oplus \mathbb{Z}/(m_k)$$

where $k \geq 1, m_1 > 1$ and $m_1 \mid \cdots \mid m_k$. Since $m_k \sum_{i=1}^{k} \mathbb{Z}/(m_i) = 0$, u is a root of the polynomial $x^{m_k} - 1 \in F[x]$ for all $u \in G$. By Corollary 2.5.6, this polynomial has at most m_k distinct roots in F, so $|G| \leq m_k$. Hence, we must have $k = 1$ and $G \cong \mathbb{Z}/(m_1)$ which is a cyclic group. □

Remark. The finite multiplicative subgroup of a division ring may not be cyclic. E.g., $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ is a subgroup of the ring of real quaternions $\mathbb{H}$ and Q_8 is not cyclic.

Let R be an integral domain and $f(x) \in R[x]$. If α is a root of $f(x)$, then there exist $m \in \mathbb{N}$ and $g(x) \in R[x]$ such that $f(x) = (x-\alpha)^m g(x)$ and $g(\alpha) \neq 0$. m is called the **multiplicity** of the root α of $f(x)$ and if $m > 1$, α is called a **multiple root** of $f(x)$.

If $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in R[x]$, we define $f'(x) \in R[x]$, the **derivative of** $f(x)$, to be the polynomial

$$f'(x) = a_1 + a_2 x + \cdots + n a_n x^{n-1}.$$

We record the immediate properties of the derivative of polynomials in the next lemma

Lemma 2.6.9. *If $f(x)$ and $g(x)$ are polynomials over an integral domain R and $c \in R$, then*

1. $(cf(x))' = cf'(x)$,
2. $(f(x) + g(x))' = f'(x) + g'(x)$,
3. $(f(x)g(x))' = f(x)g'(x) + f'(x)g(x)$,
4. $((f(x))^n)' = n(f(x))^{n-1} f'(x)$ *where* $n \in \mathbb{N}$.

Characterizations of polynomials with multiple roots using derivatives are as follows.

Theorem 2.6.10. *Let E be an extension of a field F and $f(x) \in F[x]$.*

1. *For $\alpha \in E$, α is a multiple root of $f(x)$ if and only if α is a root of both $f(x)$ and $f'(x)$.*
2. *If $f(x)$ and $f'(x)$ are relatively prime, then $f(x)$ has no multiple root.*
3. *If $f(x)$ is irreducible over F having a root in E, then $f(x)$ has no multiple root in E if and only if $f'(x) \neq 0$.*

Proof. (1) is clear.
(2) Since $f(x)$ and $f'(x)$ are relatively prime, there exist $h(x)$ and $k(x)$ in $F[x]$ such that $1 = f(x)h(x) + f'(x)k(x)$. If $\alpha \in E$ is a multiple root of $f(x)$, by (1), $f(\alpha) = 0 = f'(\alpha)$, so $1 = 0$, a contradiction.
(3) Since $f(x)$ is irreducible, $f'(x) \neq 0$ and $\deg f'(x) < \deg f(x)$, we have $f(x)$ and $f'(x)$ are relatively prime, so $f(x)$ has no multiple roots. Conversely, if $f'(x) = 0$, then $f(\alpha) = 0 = f'(\alpha)$ for some $\alpha \in E$ since $f(x)$ has a root in E. Hence, by (1), α is a multiple root of $f(x)$. □

Theorem 2.6.11. [Number of Roots] *If $f(x) \in F[x]$ and $\deg f(x) = n > 1$, then $f(x)$ can have at most n roots counting multiplicities in any extension field of F.*

Proof. We shall prove the theorem by induction on the degree of $f(x)$. If $\deg f(x) = 1$, then $f(x) = ax + b$ for some $a, b \in F$ and $a \neq 0$. Then $-b/a$ is the unique root of $f(x)$ and $-b/a \in F$, so we are done.

Let $\deg f(x) = n > 1$ and assume that the result is true for all polynomials of degree $< n$. Let E be an extension field of F. If $f(x)$ has no roots in E, then we are done. Let $r \in E$ be a root of $f(x)$ of multiplicity $m \geq 1$. Then there exists $q(x) \in E[x]$ such that $f(x) = (x-r)^m q(x)$ and $q(r) \neq 0$. Then $\deg q(x) = n - m$. By the inductive hypothesis $q(x)$ has at most $n - m$ roots in E counting multiplicities. Hence, $f(x)$ has at most $m + (n-m)$ roots in E counting multiplicities. □

Theorem 2.6.12. [Kronocker] *If $p(t) \in F[t]$ is irreducible in $F[t]$, then there exists an extension field E of F such that $[E : F] = \deg p(t)$ and $p(t)$ has a root in E.*

Proof. We use the discussion at the beginning of the section to prove this theorem. Let $E = F[x]/(p(x))$ where x is an indeterminate. Since $p(x)$ is irreducible, E is a field containing $\{a + (p(x)) : a \in F\}$ as a subfield. But $F \cong \{a + (p(x)) : a \in F\}$ by $\varphi : a \mapsto a + (p(x))$, so E can be considered as an extension field of F by considering a as $a + (p(x))$ for all $a \in F$. Then $E = F[x]/(p(x)) = F(\bar{t})$ where $\bar{t} = x + (p(x))$ is a root of $p(t)$. Since $E = F(\bar{t})$ and $p(t)$ is irreducible over F, $[E : F] = [F(\bar{t}) : F] = \deg p(t)$ by Corollary 2.6.5. □

Corollary 2.6.13. *If $p(t) \in F[t]$ is a nonconstant polynomial, then there exists a finite extension field E of F containing a root of $p(t)$ and $[E : F] \leq \deg p(t)$.*

Proof. Since $F[t]$ is a UFD, $p(t)$ has an irreducible factor in $F[t]$, say $p_1(t)$. By Theorem 2.6.12, there exists an extension field E of F such that E contains a root of $p_1(t)$ and $[E : F] = \deg p_1(t)$. Hence, $[E : F] \leq \deg p(t)$ and E contains a root of $p(t)$. □

Exercises 2.6.

1. If $u \in K$ is algebraic of odd degree over F, prove that $F(u^2) = F(u)$.
2. Let $a, b \in K$ be algebraic over F of degree m and n, respectively. Prove that if m and n are relatively prime, then $[F(a, b) : F] = mn$.
3. Show that the degree of $\sqrt{2} + \sqrt{3}$ over $\mathbb{Q}$ is 4 and the degree of $\sqrt{2} + \sqrt[3]{5}$ is 6.
4. Let p be a prime and let $v \in \mathbb{C}$ satisfy $v \neq 1$, $v^p = 1$ (e.g., $v = \cos(2\pi/p) + i \sin(2\pi/p)$). Show that $[Q(v) : Q] = p - 1$.
5. Let $E = \mathbb{Q}(u)$ where $u^3 - u^2 + u + 2 = 0$. Express $(u^2 + u + 1)(u^2 - u)$ and $(u - 1)^{-1}$ in the form $au^2 + bu + c$ where $a, b, c \in \mathbb{Q}$.
6. Let E be an algebraic extension of a field F. Show that any subring of E/F is a subfield. Hence prove that any subring of a finite dimensional extension field E/F is a subfield.
7. Let $E = F(u)$, u transcendental and let $K \neq F$ be a subfield of E/F. Show that u is algebraic over K.
8. Let u and v be positive irrational numbers such that u is algebraic over $\mathbb{Q}$ and v is transcendental over $\mathbb{Q}$.
   (a) Show that v is transcendental over $\mathbb{Q}[u]$.
   (b) Classify whether the following elements are algebraic or transcendental over $\mathbb{Q}$.
   (i) $\dfrac{1}{u + v}$ (ii) $\sqrt{u}$ (iii) $\sqrt{v}$
9. (a) Show that there are countably many irreducible polynomials in $\mathbb{Q}[x]$.
   (b) Let A be the set of all real numbers that are algebraic over $\mathbb{Q}$. Show that A is countable, so that $\mathbb{R} \setminus A$ is uncountable.
10. Let R be an integral domain and $f(x)$ a nonconstant polynomial. Prove that:
    (a) If $\operatorname{char} R = 0$, then $f'(x) \neq 0$.
    (b) If $\operatorname{char} R = p$, a prime, then $f'(x) = 0 \Leftrightarrow \exists a_0, a_1, \ldots, a_n \in R, f(x) = a_0 + a_1 x^p + \cdots + a_n x^{np}$.
11. Suppose that F is a finite field and $f(x) \in F[x]$ a nonconstant. If $f'(x) = 0$, prove that $f(x)$ is reducible over F.
12. Let F be a finite field with q elements. Prove that if K is an extension field of F and $b \in K$ is algebraic over F, then $b^{q^m} = b$ for some $m \in \mathbb{N}$.
13. A complex number α is called an **algebraic integer** if it is a root of a monic polynomial
$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$
whose coefficients are in $\mathbb{Z}$. Let $A = \{\alpha \in \mathbb{C} : \alpha \text{ is an algebraic integer}\}$. Prove that $A \cap \mathbb{Q} = \mathbb{Z}$.
14. Let $f(x) = x^2 + x + 2$ be a polynomial in $\mathbb{Z}_3[x]$ and $E = \mathbb{Z}_3[x]/(f(x))$.
    (a) Show that $f(x)$ is irreducible in $\mathbb{Z}_3[x]$, and so E is a field of 9 elements extending $\mathbb{Z}_3$.
    (b) Find the characteristic of E and $[E : \mathbb{Z}_3]$.
    (c) Find the multiplicative inverse of $1 + x + (f(x))$.
    (d) How many generators of the cyclic multiplicative group $E \setminus \{0\}$?
15. Let E_1 and E_2 be subfields of a field K. The **composite field** of E_1 and E_2, denoted by E_1E_2, is the smallest subfield of K containing both E_1 and E_2. Prove that if $[K : F]$ is finite, then $[E_1E_2 : F] \leq [E_1 : F][E_2 : F]$.
16. Let α be an irrational number. If α is a common root of $f(x) = x^3 + ax + b$ and $g(x) = x^2 + cx + d$ for some $a, b, c, d \in \mathbb{Q}$, prove that:
    (a) $g(x)$ is a factor of $f(x)$ (b) c is a root of $f(x)$.

Project 16 (Galois ring)**.** Let p be a prime and $n, r \in \mathbb{N}$. Let $f(x)$ be a monic irreducible polynomial in $\mathbb{Z}_p[x]$ of degree r. Consider this polynomial as a polynomial in $\mathbb{Z}_{p^n}[x]$.

(a) Prove that the quotient ring $R = \mathbb{Z}_{p^n}[x]/(f(x))$ is a commutative ring of p^{nr} elements that contains the ring $\mathbb{Z}_{p^n} \cong \{c + (f(x)) : c \in \mathbb{Z}_{p^n}\}$ as a subring.
(b) Prove by the first isomorphism theorem that

$$R/(p + (f(x))) \cong \mathbb{Z}_p[x]/(f(x)).$$

Deduce that $M = (p + (f(x))$ is a maximal ideal of R.
(c) Prove that $R \setminus M$ is the unit group $R^\times$. Conclude that M is the unique maximal ideal of R and so R is a local ring.

The ring $\mathbb{Z}_{p^n}[x]/(f(x))$ is called a **Galois ring**. It is a ring extension of the ring $\mathbb{Z}_{p^n}$ similar to a Galois field that is a field extension of the field $\mathbb{Z}_p$. This finite ring has many parallel properties to the finite field and has many applications in algebraic graph theory and algebraic coding theory.

Project 17 (More on the exponents)**.** Let G be a finite group. Recall the exponent of G defined before Theorem 1.7.13 in order to obtain information on the structure of a finite abelian group. Prove that:
(a) $\exp G = \operatorname{lcm}\{o(a) : a \in G\}$ where $o(a)$ is the order of a in G.
(b) If $G = G_1 \times G_2$, then $\exp G = \operatorname{lcm}(\exp G_1, \exp G_2)$.
(c) Let $n \geq 2$. Clearly, $\exp \mathbb{Z}_n = n$. Compute $\exp \mathbb{Z}_n^\times$. The exponent of $\mathbb{Z}_n^\times$ is the **Carmichael λ-function** which was first introduced in 1910 (see [22]).
(d) Find the exponent of the unit group of the Galois ring in Project 16.
(e) Somer and Křížek [40, 41] used the Carmichael λ-function as a main tool to study the digraph of the kth power mapping of $\mathbb{Z}_n$. Meemark and Wiroonsri [36, 37] replaced it with the exponent of the group to obtain a general way to study this graphs. Their work on this digraphs influenced many articles. Unfortunately, they did not have the formula for the exponent explicitly. Let q be a prime power and $f(x)$ be a monic polynomial in $\mathbb{F}_q[x]$ of degree ≥ 1. Compute the exponent of the unit group of the ring $\mathbb{F}_q[x]/(f(x))$.
(f) Compute the exponent of the unit group of the ring $\mathbb{Z}_{p^n}[x]/(f(x)^m)$ where $m \in \mathbb{N}$ and $f(x)$ is a monic irreducible polynomial in $\mathbb{Z}_p[x]$ of degree r considered as a polynomial in $\mathbb{Z}_{p^n}[x]$.
(g) [Open] Determine the exponent of the unit group of a finite local ring.

More ring theory will be in terms of modules in Chapter 4. We shall classify extension fields and talk about the fundamental theorem of Galois theory in Chapter 5.

3 | Advanced Group Theory

Deeper results of groups are presented in this chapter. Various kinds of series of a group are studied in the first three sections. A solvable group gets its name from the Galois group of a polynomial $p(x)$ and solvability by radicals of the equation $p(x) = 0$. A nilpotent group can be considered as a generalization of an abelian group. A linear group gives an example of an infinite simple group. Finally, we discuss how to construct a group from a set of objects and presentations.

3.1 Jordan-Hölder Theorem

The ideas of normal series of a group and solvability that arose in Galois theory yield invariants of groups (the Jordan-Höder theorem), showing that simple groups are, in a certain sense, building towers of finite groups.

A **subnormal series of a group** G is a finite sequence $H_0, H_1, \ldots, H_n$ of subgroups of G such that $H_i \triangleleft H_{i+1}$ (although not necessarily normal in G) for all i with $H_0 = \{e\}$ and $H_n = G$. The groups H_{i+1}/H_i are called the **factors** associated with the series. A subnormal series is called a **normal series of** G if $H_i \triangleleft G$ for all i.

Examples 3.1.1.
1. $\{0\} < 8\mathbb{Z} < 4\mathbb{Z} < \mathbb{Z}$ and $\{0\} < 9\mathbb{Z} < \mathbb{Z}$ are normal series of $\mathbb{Z}$.
2. $\{(1)\} < A_3 < S_3$ is a normal series of S_3.
3. $\{(1)\} < A_4 < S_4$, $\{(1)\} < V_4 < S_4$ and $\{(1)\} < V_4 < A_4 < S_4$ are normal series of S_4. Here $V_4 = \{(1), (12)(34), (13)(24), (14)(23)\}$.
4. $\{(1)\} < \{(1), (12)(34)\} < V_4 < A_4 < S_4$ is a subnormal series of S_4 which is not a normal series.

A subnormal [normal] series $\{K_j\}$ is a **refinement** of a subnormal [normal] series $\{H_i\}$ of a group G if $\{H_i\} \subseteq \{K_j\}$.

Example 3.1.2. The series $\{0\} < 72\mathbb{Z} < 9\mathbb{Z} < 3\mathbb{Z} < \mathbb{Z}$ is a refinement of the series $\{0\} < 9\mathbb{Z} < \mathbb{Z}$.

Two subnormal [normal] series $\{H_i\}$ and $\{K_j\}$ of the same group G are **isomorphic** if there is a one-to-one correspondence between the collections of factor groups $\{H_{i+1}/H_i\}$ and $\{K_{j+1}/K_j\}$ such that corresponding factor groups are isomorphic. Clearly, two isomorphic subnormal [normal] series must have the same number of groups.

Example 3.1.3. The two series of $\mathbb{Z}_{15}$, $\{0\} < \langle 5\rangle < \mathbb{Z}_{15}$ and $\{0\} < \langle 3\rangle < \mathbb{Z}_{15}$ are isomorphic.

The following theorem is fundamental to the theory of series.

Theorem 3.1.1. [Schreier] *Two subnormal [normal] series of a group G have isomorphic refinements.*

Example 3.1.4. Find isomorphic refinements of the normal series

$$\{0\} < 8\mathbb{Z} < 4\mathbb{Z} < \mathbb{Z} \text{ and } \{0\} < 9\mathbb{Z} < \mathbb{Z}.$$

Consider the refinement

$$\{0\} < 72\mathbb{Z} < 8\mathbb{Z} < 4\mathbb{Z} < \mathbb{Z}$$

of $\{0\} < 8\mathbb{Z} < 4\mathbb{Z} < \mathbb{Z}$ and the refinement

$$\{0\} < 72\mathbb{Z} < 18\mathbb{Z} < 9\mathbb{Z} < \mathbb{Z}$$

of $\{0\} < 9\mathbb{Z} < \mathbb{Z}$. In both cases the refinements have four factor groups isomorphic to $\mathbb{Z}_4$, $\mathbb{Z}_2$, $\mathbb{Z}_9$, and $72\mathbb{Z}$ or $\mathbb{Z}$. The *order* in which the factor groups occurs is different to be sure.

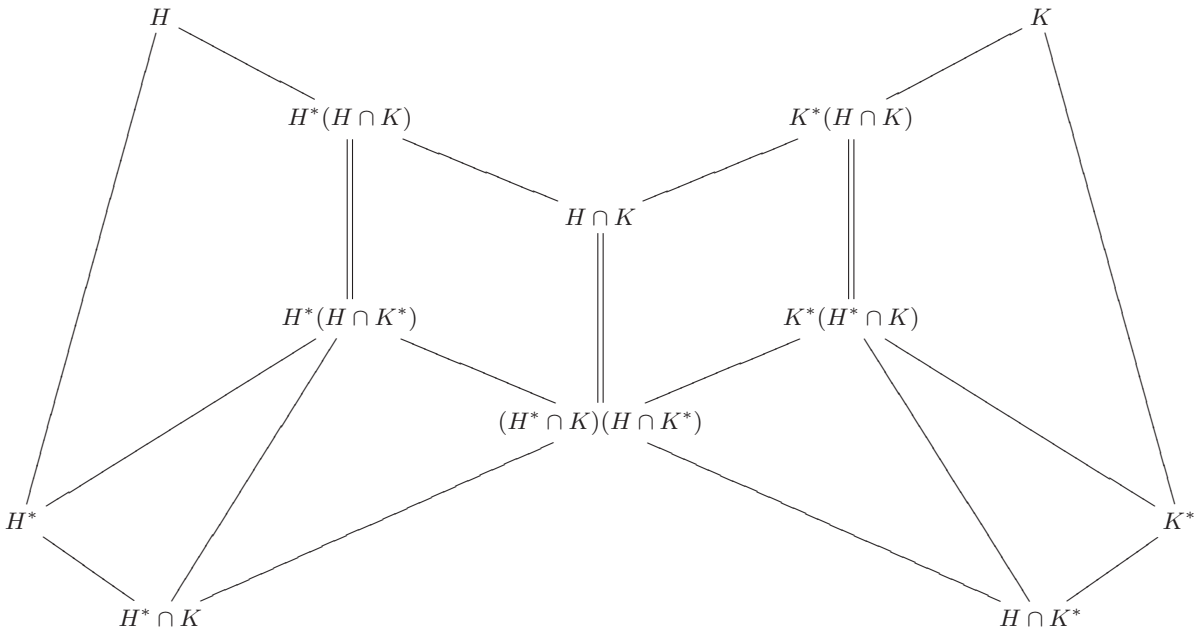
Recall the following fact.

Theorem 3.1.2. *If N is a normal subgroup of G, and if H is any subgroup of G, then $HN = NH$ is a subgroup of G. Furthermore, if $H \triangleleft G$, then $HN \triangleleft G$.*

To prove Schreier's theorem, we shall need the following lemma developed by Zassenhaus. This lemma is also called the **butterfly lemma** since the diagram which accompanies the lemma has a butterfly shape.

Lemma 3.1.3. [Zassenhaus] *Let H and K be subgroups of a group G and let H^* and K^* be normal subgroups of H and K respectively. Then*

1. *$H^*(H \cap K^*)$ is a normal subgroup of $H^*(H \cap K)$.*
2. *$K^*(H^* \cap K)$ is a normal subgroup of $K^*(H \cap K)$.*
3. *$H^*(H \cap K)/H^*(H \cap K^*) \cong K^*(H \cap K)/K^*(H^* \cap K) \cong (H \cap K)/[(H^* \cap K)(H \cap K^*)]$.*



Proof. We first note that

$$H^*(H \cap K), H^*(H \cap K^*), K^*(H \cap K) \text{ and } K^*(H^* \cap K)$$

are groups. It is easy to show that $H^* \cap K$ are $H \cap K^*$ are normal subgroups of $H \cap K$. Apply Theorem 3.1.2 to $H^* \cap K$ and $H \cap K^*$ as normal subgroups of $H \cap K$, we have $L = (H^* \cap K)(H \cap K^*)$ is a normal subgroup of $H \cap K$. Thus we have the lattice of subgroups shown above.

Let $\phi : H^*(H \cap K) \to (H \cap K)/L$ be defined as follows. For $h \in H^*$ and $x \in H \cap K$, let $\phi(hx) = xL$. We show ϕ is well defined and a homomorphism. Let $h_1, h_2 \in H^*$ and $x_1, x_2 \in H \cap K$.

If $h_1x_1 = h_2x_2$, then $h_2^{-1}h_1 = x_2x_1^{-1} \in H^* \cap (H \cap K) = H^* \cap K \subseteq L$, so $x_1L = x_2L$. Thus ϕ is well defined. Since H^* is normal in H, there is h_3 in H^* such that $x_1h_2 = h_3x_1$. Then

$$\begin{aligned}\phi((h_1x_1)(h_2x_2)) &= \phi((h_1h_3)(x_1x_2)) = (x_1x_2)L \\ &= (x_1L)(x_2L) = \phi(h_1x_1)\phi(h_2x_2)\end{aligned}$$

Thus, ϕ is a homomorphism.

Obviously ϕ is onto $(H \cap K)/L$. Finally if $h \in H^*$ and $x \in H \cap K$, then $\phi(hx) = xL = L$ if and only if $x \in L$, or if and only if $hx \in H^*L = H^*(H^* \cap K)(H \cap K^*) = H^*(H \cap K^*)$. Hence, $\ker\phi = H^*(H \cap K^*)$. Another similar result follows by symmetry. □

Proof of Schreier's theorem. Let G be a group and let

$$\{e\} = H_0 < H_1 < H_2 < \cdots < H_n = G$$

and

$$\{e\} = K_0 < K_1 < K_2 < \cdots < K_m = G$$

be two subnormal series for G. For i where $0 \leq i \leq n-1$, we form the chain of (not necessarily distinct) groups

$$H_i = H_i(H_{i+1} \cap K_0) \leq H_i(H_{i+1} \cap K_1) \leq \cdots \leq H_i(H_{i+1} \cap K_m) = H_{i+1}.$$

We refine the first subnormal series by inserting the above chain between H_i and H_{i+1}. In a symmetric fashion, for $0 \leq j \leq m-1$, we insert the chain

$$K_j = K_j(K_{j+1} \cap H_0) \leq K_j(K_{j+1} \cap H_1) \leq \cdots \leq K_j(K_{j+1} \cap H_n) = K_{j+1}$$

between K_j and K_{j+1}. Thus we get two refinement having mn terms. By Zassenhaus's Lemma, we have

$$H_i(H_{i+1} \cap K_{j+1})/H_i(H_{i+1} \cap K_j) \cong K_j(K_{j+1} \cap H_{i+1})/K_j(K_{j+1} \cap H_i)$$

for $0 \leq i \leq n-1$ and $0 \leq j \leq m-1$. Hence, this two refinements are isomorphic.

For normal series, where all H_i and K_j are normal in G, we merely observe that all the groups $H_i(H_{i+1} \cap K_j)$ and $K_j(K_{j+1} \cap H_i)$ are normal in G, so the same proof applies. □

A normal subgroup $M(\neq G)$ is called a **maximal normal subgroup** of G if there exists no normal subgroup N, other than G or M, such that $M \triangleleft N \triangleleft G$. Recall that a group G is **simple** if G and $\{e\}$ are the only normal subgroups of G. For example, $\mathbb{Z}_p$, p a prime, and A_n, $n \neq 4$, are simple. We also have an obvious fact.

Theorem 3.1.4. *G is a simple abelian group if and only if G is cyclic of prime order.*

The next criterion follows directly from the third isomorphism theorem (Theorem 1.4.4).

Theorem 3.1.5. *M is a maximal normal subgroup of a group G if and only if G/M is simple.*

A subnormal series $\{H_i\}$ of a group G is a **composition series** if all the factor groups H_{i+1}/H_i are simple. A normal series $\{H_i\}$ of G is a **principal** or **chief series** if all the factor groups H_{i+1}/H_i are simple.

Observe that by Theorem 3.1.5 H_{i+1}/H_i is simple if and only if H_i is a maximal normal subgroup of H_{i+1}. Thus for a composition series, each H_i must be a maximal normal subgroup of H_{i+1}. *To form a composition series of a group G, we just look for a maximal normal subgroup H_{n-1} of G, then for a maximal normal subgroup of H_{n-1}, and so on. If this process terminates in finite number of steps, we have a composition series.* Hence, we have first shown:

Theorem 3.1.6. *If G is a finite group, then G has a composition series.*

Note that by Theorem 3.1.5 a composition series cannot have any further refinement. *To form a principal series, we have to hunt for a maximal normal subgroup H_{n-1} of G, then for a maximal normal subgroup of H_{n-1} that is also normal in G, and so on*. The main theorem is as follows.

Theorem 3.1.7. [Jordan-Hölder] *Any two composition [principal] series of a group G are isomorphic.*

Proof. Let $\{H_i\}$ and $\{K_i\}$ be two composition [principal] series of G. By Schreier's theorem, they have isomorphic refinements. But since all factor groups are already simple, Theorem 3.1.5 shows that neither series has any further refinement. Hence, $\{H_i\}$ and $\{K_i\}$ must already be isomorphic. □

Examples 3.1.5 (Examples of composition series).

1. If G is simple, then $\{e\} \triangleleft G$ is the only normal series of G. It is a composition series for G and its associated factor is $G = G/\{e\}$.
2. If $n \neq 4$, then $\{(1)\} < A_n < S_n$ is a composition series of S_n.
3. $\mathbb{Z}$ has many normal series. For example, let $m_1, \ldots, m_n$ be positive integers. Then

$$\mathbb{Z} > m_1\mathbb{Z} > m_1m_2\mathbb{Z} > \cdots > m_1m_2 \ldots m_n\mathbb{Z} > \{0\}$$

   is a normal series for $\mathbb{Z}$ whose associated factors are $\mathbb{Z}_{m_1}$, $\mathbb{Z}_{m_2}$, $\ldots$, $\mathbb{Z}_{m_n}$, $\mathbb{Z}$. Note that since any nontrivial subgroup of $\mathbb{Z}$ is isomorphic to $\mathbb{Z}$, any normal series for $\mathbb{Z}$ must have one associated factor isomorphic to $\mathbb{Z}$. Hence, $\mathbb{Z}$ *has no composition series*.
4. Let p be prime and $G = \mathbb{Z}_p \times \mathbb{Z}_p$. If $(x, y) \neq (0, 0)$ in G, then $\langle (x, y)\rangle \cong \mathbb{Z}_p$ and $\{(0, 0)\} < \langle (x, y)\rangle < G$ is a composition series for G. The composition factors are $G/\langle (x, y)\rangle \cong \mathbb{Z}_p$ and $\langle (x, y)\rangle/(0, 0) \cong \mathbb{Z}_p$, i.e., $\mathbb{Z}_p$ with multiplicity 2. Note that G has $(p^2 - 1)/(p - 1) = p + 1$ subgroups of order p, so G has $p + 1$ distinct composition series. But in all cases they have the same composition factors: $\mathbb{Z}_p$ with multiplicity 2.
5. Let p and q be prime and $G = \mathbb{Z}_p \times \mathbb{Z}_q = \langle a\rangle \times \langle b\rangle$. Then the only proper subgroup of G are $\langle a\rangle = \mathbb{Z}_p$ and $\langle b\rangle = \mathbb{Z}_q$. Thus G has two composition series

$$\{e\} < \langle a\rangle < G \text{ and } \{e\} < \langle b\rangle < G$$

   In both cases, the associated composition factors are $\mathbb{Z}_p$ and $\mathbb{Z}_q$ both with multiplicity one.
6. Consider $\mathbb{Z}_{p^3}$, $\mathbb{Z}_{p^2} \times \mathbb{Z}_p$ and $\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$. In any composition series for these groups the same composition factors, namely $\mathbb{Z}_p$ with multiplicity 3, occur.

Exercises 3.1.

1. Suppose G has precisely two subgroups. Show that G has prime order.
2. A proper subgroup M of G is **maximal** if whenever $M \subseteq H \subseteq G$, we have $H = M$ or $H = G$. Suppose G is finite and has only one maximal subgroup. Show that $|G|$ is a power of prime.
3. Let $G = \mathbb{Z}_{36}$. Consider two normal series $\{0\} < \langle 12\rangle < \langle 3\rangle < \mathbb{Z}_{36}$ and $\{0\} < \langle 18\rangle < \mathbb{Z}_{36}$. Find two isomorphic chains and exhibit the isomorphic factor groups as described in the proof of Schreier's Theorem.
4. Find a composition series for the dihedral group $D_4 = \{\sigma, \rho : \sigma^4 = \rho^2 = e \text{ and } \rho\sigma\rho^{-1} = \sigma^{-1}\}$ and for the quaternion group $Q = \{\pm 1, \pm i, \pm j, \pm k\}$. Determine the composition factor in each case.
5. Prove that if G has a composition [resp. principal] series and if N is a proper normal subgroup of G, then there exists a composition [resp. principal] series containing N. Hence, show that N and G/N have composition [principal] series.
6. Show that if $H_0 = \{e\} < H_1 < H_2 < \cdots < H_n = G$ is a subnormal [normal] series of G, and if H_{i+1}/H_i is of finite order s_{i+1}, then G is of finite order $s_1s_2 \ldots s_n$.
7. Show that an infinite abelian group can have no composition series.

3.2 Solvable Groups

Let G be a group. For $g, h \in G$, $[g,h] = ghg^{-1}h^{-1}$ is called a **commutator of** G. The **derived subgroup** of G, denoted by G', is the group generated by all commutators of elements of G, i.e.,

$$G' = \langle ghg^{-1}h^{-1} : g, h \in G \rangle.$$

The n**-th derived subgroup** of G, denoted by $G^{(n)}$ is defined inductively by $G^{(0)} = G$ and $G^{(n)} = (G^{(n-1)})'$ for all $n \geq 1$.

Theorem 3.2.1. *Let G be a group.*

1. *If N is a subgroup, then (N is normal and G/N is abelian) if and only if $G' \subseteq N$.*
2. *G' is a normal subgroup of G and G/G' is abelian.*
3. *Every homomorphism $\theta : G \to A$, where A is an abelian group, factors through G/G'. More precisely, there is a map $\bar{\theta} : G/G' \to A$ such that $\theta = \bar{\theta} \circ \pi$, where $\pi : G \to G/G'$ is the canonical projection.*

Proof. (1) Assume that N is normal and G/N is abelian. Let $x, y \in G$. Then $xyN = yxN$, so $xyx^{-1}y^{-1} \in N$. Thus $G' \subseteq N$. Conversely, suppose that $G' \subseteq N$. Let $x, y \in G$ and $n \in N$. Then $xnx^{-1}n^{-1} \in G' \subseteq N$ which implies that $xnx^{-1} \in Nn = N$. Hence, $N \lhd G$. Since $(xy)(yx)^{-1} = xyx^{-1}y^{-1} \in G' \subseteq N$, $xyN = yxN$, so G/N is abelian.
(2) follows from 1 by taking $N = G'$.
(3) Define $\bar{\theta}(xG') = \theta(x)$ for all $x \in G$. Clearly, $\theta = \bar{\theta} \circ \pi$ and is a homomorphism. Since $\theta(G') = \{e\}$, $\bar{\theta}$ is well defined. □

Remark. The quotient G/G' is the largest abelian homomorphic image of G.

The **derived series** of a group G is the sequence of groups

$$G = G^{(0)} \geq G^{(1)} \geq G^{(2)} \geq \cdots \geq G^{(n)} \geq \ldots.$$

A group G is said to be **solvable (of derived length $\leq n$)** if $G^{(n)} = \{e\}$ for some n. A solvable group arises from the study of the Galois group of a polynomial in order to obtain a criterion to determine if it is solvable by radicals. We shall see this in Section 5.7.

A subgroup H of a group G which is invariant under all automorphisms, that is, $\varphi(H) \leq H$ for all $\varphi \in \operatorname{Aut} G$, is called a **characteristic subgroup of** G. Using the inner automorphisms $\varphi_a(x) = axa^{-1}$ for all $a \in G$, we deduce that every characteristic subgroup is normal in G.

Lemma 3.2.2. *Let $\varphi : G \to H$ be a surjective homomorphism. Then $\varphi(G^{(i)}) = H^{(i)}$ for every $i \geq 0$. Also, $G^{(i)}$ is a characteristic subgroup for all i, and is thus normal in G.*

Proof. We have $\varphi([x,y]) = [\varphi(x), \varphi(y)]$, and since φ is onto, we see that φ maps the set of commutators in G onto those in H. It follows that $\varphi(G') = H'$, and repeated application of this argument yields that $\varphi(G^{(i)}) = H^{(i)}$, as required. That the terms of the derived series of G are characteristic follows from the first part of the lemma when we take $H = G$ and $\varphi \in \operatorname{Aut} G$. □

Theorem 3.2.3. *Let G be a group. Then G is solvable if and only if G has a subnormal series with abelian factors.*

Proof. If G is solvable, $G = G^{(0)} > G^{(1)} > \cdots > G^{(n)} = \{e\}$ is a subnormal series with abelian factors. Conversely, suppose $G = G_0 > G_1 > \cdots > G_m = \{e\}$ is a subnormal series for G with abelian factors. Since G_i/G_{i+1} is abelian, $G_{i+1} \geq G_i'$. We claim $G_i \geq G^{(i)}$ for $i = 0, 1, \ldots, m$ by induction on i. For $i = 0$, $G_0 = G = G^{(0)}$. Assume $G_i \geq G^{(i)}$. Then $G_{i+1} \geq G_i' \geq (G^{(i)})' = G^{(i+1)}$, which completes the induction. Hence, $\{e\} = G_m \geq G^{(m)}$, so $G^{(m)} = 1$ and G is solvable. □

Remark. From Lemma 3.2.2, we know that $G^{(i)} \triangleleft G$ for all i. Then the above derived series

$$G = G^{(0)} > G^{(1)} > \cdots > G^{(n)} = \{e\}$$

is indeed a normal series with abelian factors for G. Also, if G is solvable, its **derived length**, $\mathrm{dl}(G)$, is the smallest positive integer n such that $G^{(n)} = \{e\}$.

Examples 3.2.1 (Examples of solvable groups)**.** 1. An abelian group G is solvable of derived length 1 because $G' = \{e\}$. In addition, the groups with derived length 1 are exactly the abelian groups. Hence, a group G is abelian if and only if G is solvable of derived length 1.

2. Let D_n be the dihedral group of order $2n$, i.e.,

$$D_n = \{\sigma, \rho : \sigma^n = \rho^2 = e \text{ and } \rho\sigma\rho^{-1} = \sigma^{-1}\}.$$

Here, σ is the $2\pi/n$ rotation and ρ is the reflection of the regular n-gon. For example, $D_1 = \mathbb{Z}_2$, $D_2 = \mathbb{Z}_2 \times \mathbb{Z}_2$ and $D_3 = S_3$. Then $D_n' = \langle\sigma^2\rangle$, an abelian group. Thus, $D_n^{(2)} = \{e\}$. For $n = 1$ or 2, D_n is abelian and hence has derived length one. For $n \geq 3$, D_n is solvable of derived length two.

Proof. Observe that $D_n = \{e, \sigma, \sigma^2, \ldots, \sigma^{n-1}, \rho, \rho\sigma, \rho\sigma^2, \ldots, \rho\sigma^{n-1}\}$. For $x, y \in D_n$, we distinguish four cases

$$xyx^{-1}y^{-1} = \begin{cases} \sigma^k\sigma^l\sigma^{-k}\sigma^{-l} = e \\ (\rho\sigma^k)(\rho\sigma^l)(\sigma^{-k}\rho^{-1})(\sigma^{-l}\rho^{-1}) = \rho\sigma^k\sigma^k\sigma^{-l}\sigma^{-l}\rho^{-1} = \sigma^{-2k}\sigma^{2l} \\ (\rho\sigma^k)\sigma^l(\sigma^{-k}\rho^{-1})\sigma^{-l} = \sigma^{-l}\sigma^{-l} = \sigma^{-2l} \\ \sigma^k(\rho\sigma^l)\sigma^{-k}(\sigma^{-l}\rho^{-1}) = \sigma^k\sigma^k = \sigma^{2k}. \end{cases}$$

This implies that $D_n' \subseteq \langle\sigma^2\rangle$. On the other hand, we have $\sigma^2 = \rho\sigma^{-1}\rho^{-1}\sigma$. Thus, $D_n' = \langle\sigma^2\rangle$. □

3. The groups $S_1 = \{(1)\}$ and $S_2 = \mathbb{Z}_2$ are abelian groups. The group $S_3 = D_3$ is solvable of derived length two. Since $S_4' = A_4$, $A_4' = V_4$ and $V_4' = \{(1)\} = S_4^{(3)}$, we can conclude that the group S_4 is solvable of derived length 3. For $n \geq 5$, $S_n' = A_n$ and $A_n' = S_n^{(2)} = A_n$ since A_n is simple and non-abelian. Therefore $S_n \geq A_n \geq A_n \geq \ldots$ is the derived series of S_n and S_n is not solvable for $n \geq 5$. These facts are important in Galois theory (Section 5.7) and relate to the famous formula for the solution of quadratic, cubic and quartic equations (by using square roots, cube roots, etc.), and the historic proof by Abel in 1824 that there are no such formula for the quintic equation.

Proof. It is easy to see that any group of order two in A_4 are not normal. Since A_4 has more than one Sylow 3-subgroup, any subgroups of A_4 of order three are not normal. Moreover, A_4 has no subgroup of order six (see Exercises 1.5). Hence, the normal subgroups of A_4 are A_4, V_4 and $\{(1)\}$. Note that S_4' is a subgroup of A_4. Moreover, it is normal in A_4. Since S_4 and S_4/V_4 are not abelian, S_4' must be A_4. Since A_4 is not abelian and A_4/V_4 is abelian, we have $A_4' = V_4$. Hence, $S_4 \triangleright A_4 \triangleright V_4 \triangleright \{(1)\}$ is the derived series of S_4. Next, let $n \geq 5$ and $K = S_n' \triangleleft S_n$. Then $K \cap A_n \triangleleft S_n$, so $K \cap A_n \triangleleft A_n$. Since A_n is simple, $K \cap A_n = \{(1)\}$ or $K \cap A_n = A_n$. But $K \subseteq A_n$ and $K \neq \{(1)\}$ (since S_n is non-abelian), we get $K = A_n$. Hence, $S_n' = A_n$. □

The following theorem is often useful to decide if a group is solvable.

Theorem 3.2.4. *1. If G is solvable and H is a subgroup of G, then H is solvable.*
2. If G is solvable and N is a normal subgroup G, then G/N is solvable.
3. A homomorphic image of a solvable group is solvable.
4. If $N \triangleleft G$ and N and G/N are solvable, then G is solvable and $\mathrm{dl}(G) \leq \mathrm{dl}(N) + \mathrm{dl}(G/N)$.
5. If G and H are solvable, then $G \times H$ is solvable.

Proof. (1) Since $H^{(i)} \leq G^{(i)}$ for all i, $H^{(n)} = \{e\}$ if $G^{(n)} = \{e\}$.
(2) The application of Lemma 3.2.2 to the canonical homomorphism $\pi : G \to G/N$ yields that $(G/N)^{(i)} = \pi(G^{(i)})$ for all i, and hence if $G^{(n)} = \{e\}$, we have $(G/N)^{(n)} = \{N\}$.
(3) follows from (2).
(4) Let $\mathrm{dl}(N) = n$ and $\mathrm{dl}(G/N) = m$. Since the canonical homomorphism $\varphi : G \to G/N$ maps $G^{(m)}$ to $(G/N)^{(m)} = \{N\}$, we see that $G^{(m)} \subseteq N$. Thus $G^{(m+n)} = (G^{(m)})^{(n)} \subseteq N^{(n)} = \{e\}$, and hence G is solvable.
(5) follows from (4). □

Some additional conditions under which finite groups are solvable are as follows.

Theorem 3.2.5. *Let G be a finite group.*

1. [Burnside] *If $|G| = p^a q^b$ for some primes p and q, then G is solvable.*
2. [Philip Hall] *If for every prime p dividing $|G|$ we factor the order of G as $|G| = p^a m$ where $(p, m) = 1$, and G has a subgroup of order m, then G is solvable, i.e., if for all primes p, G has a subgroup whose index equals the order of a Sylow p-subgroup, then G is solvable—such subgroups are called* **Sylow p-complements**.
3. [Feit-Thompson] *If G is odd, then G is solvable.*
4. [Thompson] *If for every pair of elements $x, y \in G$, $\langle x, y\rangle$ is a solvable groups, then G is solvable.*

Burnside's and Philip Hall's Theorems were proved by using Character Theory. The proof of the Feit-Thompson Theorem takes 255 pages of hard mathematics (Solvability of groups of odd order, *Pacific Journal of Mathematics*, 13 (1963), pp. 775–1029). Thompson's Theorem was first proved as a consequence of 475-page paper (that in turn relies ultimately on the Feit-Thompson Theorem).

Exercises 3.2.

1. (a) Give an example of a normal subgroup of G which is not characteristic.
   (b) Prove that $Z(G)$ is a characteristic subgroup of G.
   (c) If H is a characteristic subgroup of N and $N \triangleleft G$, show that $H \triangleleft G$.
2. Show that if G is a solvable simple group, then G is abelian.
3. Let $\{e\} = H_0 < H_1 < H_2 < \cdots < H_{n-1} < H_n = G$ be a composition for G. Prove that G is solvable if and only if the composition factors H_{i+1}/H_i all have prime order. Deduce that if G is solvable with a composition series, then G is finite.
4. Find a composition series of $S_3 \times S_3$. Is $S_3 \times S_3$ solvable?
5. Show that a group of order 1995 is solvable.
6. Let $p < q < r$ be primes and let G_1 be a group of order pq and let G_2 be a group of order pqr. Prove that both of them are solvable. [*Hint*. G_1 has a unique subgroup of order q.]
7. Let G be a group of order $495 = 3^2 \cdot 5 \cdot 11$.
   (a) Prove that a Sylow 5-subgroup or a Sylow 11-subgroup of G is normal in G.
   (b) Let P be a Sylow 5-subgroup and Q a Sylow 11-subgroup of G. Prove that PQ is normal in G.
   (c) Prove that G is solvable.
8. Prove (without using the Feit-Thompson Theorem) that the following statements are equivalent:
   (i) every group of odd order is solvable
   (ii) the only simple groups of odd order are those of prime order.

3.3 Nilpotent Groups

In this section, we shall introduce a class of groups whose structure, next to those of abelian groups, is most amenable to analysis. We begin by generalizing the notion of a commutator.

If A and B are subsets of G, $[A, B]$ is the subgroup of G generated by all commutators $[a, b] = aba^{-1}b^{-1}$ where $a \in A$ and $b \in B$, that is,

$$[A, B] = \langle [a, b] : a \in A \text{ and } b \in B \rangle.$$

Note that $[A, B] = [B, A]$.

Example 3.3.1. $G' = [G, G]$, $G^{(2)} = [G', G']$, $\ldots$, $G^{(n+1)} = [G^{(n)}, G^{(n)}]$.

The **lower central series of a group** G is defined inductively by $\Gamma_1(G) = G$ and $\Gamma_{n+1}(G) = [G, \Gamma_n(G)]$ for all $n \geq 1$, so we get

$$G = \Gamma_1(G) \geq \Gamma_2(G) \geq \ldots$$

and $\Gamma_n(G)$ is called the **n-th term of the lower central series of** G.

A group G is said to be **nilpotent** of class $\leq n$ if $\Gamma_{n+1} = \{e\}$.

Remarks. 1. Since $\Gamma_2(G) = [G, G] = G'$, G is abelian if and only if G is nilpotent of class ≤ 1.

2. Note that the derived series commences

$$G = G^{(0)} \geq G^{(1)} \geq \ldots$$

while the lower central series commences

$$G = \Gamma_1(G) \geq \Gamma_2(G) \geq \ldots.$$

Note however that G is abelian if and only if $\{e\} = [G, G] = G' = \Gamma_2(G)$, so

$$G \text{ is abelian} \Leftrightarrow G \text{ is solvable of length} \leq 1 \Leftrightarrow G \text{ is nilpotent of class} \leq 1.$$

Examples 3.3.2 (Examples of nilpotent groups). 1. S_3 has the derived series $S_3 > A_3 > \{(1)\}$ and has the lower central series $S_3 > A_3 \geq A_3 \geq \ldots$, so S_3 is solvable (of length 2) but not nilpotent.

2. S_4 has the derived series $S_4 > A_4 > V_4 > \{(1)\}$ and has the lower central series $S_4 > A_4 \geq A_4 \geq \ldots$, so S_4 is solvable (of length 3) but not nilpotent.

3. $D_n = \langle \rho, \tau : \rho^n = \tau^2 = e \text{ and } \tau\rho\tau^{-1} = \rho^{-1} \rangle$ has the derived series $D_n > \langle \rho^2 \rangle > \{e\}$ and has a lower central series $D_n \geq \langle \rho^2 \rangle \geq \langle \rho^4 \rangle \geq \langle \rho^8 \rangle \geq \ldots$. Hence, D_n is solvable (of length 2) unless D_1 or D_2 which is abelian. But D_n is nilpotent if and only if $\rho^{2^r} = e$ for some r if and only if n is a power of 2.

Theorem 3.3.1. *Let G be a group. Then $\Gamma_{n+1}(G) \geq G^{(n)}$ for all $n \geq 0$. Hence, a nilpotent group is solvable. Therefore, S_n is not nilpotent for all $n \geq 5$.*

Proof. We shall use induction on n. For $n = 0$, $\Gamma_1(G) = G = G^{(0)}$. For the inductive step, we suppose $\Gamma_{n+1}(G) \geq G^{(n)}$. Thus

$$\Gamma_{n+2}(G) = [G, \Gamma_{n+1}(G)] \geq [G^{(n)}, G^{(n)}] = G^{(n+1)}.$$

Finally, assume that G is nilpotent. Then $\Gamma_{n+1}(G) = \{e\}$ for some n, so $G^{(n)} = \{e\}$. Hence, G is solvable. □

Remark. In fact, we have $\Gamma_1(G) \geq G^{(0)}$, $\Gamma_2(G) \geq G^{(1)}$, $\Gamma_4(G) \geq G^{(2)}$, $\Gamma_8(G) \geq G^{(3)}$, $\ldots$, $\Gamma_{2^n}(G) \geq G^{(n)}$, $\ldots$ but this is more difficult to prove.

Recall that if N is a normal subgroup of G, then $H \leftrightarrow H/N$ gives a 1-1 correspondence between subgroups of G containing N and subgroups of G/N. Moreover, this correspondence carries normal subgroups to normal subgroups.

Now let $Z(G)$ denote the center of a group G. Then $Z(G)$ is a normal subgroup of G and $Z(G/Z(G))$ is a normal subgroup of $G/Z(G)$. Hence,

$$Z(G/Z(G)) = Z_2(G)/Z(G)$$

where $Z_2(G)$ is a normal subgroup of G containing $Z(G)$. We generalize this construction to make the following definition.

The **upper central series of a group** G is defined inductively by $Z_0(G) = \{e\}$ and $Z_{n+1}(G)/Z_n(G) = Z(G/Z_n(G))$ for all $n \geq 1$, so we get

$$\{e\} = Z_0(G) \leq Z_1(G) \leq Z_2(G) \leq \dots$$

and $Z_n(G)$ is called the **n-th term of the upper series of** G.

Remarks. 1. $Z_1(G)$ is the center of G and $Z_{i+1}(G)/Z_i(G)$ is the center of $G/Z_i(G)$.

2. $Z_{i+1}(G)/Z_i(G) = Z(G/Z_i(G))$ is equivalent to $Z_{i+1}(G) = \{g \in G : [G, g] \leq Z_i(G)\}$ because

$$\begin{aligned} Z_{i+1}(G)/Z_i(G) =& Z(G/Z_i(G)) \\ &\iff \forall g \in G, [g \in Z_{i+1}(G) \Leftrightarrow \forall x \in G, gxZ_i(G) = xgZ_i(G)] \\ &\iff \forall g \in G, [g \in Z_{i+1}(G) \Leftrightarrow \forall x \in G, xgx^{-1}g^{-1} \in Z_i(G)] \\ &\iff \forall g \in G, [g \in Z_{i+1}(G) \Leftrightarrow [G, g] \subseteq Z_i(G)] \\ &\iff Z_{i+1}(G) = \{g \in G : [G, g] \leq Z_i(G)\}. \end{aligned}$$

3. We can show by induction that $Z_i(G)$ is a characteristic subgroup of G for all $i \in \mathbb{N}$.

A subnormal series $G = G_1 \geq G_2 \geq \dots$ is called a **central series** for G if $[G, G_i] \leq G_{i+1}$ for all i.

Remarks. 1. Since $[G, \Gamma_i(G)] = \Gamma_{i+1}(G)$, the lower central series is a central series for G.

2. Note that the condition $Z_{i+1}(G)/Z_i(G) = Z(G/Z_i(G))$ implies the inclusion $[G, Z_{i+1}(G)] \leq Z_i(G)$. Thus, if $Z_n(G) = G$ for some n, then the upper central series (in reverse order) is a central series for G:

$$G = Z_n(G) \geq Z_{n-1}(G) \geq \dots \geq Z_1(G) \geq Z_0(G) = \{e\}.$$

Now, we wish to collect equivalence definitions of a nilpotent group in terms of lower central series, upper central series and central series.

Theorem 3.3.2. *Let G be a group.*

1. *If $G = G_1 \geq G_2 \geq G_3 \geq \dots$ is a central series for G, then $G_n \geq \Gamma_n(G)$ for all n.*
2. *G has a central series $G = G_1 > G_2 > \dots > G_{n+1} = \{e\}$ if and only if G is nilpotent of class $\leq n$.*

Proof. (1) We shall use induction on n. For $n = 1$, $G_1 = G = \Gamma_1(G)$. For the inductive step, we suppose $G_n \geq \Gamma_n(G)$. Then

$$G_{n+1} \geq [G, G_n] \geq [G, \Gamma_n(G)] = \Gamma_{n+1}(G).$$

(2) If $G = G_1 \leq G_2 \leq \dots \leq G_{n+1} = \{e\}$ is a central series for G, then $\{e\} = G_{n+1} \geq \Gamma_{n+1}(G)$, so G is nilpotent of class $\leq n$. Conversely, if G is nilpotent of class $\leq n$, then $G = \Gamma_1(G) \geq \dots \geq \Gamma_{n+1}(G) = \{e\}$ is a central series of the required length. □

Theorem 3.3.3. *Let G be a group.*
1. *Suppose $G = G_1 \geq G_2 \geq \ldots \geq G_{n+1} = \{e\}$ is a central series for G. Then $Z_k(G) \geq G_{n-k+1}$ for all $k \in \{0, 1, \ldots, n\}$.*
2. *If $Z_n(G) = G$, then $G = Z_n(G) \geq Z_{n-1}(G) \geq \ldots \geq Z_1(G) \geq Z_0(G) = \{e\}$ is a central series for G.*
3. *G is nilpotent of class $\leq n$ if and only if $Z_n(G) = G$.*

Proof. (1) We shall show that $Z_k(G) \geq G_{n-k+1}$ by induction on k. For $k = 0$, $Z_0(G) = \{e\} = G_{n+1}$. Suppose $Z_k(G) \geq G_{n-k+1}$. Let $g \in G_{n-(k+1)+1} = G_{n-k}$, then $[G, g] \leq G_{n-k+1} \leq Z_k(G)$, so $g \in Z_{k+1}(G)$. Hence, $Z_{k+1}(G) \geq G_{n-(k+1)+1}$.
(2) Since $Z_{i+1}(G)/Z_i(G) = Z(G/Z_i(G))$, $[G, Z_{i+1}(G)] \leq Z_i(G)$. Hence, the given series is a central series.
(3) follows from (1) and (2) using Theorem 3.3.2. □

Suppose that G is nilpotent of class $\leq n$ and that $G = G_1 \geq G_2 \geq \ldots \geq G_{n+1} = \{e\}$ is *any* central series for G. Theorems 3.3.2 and 3.3.3 show that we have the following inclusions

$$\begin{array}{ccccccccc}
G = & Z_n(G) & \geq & Z_{n-1}(G) & \geq \cdots \geq & Z_k(G) & \geq \cdots \geq & Z_0(G) & = \{e\} \\
 & \bigcup| & & \bigcup| & & \bigcup| & & \bigcup| & \\
G = & G_1 & \geq & G_2 & \geq \cdots \geq & G_{n-k+1} & \geq \cdots \geq & G_{n+1} & = \{e\} \\
 & \bigcup| & & \bigcup| & & \bigcup| & & \bigcup| & \\
G = & \Gamma_1(G) & \geq & \Gamma_2(G) & \geq \cdots \geq & \Gamma_{n-k+1}(G) & \geq \cdots \geq & \Gamma_{n+1}(G) & = \{e\}
\end{array}$$

In other words, of all central series for G, the upper central series has the largest groups and the lower central series has the smallest groups. We can restate some of the conclusions of Theorems 3.3.2 and 3.3.3 as follows.

Theorem 3.3.4. *Let G be a group. Then the following statements are equivalent.*
(i) *G is nilpotent of class $\leq n$.*
(ii) *$\Gamma_{n+1}(G) = \{e\}$.*
(iii) *G has a central series $G = G_1 \geq G_2 \geq \ldots \geq G_{n+1} = \{e\}$.*
(iv) *$Z_n(G) = G$.*

Next, we shall see that a finite nilpotent group behaves like a finite abelian group. We show that it is a direct product of its Sylow p-subgroups. We recall Theorem 1.6.2.

Theorem 3.3.5. *Let p be a prime. If $G \neq \{e\}$ is a finite p-group, then $Z(G) \neq \{e\}$.*

We can thus prove another important fact.

Theorem 3.3.6. *Let G be a finite p-group. Then G is nilpotent, and hence G is solvable.*

Proof. Consider the upper central series $\{e\} = Z_0(G) \leq Z_1(G) \leq Z_2(G) \leq Z_3(G) \leq \ldots$. If $Z_i(G) \neq G$, then $G/Z_i(G)$ is a p-group, so $Z_{i+1}(G)/Z_i(G) = Z(G/Z_i(G)) \neq \{Z_i(G)\}$. That is, $Z_{i+1}(G) \gneqq Z_i(G)$. Since G is finite, the central series cannot increase for all i. Hence, $Z_n(G) = G$ for some n, so G is nilpotent. □

Theorem 3.3.7. *Let G be a nilpotent group and let $\{e\} < Z_1(G) < \cdots < Z_n(G) = G$ be the upper central series of G. Suppose H is a subgroup of G and define inductively $N_0(H) = H$, $N_1(H) = N(H) = \{g \in G : gHg^{-1} \subseteq H\}$, the normalizer of H and $N_{k+1} = N(N_k(H))$ for all $k \geq 0$. Then $N_n(H) = G$.*

Proof. We shall prove by induction on i that $N_i(H) \geq Z_i(G)$. For $i = 0$, $N_0(H) = H \geq \{e\} = Z_0(G)$. Suppose $N_i(H) \geq Z_i(G)$. Let $g \in Z_{i+1}(G)$. Then $[g, G] \subseteq Z_i(G)$. To show that $g \in N(N_i(H))$, let $x \in N_i(H)$. Then $gxg^{-1}x^{-1} \in Z_i(G) \leq N_i(H)$, so $gxg^{-1} \in N_i(H)x = N_i(H)$. Hence, $g \in N_{i+1}(H)$. □

From the above theorem, we can deduce the following:

Theorem 3.3.8. *Suppose G is nilpotent and H is a proper subgroup of G. Then $N(H) \gneqq H$.*

Before we discuss the main characterization theorem, we study some auxiliary results.

Theorem 3.3.9. *1. If G is nilpotent and H is a subgroup of G, then H is nilpotent.*
2. If G is nilpotent and N is a normal subgroup G, then G/N is nilpotent.
3. If G and H are nilpotent, then $G \times H$ is nilpotent.

Proof. (1) and (2) are analogous to the proofs of 3.2.4 for G is solvable.
(3) Suppose that G and H are nilpotent. Then there exist $r, s > 0$ so that $\Gamma_r(G) = \{e_G\}$ and $\Gamma_s(H) = \{e_H\}$. Thus $\Gamma_k(G \times H) = \Gamma_k(G) \times \Gamma_k(H) = \{(e_G, e_H)\}$ where $k = \max\{r, s\}$. Hence, $G \times H$ is nilpotent. □

Finally, we shall that a finite nilpotent group behaves like a finite abelian group as we have seen in Theorem 1.7.12. This theorem characterizes all finite nilpotent groups.

Theorem 3.3.10. [Finite Nilpotent Groups] *Let G be a finite group. Then the following statements are equivalent.*
(i) G is nilpotent.
(ii) All Sylow p-subgroups of G are normal in G.
(iii) G is the direct product of its Sylow p-subgroups.

Proof. (i) ⇒ (ii). Assume that G is nilpotent. Recall Theorem 1.6.10 that if P is a Sylow p-subgroup, then $N(N(P)) = N(P)$. But Theorem 3.3.8 asserts that if H is a proper subgroup of G, then $N(H) \gneqq H$. Thus we must have $N(P) = G$, that is, P is normal in G since $P \triangleleft N(P)$.
(ii) ⇒ (iii). Note that if a Sylow p-subgroup P of G is normal in G, then it is the unique Sylow p-subgroup of G. Let $p_1, p_2, \ldots, p_k$ be the distinct prime divisors of $|G|$ and let P_i be the Sylow p_i-subgroup of G. Suppose $x \in P_i$ and $y \in P_j$ where $i \neq j$. Then $xyx^{-1}y^{-1} \in P_i \cap P_j = \{e\}$, so x and y commute. It follows that $\phi : P_1 \times \cdots \times P_k \to G$ defined by $\phi(x_1, \ldots, x_k) = x_1 \ldots x_k$ is a homomorphism. It is easy to show that ϕ is a bijection. Hence, G is the direct product of its Sylow p-subgroups.
(iii) ⇒ (i). A finite p-group is nilpotent (Theorem 3.3.6) and a finite direct product of nilpotent groups is nilpotent (Theorem 3.3.9). Hence, if G is the direct product of its Sylow p-subgroups, then G is nilpotent. □

The next corollary is just a restatement of Theorem 1.7.12.

Corollary 3.3.11. *A finite abelian group is the direct product of its Sylow subgroups.*

Exercises 3.3. 1. (a) [P. Hall] Let G be a group and $x, y, z \in G$. Write $[x, y, z]$ for $[[x, y], z]$. Prove that $[x, y^{-1}, z]^y[y, z^{-1}, x]^z[z, x^{-1}, y]^x = e$.
(b) Let $X, Y, Z \subseteq G$ and assume $[X, Y, Z] = \{e\} = [Y, Z, X]$. Prove that $[Z, X, Y] = \{e\}$.
2. Prove that if $N \leq Z(G)$ and N and G/N are nilpotent, then G is nilpotent. Give an example of a group G with a normal subgroup N such that N and G/N are nilpotent but G is not nilpotent.

3. Let G be nilpotent of class 3. Show that if $v \in G'$ and $x \in G$, then $v^{-1}xv = cx$ where $c \in Z(G)$. Deduce that G' is abelian.
4. Show that if G is a nilpotent group and N is a normal subgroup of G where $N \neq \{e\}$, then $N \cap Z(G) \neq \{e\}$.
5. Prove that if M is a maximal subgroup of a nilpotent group G, then M is normal and $|G/M| = p$ where p is a prime. (A maximal subgroup is a proper subgroup which is not contained in any other proper subgroup. Infinite groups needs not possess maximal subgroups.)
6. Prove that if G is a nilpotent group and N is a minimal normal subgroup of G ($\{e\} \neq N$ is normal and simple), then $N \leq Z(G)$ and $|N| = p$ for some p.

Project 18 (Metabelian groups)**.** A group G is **metabelian** if it admits a proper normal subgroup N such that both N and G/N are abelian. Prove the following statements.

(a) All abelian groups are metabelian.
(b) A group G is metabelian if and only if $G'' = \{e\}$. Deduce that if G is a metabelian group, then G is solvable. Give an example of a solvable group which is not metabelian.
(c) Every subgroup of a metabelian group is metabelian.
(d) All nilpotent groups of class 3 or less are metabelian.

3.4 Linear Groups

In this section, we talk about linear groups over a field. They have many interesting properties and provide us an example of an infinite simple group (Jordan-Moore's theorem).

Let K be a field and $M_n(K)$ be the set of $n \times n$ matrices with entries in K. Then $M_n(K)$ is a ring. Let $\mathrm{GL}_n(K)$ denote the set of multiplicatively invertible elements in $M_n(K)$, called the **general linear group of degree** n, that is,

$$\mathrm{GL}_n(K) = M_n(K)^\times = \{A \in M_n(K) : \det(A) \neq 0\}.$$

Since $\det(AB) = \det A \det B$, $\det : \mathrm{GL}_n(K) \to K^\times$ is a homomorphism (of two groups). Its kernel consists of determinant one matrices, denoted by $\mathrm{SL}_n(K)$ and called the **special linear group of degree** n. It is a normal subgroup of $\mathrm{GL}_n(K)$ with quotient $\mathrm{GL}_n(K)/\mathrm{SL}_n(K)$ isomorphic to $K^\times = K \setminus \{0\}$.

Geometrically, let V be a vector space over K of dimension n. Upon choosing a basis of V, we can represent all linear transformations from V to V via $n \times n$ matrices with entries in K. Then $\mathrm{GL}_n(K)$ represents the invertible linear transformations on V, i.e., those which are one-to-one or equivalently those which are onto.

Theorem 3.4.1. *Let K be a field. Then*

$$Z(\mathrm{GL}_n(K)) = \{\lambda I_n : \lambda \in K^\times\} \quad \textit{and} \quad Z(\mathrm{SL}_n(K)) = \{\lambda I_n : \lambda \in K \textit{ and } \lambda^n = 1\},$$

where I_n is the $n \times n$ identity matrix.

Proof. For M to be in the center of $G = \mathrm{GL}_n(K)$, it must commute with every N in G. In particular, M commutes with the elementary matrices. Recall that multiplying M on the left by an elementary matrix corresponds to performing an elementary row operation; multiplying M on the right by an elementary matrix corresponds to performing an elementary column operation. Thus, multiplying the ith row of M by a nonzero a gives you the same matrix as multiplying the ith column of M by a. This implies that the matrix is diagonal. Then, since interchanging the ith and jth row of M gives us the same matrix as swapping the ith and jth column of M, the ith entry along the diagonal must equal the jth entry along the diagonal, for all i and j. Therefore,

M must be a multiple of I_n. Finally, it is easy to see that all nonzero multiples of I_n do commute with all $N \in G$. Hence, the theorem is proved for $\mathrm{GL}_n(K)$.

For the center of $\mathrm{SL}_n(K)$, we need to use the elementary matrices $X_{ij}(a)$, $i \neq j$, whose entries are the same as that of the identity matrix I_n except for an $a \in K$ in the (i,j) location. It is obtained by performing the row operation $R_i + aR_j$, $i \neq j$ or the column operation $C_j + aC_i$ on I_n. Clearly, $X_{ij}(a) \in \mathrm{SL}_n(K)$ for all $a \in K$ and $i \neq j$.

If M is in the center of $\mathrm{SL}_n(K)$, then M must commute with $X_{ij}(1)$ for all $i \neq j$, so the ith and jth columns and rows must be all zeros except for the (i,i) and (j,j) entries which must be equal. Moreover, the product of the diagonal entries is the determinant which is equal to 1. □

From the above theorem, the center of $\mathrm{GL}_n(K)$ consists of scalar matrices λI_n with $\lambda \in K^\times$ and the center of $\mathrm{SL}_n(K)$ consists of scalar matrices λI_n with $\lambda \in K$ and $\lambda^n = 1$. They are normal and lead to the next definitions.

The quotient group $\mathrm{GL}_n(K)/Z(\mathrm{GL}_n(K)) = \mathrm{PGL}_n(K)$, called the **projective linear group of degree** n. The quotient $\mathrm{SL}_n(K)/Z(\mathrm{SL}_n(K)) = \mathrm{PSL}_n(K)$ is called the **projective special linear group of degree** n.

If K is finite, we may determine the cardinality of each linear group as follows.

Theorem 3.4.2. *If $|K| = q < \infty$, then*

$$|\mathrm{GL}_n(K)| = (q^n - 1)(q^n - q)(q^n - q^2)\dots(q^n - q^{n-1}).$$

Proof. Let $A \in GL_n(K)$. Then the columns of A are linearly independent vectors in K^n. Thus the first column of A can be any nonzero vectors in K^n. The second column must not be multiple of the first column, and the jth column must not be a linear combination of the previous $j-1$ columns for all $j = 2, \dots, n$. By the product rule, we obtain the theorem. □

Corollary 3.4.3. *Let K be a finite field with q elements. Then*

$$|\mathrm{SL}_n(K)| = |\mathrm{PGL}_n(K)| = (q^n - 1)(q^n - q)(q^n - q^2)\dots(q^n - q^{n-2})q^{n-1}$$

and $|\mathrm{PSL}_2(K)| = |\mathrm{SL}_2(K)|$ if $\mathrm{char}K = 2$ and $|\mathrm{PSL}_2(K)| = |\mathrm{SL}_2(K)|/2$ if $\mathrm{char}K \neq 2$.

Proof. They follow from their definitions and Theorem 3.4.2. □

Lemma 3.4.4. *Let K be a field. The group $\mathrm{SL}_2(K)$ is generated by the union of the two subgroups $\left\{ \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix} : \lambda \in K \right\}$ and $\left\{ \begin{bmatrix} 1 & 0 \\ \mu & 1 \end{bmatrix} : \mu \in K \right\}$. Hence, every matrix, in $\mathrm{SL}_2(K)$ is a finite product of matrices which either upper triangular or lower triangular and which have 1's along the diagonal. These matrices are called* **unipotent matrices or transvections**.

Proof. Let $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(K)$. Assume that $c \neq 0$. Perform the following row/column transformations:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \xrightarrow[\text{because } ad-bc=1]{R_1 + \frac{1-a}{c}R_2} \begin{bmatrix} 1 & \frac{d-1}{c} \\ c & d \end{bmatrix} \xrightarrow{R_2 - cR_1} \begin{bmatrix} 1 & \frac{d-1}{c} \\ 0 & 1 \end{bmatrix} \xrightarrow{C_2 + \frac{1-d}{c}C_1} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus,

$$\begin{bmatrix} 1 & 0 \\ -c & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1-a}{c} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 & \frac{1-d}{c} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Hence, $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is a product of transvections.

If $c = 0$, then $d \neq 0$ and the matrix $\begin{bmatrix} a+b & b \\ d & d \end{bmatrix} \in \mathrm{SL}_2(K)$ can be treated as in the first case. However,

$$\begin{bmatrix} a & b \\ 0 & d \end{bmatrix} = \begin{bmatrix} a+b & b \\ d & d \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

and the result follows. □

Theorem 3.4.5. *Let K be a field. The elementary matrices $X_{ij}(a)$, defined in the proof of Theorem 3.4.1, generate $\mathrm{SL}_n(K)$.*

Proof. If $n = 1$, then $\mathrm{SL}_1(K) = \{1\}$ is trivial. Lemma 3.4.4 gives the case $n = 2$. For $n > 2$, the theorem follows from the mathematical induction in a similar manner. □

Lemma 3.4.6. *The elementary matrices $X_{ij}(a)$, defined in the proof of Theorem 3.4.1, are commutators in $\mathrm{SL}_n(K)$ except in the case $n = 2$ and ($|K| = 2$ or 3).*

Proof. If $n \geq 3$, this is easy since there is a third index k and $[X_{ik}(a), X_{kj}(a)] = X_{ij}(a)$. If $n = 2$, we use the commutator relation

$$\left[\begin{bmatrix} \alpha & 0 \\ 0 & \alpha^{-1} \end{bmatrix}, \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix}\right] = \begin{bmatrix} 1 & (\alpha^2 - 1)\beta \\ 0 & 1 \end{bmatrix}.$$

However, given any $\lambda \in K$, the equation $\lambda = (\alpha^2 - 1)\beta$ can be solved for β if and only if there exists a nonzero $\alpha \in K$ so that $\alpha^2 \neq 1$ (i.e., $\alpha \neq \pm 1$). This works as long as $K^\times$ has at least three elements. □

Corollary 3.4.7. *Let K be a field. If $n \geq 2$, then $\mathrm{SL}_n(K)$ is not solvable except in the cases $\mathrm{SL}_2(\mathbb{F}_2)$ and $\mathrm{SL}_2(\mathbb{F}_3)$.*

Remark. $\mathrm{SL}_2(\mathbb{F}_2) \cong \mathrm{PSL}_2(\mathbb{F}_2) \cong S_3$, $\mathrm{SL}_2(\mathbb{F}_3) \cong S_4$ and $\mathrm{PSL}_2(\mathbb{F}_3) \cong A_4$. They are solvable groups. Moreover, $\mathrm{PSL}_2(\mathbb{F}_2)$ and $\mathrm{PSL}_2(\mathbb{F}_3)$ are not simple.

The following theorem was proved by C. Jordan in 1870 for $|K|$ prime. In 1893, after F. Cole discovered a simple group G of order 504, E. H. Moore recognized G as $\mathrm{PSL}_2(\mathbb{F}_8)$, and then proved the simplicity of $\mathrm{PSL}_2(K)$ for all K of size ≥ 4. It provides an example of infinite simple groups.

Theorem 3.4.8. [Jordan-Moore] *Let K be a field with $|K| \geq 4$. Then $\mathrm{PSL}_2(K)$ is a simple group.*

Proof. Using the third isomorphism theorem, it suffices to prove that a normal subgroup N of $\mathrm{SL}_2(K)$ containing a matrix other than $\pm I_2$ must be all of $\mathrm{SL}_2(K)$. Let $A \neq \pm I_2$ be a matrix in N. Then there is a vector $\vec{v}$ in K^2 so that $\vec{v}$ and $A\vec{v}$ are linearly independent over K. This means that $\{\vec{v}, A\vec{v}\}$ is a basis of K^2. The matrix representation of A with respect to this basis is $\begin{bmatrix} 0 & b \\ 1 & d \end{bmatrix}$ (since $A\vec{v} = 0 \cdot \vec{v} + 1 \cdot A\vec{v}$ and $A(A\vec{v}) = b \cdot \vec{v} + d \cdot A\vec{v}$ for some $b, d \in K$). Since $\det A = 1$, we actually have $b = -1$. That is, A is conjugate to $\begin{bmatrix} 0 & -1 \\ 1 & d \end{bmatrix}$. Since N is normal, $\begin{bmatrix} 0 & -1 \\ 1 & d \end{bmatrix}$ is also in N.

Our strategy is to show that N contains all unipotent elements in $\mathrm{SL}_2(K)$ by repeatedly using the fact that "$C^{-1}B^{-1}CB \in N$ for all $C \in \mathrm{SL}_2(K)$ and $B \in N$". First, apply this trick with $B = A$ and $C = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha^{-1} \end{bmatrix}$ ($\alpha \in K^\times$) to get

$$C^{-1}A^{-1}CA = \begin{bmatrix} \alpha^{-2} & d(\alpha^{-2} - 1) \\ 0 & \alpha^2 \end{bmatrix} \in N.$$

Next, repeat the fact with $B' = \begin{bmatrix} \alpha^{-2} & d(\alpha^{-2}-1) \\ 0 & \alpha^2 \end{bmatrix}$ and $C' = \begin{bmatrix} 1 & \mu \\ 0 & 1 \end{bmatrix}$ ($\mu \in K$), we get

$$C'^{-1}B'^{-1}C'B' = \begin{bmatrix} 1 & \mu(\alpha^4-1) \\ 0 & 1 \end{bmatrix} \in N.$$

We get all upper triangular unipotent elements in N as long as there exists an $\alpha \in K^\times$ such that $\alpha^4 \neq 1$. This happens if $|K| \geq 6$ since the polynomial $x^4 - 1$ has at most four distinct roots in $K^\times$ or if $|K| = 4$ since $\mathbb{F}_4^\times$ is cyclic of order 3 and $\alpha^4 = \alpha$ for all $\alpha \in \mathbb{F}_4^\times$. Observe that

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \mu \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ \mu & 1 \end{bmatrix}$$

for all $\mu \in K^\times$. This proves that $N = \mathrm{SL}_2(K)$ if $|K| \geq 4$ and $|K| \neq 5$.

It remains to deal with the case $K = \mathbb{F}_5$. We still have $\begin{bmatrix} \alpha^{-2} & d(\alpha^{-2}-1) \\ 0 & \alpha^2 \end{bmatrix} \in N$ for all $\alpha \in K^\times$. Take $\alpha = 2$ to get $\begin{bmatrix} -1 & -2d \\ 0 & -1 \end{bmatrix} \in N$, and hence $\begin{bmatrix} -1 & -2d \\ 0 & -1 \end{bmatrix}^2 = \begin{bmatrix} 1 & -d \\ 0 & 1 \end{bmatrix} \in N$. Two cases are possible:

(a) $d \neq 0$. The powers of $\begin{bmatrix} 1 & -d \\ 0 & 1 \end{bmatrix}$ give all upper triangular unipotent elements. By conjugating with $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, the lower triangular ones appear. Thus, $N = \mathrm{SL}_2(K)$.

(b) $d = 0$, so $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. We then perform the standard trick with $B = A$ and $C'' = \begin{bmatrix} \delta & 1 \\ -1 & 0 \end{bmatrix}$ ($\delta \in \mathbb{F}_5^\times$), so that

$$A_\delta = C''^{-1}A^{-1}CA = \begin{bmatrix} 1 & -\delta \\ -\delta & \delta^2+1 \end{bmatrix} \in N.$$

Since $\delta \neq 0$, this element is not in the center. Note that its trace is $\delta^2 + 2$ is never zero. Choose $\delta = 1$, say. Then $A_1 \in N$ and A is conjugate to $A' = \begin{bmatrix} 0 & -1 \\ 1 & 3 \end{bmatrix}$ (as at the beginning of the proof and the trace remains the same under conjugation). Apply Case (a), to A', and the proof is complete. □

Exercises 3.4. 1. Show that there is no non-abelian finite simple group of order less than 60. (*Hint*. We may focus on groups of the following orders: $24, 30, 40, 48, 54$ and 56.)

2. Suppose G is a simple group of order 60. Show that:
   (a) G has a subgroup A of order 12
   (b) A has exactly five different conjugates
   (c) there is an injective homomorphism from G to S_5
   (d) both A_5 and H contain every element of S_5 of the form g^2 and therefore every 5-cycle and every 3-cycle
   (e) $H = A_5$.

   Deduce that any simple group of order 60 must be isomorphic to A_5 and hence $\mathrm{PSL}_2(\mathbb{F}_4)$ and $\mathrm{PSL}_2(\mathbb{F}_5)$ are isomorphic to A_5.

Project 19 (The groups $\mathrm{GL}_2(\mathbb{Z}/N\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$)**.** In this project, we determine the structure and the cardinality of the groups $\mathrm{GL}_2(\mathbb{Z}/N\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$.

(a) Prove that for any integer N, the map $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$ obtained by reducing the matrix entries modulo N is a surjective group homomorphism.

(b) Prove that for positive integers M and N, the maps ("reduction modulo N") from $\mathrm{SL}_2(\mathbb{Z}/MN\mathbb{Z})$ to $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$ and from $\mathrm{GL}_2(\mathbb{Z}/MN\mathbb{Z})$ to $\mathrm{GL}_2(Z/N\mathbb{Z})$ are surjective group homomorphisms.
(c) What is the kernel of the homomorphism $\mathrm{GL}_2(\mathbb{Z}/p^e\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{Z}/p\mathbb{Z})$?
(d) What are the order of the groups $\mathrm{GL}_2(\mathbb{Z}/p^e\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/p^e\mathbb{Z})$?
(e) Let $N = p_1^{e_1} \dots p_r^{e_r}$ be the prime factorization of the positive integer N. Show that the reductions modulo $p_j^{e_j}, j = 1, \dots, r$, give isomorphisms

$$\mathrm{GL}_2(\mathbb{Z}/N\mathbb{Z}) \cong \prod_j \mathrm{GL}_2(\mathbb{Z}/p_j^{e_j}) \quad \text{and} \quad \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z}) \cong \prod_j \mathrm{SL}_2(\mathbb{Z}/p_j^{e_j}).$$

(f) What are the order of the groups $\mathrm{GL}_2(\mathbb{Z}/N\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$?

3.5 Free Groups and Presentations

There is a basic method of defining a group G, called a *presentation of G by generators and defining relations*. We have used this method without defining it precisely. For example, $\langle a \rangle$ means the cyclic group generated by a. If a happened to be an element of some larger group G, then $\langle a \rangle$ means the subgroup of G generated by $\langle a \rangle$. It could be infinite cyclic or finite cyclic. More generally, if we were working a particular group G, and $a_1, \dots, a_k \in G$, then $\langle a_1, \dots a_k \rangle$ denoted the subgroup of G generated by $a_1, \dots a_k$.

However, when we were not talking about subgroups of a particular group G, then the brackets $\langle\ \rangle$ had a different meaning as shown by the following examples.

Examples 3.5.1.
1. $\langle a \rangle \cong \mathbb{Z}$ and $\langle a : a^n = e \rangle \cong \mathbb{Z}_n$.
2. $\langle a, b : a^n = e, b^m = e, ab = ba \rangle = \langle a, b : a^n = e, b^m = e, aba^{-1}b^{-1} = e \rangle \cong \mathbb{Z}_n \times \mathbb{Z}_m$.
3. $\langle a_1, \dots, a_k : a_1^{n_1} = \cdots = a_k^{n_k} = e, a_i a_j a_i^{-1} a_j^{-1} = e \text{ if } i \neq j \rangle \cong \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$.
4. $\langle a_1, \dots, a_k : a_i a_j a_i^{-1} a_j^{-1} = e \text{ if } i \neq j \rangle \cong \mathbb{Z} \times \cdots \times \mathbb{Z}$ (k copies).
5. $D_n = \langle a, b : a^n = b^2 = e, bab^{-1} = a^{-1} \rangle = \langle a, b : a^n = b^2 = e, bab^{-1}a = e \rangle$ is the dihedral group of order $2n$.
6. $D_\infty = \langle a, b : b^2 = e, bab^{-1} = a^{-1} \rangle = \langle a, b : b^2 = e, bab^{-1}a = e \rangle$ is the infinite dihedral group.
7. $\langle a, b : a[a,b] = [a,b]a, b[a,b] = [a,b]b \rangle \cong \left\{ \begin{bmatrix} 1 & p & q \\ 0 & 1 & r \\ 0 & 0 & 1 \end{bmatrix} : p, q, r \in \mathbb{Z} \right\}$. An isomorphism is given by

$$a \mapsto \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad b \mapsto \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Observe that $c = a^{-1}b^{-1}ab = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

In each of the above examples the data inside the brackets $\langle\ \rangle$ is sufficient to describe a group, that is, it gives the multiplication table for a groups. We call such an expression a presentation for the group. It turns out that every group has a presentation and every presentation defines a group. However, it is generally difficult to decide if a group defined by a presentation is isomorphic to an explicitly given group.

Let A be any (not necessarily finite) set of elements a_i for $i \in I$. We think of A as an **alphabet set** and of the a_i as letters in the alphabet set. Any symbol of the form a_i^n with $n \in \mathbb{Z}$ is a **syllable** and a finite string w of syllables written in juxtaposition is a **word**. We also introduce the **empty word** 1, which has no syllables. A word on A is **reduced** if $w = 1$ or the string $a^i a^{-i}$ or $a^{-i}a^i$ does not appear in w for all $a \in A$ and $i \in \mathbb{N}$.

Let A be a set. Write $F[A]$ for the set of all reduced words formed from our alphabet A. For convenience, we may let $F[\emptyset] = \{1\}$. We make $F[A]$ into a group by the juxtaposition $w_1 w_2$ of

two words w_1 and w_2 with reduction of strings $a^i a^{-i}$ or $a^{-i}a^i$ (if any) for all $a \in A$ and $i \in \mathbb{N}$. It is called the **free group generated by** A.

Example 3.5.2. The only example of a free group that has occurred before is $\mathbb{Z}$, which is free on one generators. Clearly, every free group is infinite.

Example 3.5.3. $F_2 = \langle x, y\rangle$. The element of F_2 are all words in x and y. More precisely, F_2 is the disjoint union of the following seven sets.

1. $\{1\}$
2. $\{x^i : i \in \mathbb{Z} \setminus \{0\}\}$
3. $\{y^i : i \in \mathbb{Z} \setminus \{0\}\}$
4. $\{x^{i_1}y^{j_1} \dots x^{i_k}y^{j_k} : k > 0, i_r, j_r \in \mathbb{Z} \setminus \{0\}\}$
5. $\{x^{i_1}y^{j_1} \dots x^{i_k}y^{j_k}x^{i_{k+1}} : k > 0, i_r, j_r \in \mathbb{Z} \setminus \{0\}\}$
6. $\{y^{j_1}x^{i_1} \dots y^{j_k}x^{i_k} : k > 0, i_r, j_r \in \mathbb{Z} \setminus \{0\}\}$
7. $\{y^{j_1}x^{i_1} \dots y^{j_k}x^{i_k}y^{j_{k+1}} : k > 0, i_r, j_r \in \mathbb{Z} \setminus \{0\}\}$

Let G be a group and let A be a subset of G such that $\langle A\rangle = G$. If G is isomorphic to $F[A]$ under the map $\varphi : G \to F[A]$ such that $\varphi(a) = a$ for all $a \in A$, then G is said to be **free on** A. A group is **free** if it is free on some nonempty set A.

Theorem 3.5.1. [Universal Mapping Property of a Free Group] *Let A be a nonempty set. Suppose H is any group and there is a function $\phi : A \to H$.*

1. *There is a unique homomorphism $\Phi : F[A] \to H$ extending ϕ.*
2. *If $\operatorname{im}\phi$ generates H, then $\Phi : F[A] \to H$ is a surjection.*
3. *If G is a group and $\theta : G \to F[A]$ is an onto homomorphism, then there is a homomorphism $\Phi : F[A] \to G$ such that $\theta \circ \Phi = \operatorname{id}_{F[A]}$, the identity map on $F[A]$.*

Proof. (1) is clear and (2) follows immediately from (1).
(3) Since θ is onto, for each $a \in A$, there is a $g_a \in G$ such that $\theta(g_a) = a$. By (1), there is a unique homomorphism $\Phi : F[A] \to H$ with $\Phi(a) = g_a$ for all $a \in A$. Then $\theta \circ \Phi : F[A] \to F[A]$ is the identity map. □

Similarly, we can show that

Corollary 3.5.2. *Let S be a set. Then there is a unique free group on S.*

Proof. Let G_1 and G_2 be free groups on S. Then S is a subset of both G_1 and G_2. Consider the inclusion maps $\iota_1 : S \to G_1$ and $\iota_2 : S \to G_2$ and the result follows from the uniqueness of the universal mapping property. □

Corollary 3.5.3. *Every group H is a homomorphic image of a free group.*

Proof. Let A be a set for which there exists a bijection $\phi : A \to H$ (e.g., take $A = H$ and $\phi = \operatorname{id}_H$), and let $G = F[A]$. By the universal mapping property, there is an onto homomorphism $\Phi : G \to H$ extending ϕ. Therefore, $G/(\ker\Phi) \cong H$. □

We refer the reader to reference textbooks for proofs of the next three theorems. They are stated simply to inform us of these interesting facts.

Theorem 3.5.4. *If a group G is free on A and also on B (not necessarily finite), then the sets A and B have the same number of elements; that is, any two sets of generators of a free group have the same cardinality.*

We shall prove this theorem for the finite basis case with some result on finitely generated free abelian group (Corollary 4.2.7) in the next chapter.

If G is free on a set A, the number of elements in A is called the **rank of** G.

Theorem 3.5.5. *Two free groups are isomorphic if and only if they have the same rank.*

Theorem 3.5.6. [Schreier] *A nontrivial proper subgroup of a free group is free.*

This is not trivial to prove. There is a nice proof of this result using covering spaces (cf. J.-P. Serre, *Trees*, Springer-Verlag, 1980).

Example 3.5.4. Let $y_l = x^l y x^{-l}$ for $l \geq 0$. Then y_l, $l \geq 0$, are free generators for the subgroup of $F_2 = \langle x, y \rangle$ that they generate. This illustrates that although a subgroup of a free group is free, the rank of the subgroup may be much greater than the rank of the whole group!

Let $G \xrightarrow{\theta} H \xrightarrow{\phi} K$ be a sequence of groups homomorphisms. We say that it is **exact** at H if $\operatorname{im}\theta = \ker\phi$. A **short exact sequence of groups** is a sequence of groups and homomorphisms

$$1 \longrightarrow G \xrightarrow{\theta} H \xrightarrow{\phi} K \longrightarrow 1$$

which is exact at G, H and K. In other words, if θ is 1-1, ϕ is onto and $\operatorname{im}\theta = \ker\phi$.

Remark. If N is a normal subgroup of G, then $1 \to N \to G \to G/N \to 1$ is exact. Conversely, if $1 \to N \xrightarrow{\iota} G \to H \to 1$ is exact, then N is normal in G and $H \cong G/N$. Thus short exact sequences are just another notation for normal subgroups and factor groups.

A **presentation for a group** G is an expression

$$G = \langle g_1, \ldots, g_r : w_1 = \cdots = w_t = 1 \rangle$$

where $w_1, \ldots, w_t$ are words in $g_1, \ldots, g_r$ such that the following two properties are satisfied: (1) $g_1, \ldots, g_r$ generate G and (2) the conditions that $w_1 = w_2 = \cdots = w_t = 1$ are sufficient to define the multiplication table of G. Here, $g_1, \ldots, g_r$ are called **generators** of G in the presentation and $w_1, w_2, \ldots, w_t$ are called **defining relations**.

Note that the free group of rank n is the group $F_n = \langle x_1, \ldots, x_n : \ \rangle$ given by a presentation with n generators and zero defining relation.

Remark. The elements of $\langle x_1, \ldots, x_n \rangle$ are words in $x_1, \ldots, x_n$. Suppose $w = w(x_1, \ldots, x_n)$ is any such word. Then if G is any group, we can think of w as a *function* $G \times \cdots \times G \to G$ such that $(g_1, \ldots, g_n) \mapsto w(g_1, \ldots, g_n)$. For example, if $w(x_1, x_2) = [x_1, x_2] = x_1 x_2 x_1^{-1} x_2^{-1}$, then $w(g_1, g_2) = g_1 g_2 g_1^{-1} g_2^{-1}$.

Remark. If we have $F_n = \langle x_1, \ldots, x_n \rangle$ is a free group and $G = \langle y_1, \ldots, y_n : w_1 = \cdots = w_t = 1 \rangle$, then the universal mapping property of a free group says that $\phi(x_i) = y_i$ defines an onto homomorphism $\phi : F_n \to G$. This means we have a short exact sequence

$$1 \longrightarrow \ker\phi \xrightarrow{\iota} F_n \xrightarrow{\phi} G \longrightarrow 1.$$

What is the kernel of ϕ? $\ker\phi$ is a normal subgroup of F_n and contains $w_i(x_1, \ldots, x_n)$ for $i = 1, \ldots, t$. In fact, $\ker\phi$ is the smallest normal subgroup of F_n which contains $w_i(x_1, \ldots, x_n)$ for $i = 1, \ldots, t$.

Let G be a group and S a subset of G. The **normal closure of** S **in** G, denoted by $\langle S \rangle^G$, is the smallest normal subgroup of G containing S. It is the subgroup of G generated by all conjugates of elements of S by elements of G. That is,

$$\langle S \rangle^G = \langle xyx^{-1} : x \in G \text{ and } y \in S \rangle$$

and so

Theorem 3.5.7. *Let $G = \langle x_1, \ldots, x_n : w_1 = \cdots = w_t = 1\rangle$. Then $G \cong F/N$ where N is the normal closure of $\{w_1, \ldots, w_t\}$ in the free group $F = F[\{x_1, \ldots, x_n\}]$.*

Example 3.5.5. Consider the free group $F_2 = \langle x, y\rangle$. Let $G = \langle x, y : xyx^{-1}y^{-1} = 1\rangle \cong F_2/N$. Since G is abelian, $F_2' \subseteq N$. But $xyx^{-1}y^{-1} \in N$ and N is the smallest, so $N = F_2'$.

Example 3.5.6. Consider the quaternion group $Q_8 = \langle a, b : a^4 = 1, a^2 = b^2, ba = a^3b\rangle$ of order eight. We shall determine the structure of Q_8/Q_8'. Since $|a| = 4$ and $ba = a^3b$, $\langle a\rangle \triangleleft Q_8$ and $b \notin \langle a\rangle$, so $Q_8/\langle a\rangle = \{\langle a\rangle, b\langle a\rangle\}$. Then $Q_8' \subseteq \langle a\rangle$. Since $a^2 = a^-1bab^{-1} \in Q_8'$, $\langle a^2\rangle \subseteq Q_8' \subseteq \langle a\rangle$. In addition, $ba^2b^{-1} = a^{-2}$. Thus, $\langle a^2\rangle$ is normal in Q_8. Since $|Q_8/\langle a^2\rangle| = 4$, it is abelian. Hence, $Q_8' = \langle a^2\rangle$. Note that $a^2Q_8' = b^2Q_8' = Q_8'$. Therefore, $Q_8/Q_8' \cong \mathbb{Z}_2 \times \mathbb{Z}_2$.

Theorem 3.5.8. [von Dyck's Theorem `/fon dike/`] *Let G be given by a presentation*

$$G = \langle x_1, \ldots, x_n : w_1 = \cdots = w_t = 1\rangle.$$

Suppose H is any group which satisfies:
1. H is generated by $h_1, \ldots, h_n$ and 2. $w_i(h_1, \ldots, h_n) = 1$ for $i = 1, \ldots, t$.
Then there is a unique onto homomorphism $\phi : G \to H$ for which $\phi(x_i) = h_i$.

Proof. By Theorem 3.5.7, $G \cong F/N$, where F is a free group on $\{x_1, \ldots, x_n\}$ and N is the normal closure of $\{w_1, \ldots, w_t\}$. By the assumption $N \subseteq \ker\phi$, so ϕ induces a (well defined) homomorphism $x_i = x_iN \mapsto h_i$ for all $i \in \{1, \ldots, n\}$. □

Example 3.5.7. Classify all groups G of order six.

Proof. Since $6 = 2 \cdot 3$, G contains elements a and b such that $|a| = 2$ and $|b| = 3$ and $G = \langle a, b\rangle$. Since $\langle b\rangle$ is normal in G, $aba^{-1} \in \langle b\rangle$. Thus, $aba^{-1} = b$ or $aba^{-1} = b^{-1}$. If $aba^{-1} = b$, then G is abelian, so $G \cong \mathbb{Z}_6$. Assume that $aba^{-1} = b^{-1}$. Then $G = \langle a, b : a^2, b^3, aba^{-1} = b^{-1}\rangle$. Note that $S_3 = \langle(12), (123)\rangle$ and $(12)(123)(12)^{-1} = (132) = (123)^{-1}$. By von Dyck's theorem, there is an onto homomorphism from G to S_3. But $|G| = 6 = |S_3|$, $G \cong S_3$. □

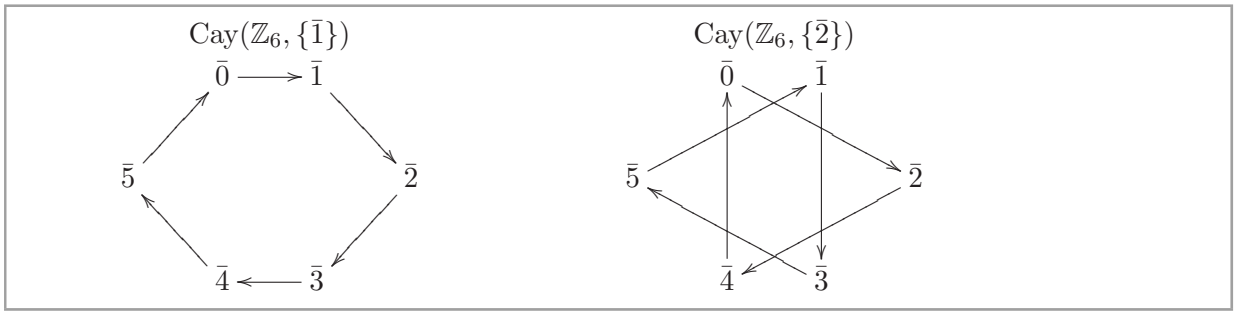
To close this section, we introduce a graphical representation of a group given by a set of generators and relations. The idea was originated by Cayley in 1878. It provided a method of visualizing a group and connects two important branches of modern mathematics—groups and graphs. This also has many applications to computer science.

Let G be a finite group and S a nonempty subset of G. To avoid loops, we shall assume that $e \notin S$. The **Cayley digraph of G with generating set S** is a digraph $\mathrm{Cay}(G, S)$ such that

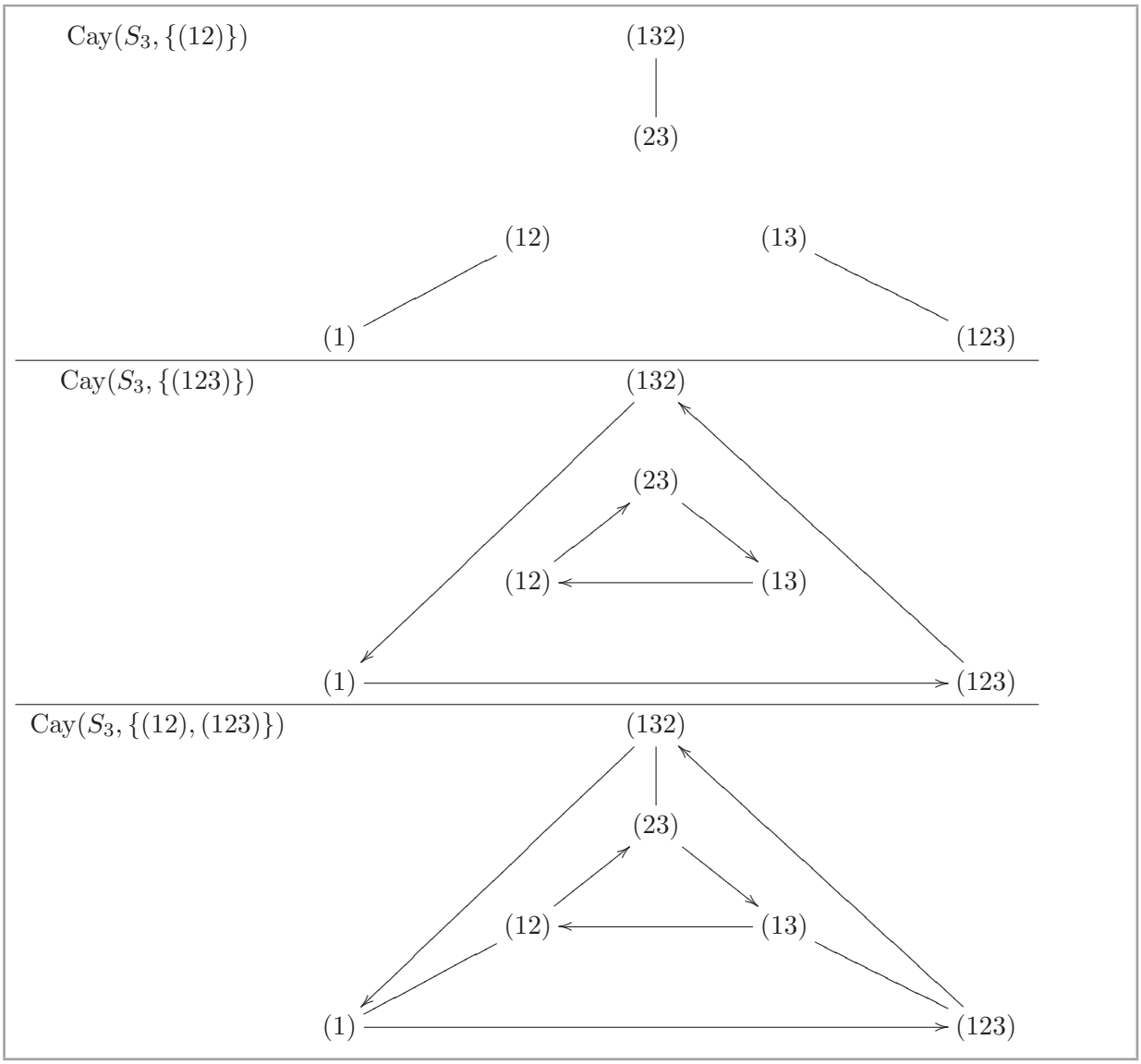
1. each element of G is a vertex of $\mathrm{Cay}(G, S)$, and
2. for $x, y \in G$, there is a directed edge from x to y if and only if $x^{-1}y \in S$.

We call the finite sequence of vertices such that $x_1 \longrightarrow x_2 \longrightarrow \cdots \longrightarrow x_n$ a **directed path**. A directed graph is **connected** if for every pair of vertices x, z, there is a directed path from x to z. A maximal connected subdigraph of a digraph is called a **connected component**.

Example 3.5.8. For the group $(\mathbb{Z}_6, +)$, we consider the following two digraphs. The first one is connected while the second one is not. The second one has two connected components.



Example 3.5.9. For the group $(S_3, \circ)$, we consider the following three digraphs. The first and the second ones are not connected. They have three connected components and two connected component, respectively. The last one is connected. Note also that since $(12)^{-1} = (12)$, some edges are undirected.



For a non-empty subset S of a group G, we let $S^{-1} = s^{-1} : s \in S$. Observe that if we assume further that $e \notin S$ and $S = S^{-1}$, then the Cayley graph $\mathrm{Cay}(G, S)$ is an undirected graph without loops. Hence, it is a graph.

A regular graph is a graph such that each vertex has the same number of neighbors which is called the **degree** of a regular graph. Note that for each $x \in G$, we have x —— xs for all $s \in S$. From the cancellative property in G, each vertex has $|S|$ neighbors, so $\mathrm{Cay}(G, S)$ is a regular graph of degree $|S|$.

Theorem 3.5.9. *Let G be a group and S a nonempty subset of G such that $e \notin S$ and $S = S^{-1}$. Then*

1. *$\mathrm{Cay}(G, S)$ is a regular graph of degree $|S|$.*
2. *For $x, z \in G$, there is a path from x to z if and only if $x^{-1}z \in \langle S \rangle$.*
3. *The number of connected components of $\mathrm{Cay}(G, S)$ is the index $[G : \langle S \rangle]$.*
4. *$\mathrm{Cay}(G, S)$ is connected if and only if $\langle S \rangle = G$.*

Proof. Clearly, (4) follows from (3) and (3) follows from (2), respectively. To prove (2), let $x, z \in G$. Since $S = S^{-1}$, $\langle S\rangle = \{s_1 s_2 \dots s_k : k \in \mathbb{N} \cup \{0\}$ and $s_1, s_2, \dots, s_k \in S\}$. Also, any path from x to z is given by

$$x \text{ —— } xs_1 \text{ —— } xs_1s_2 \text{ —— } \cdots \text{ —— } xs_1s_2 \dots s_k = z$$

for some $s_1, s_2, \dots, s_k$. Thus, there is a path from x to z if and only if $x^{-1}z = s_1 s_2 \dots s_k \in \langle S\rangle$. □

Finally, we provides some examples of Cayley graph arising from number theory and ring theory.

Example 3.5.10. For $n \geq 3$, consider the additive group $(\mathbb{Z}_n, +)$. We know that for $\bar{a} \in \mathbb{Z}_n$, $\bar{a} \in \mathbb{Z}_n^\times \Leftrightarrow -\bar{a} \in \mathbb{Z}_n^\times$, so $\mathbb{Z}_n^\times = -\mathbb{Z}_n^\times$. The Cayley graph $X_n = \mathrm{Cay}(\mathbb{Z}_n, \mathbb{Z}_n^\times)$ is called the **unitary Cayley graph**. More generally, if R is a finite ring and consider $(R, +)$ is an additive group, then the **unitary Cayley graph for** R is the Cayley graph $X_R = \mathrm{Cay}(R, R^\times)$.

Example 3.5.11. Let R be a finite commutative ring with unity $1 \neq 0$. Consider the exact sequence of groups

$$1 \longrightarrow K_R \longrightarrow R^\times \xrightarrow{\theta} (R^\times)^2 \longrightarrow 1$$

where $\theta : a \to a^2$ is the square mapping on $R^\times$ with kernel K_R and $(R^\times)^2 = \{a^2 : a \in R^\times\}$. Note that K_R consists of the identity and all elements of order two in $R^\times$. Let $T_R = K_R(R^\times)^2$. Since $-1 \in T_R$, the Cayley graphs $H_R = \mathrm{Cay}(R, T_R)$ is undirected. This graph is a subgraph of the unitary Cayley graph. It is called the **restricted unitary Cayley graphs induced from the square mapping**. It is a generalization of Paley graphs (see Project 25).

Exercises 3.5.

1. (a) Prove that the derived group of a free group consists of those words in which the sum of the exponents for each generator is equal to zero (e.g., $x_1x_2^{-1}x_1^{-2}x_2x_1$).
   (b) Let F be a free group generated by $x_1, x_2, \dots, x_r$. Show that each element of F/F' is of the form $(x_1^{m_1}x_2^{m_2}\dots x_r^{m_r})F'$. Now use (a) to show that $F/F' \cong \mathbb{Z}^r$, i.e., F/F' is the free abelian group of rank r.
2. Determine the structure of G/G', when G is given by
   (i) $a^6 = b^2 = (ab)^2 = 1$; (ii) $a^6 = 1, b^2 = (ab)^2 = a^3$.
3. Show that if G is generated by a and b subject to the relations $a^{-1}ba = b^2$ and $ab = ba^2$, then $G = \{1\}$.
4. Let G be a group. For $a, b \in G$, let $[a, b] = aba^{-1}b^{-1}$ and $a^b = bab^{-1}$.
   (a) Prove that $[a, bc] = [a, b][a, c]^b$ for all $a, b, c \in G$.
   (b) If $H = \langle x, y, z \in G : [x, y] = y, [y, z] = z$ and $[z, x] = x\rangle$, show that $H = \{e\}$.
5. If G is a non-abelian group of order eight, show that G is isomorphic to D_4 or Q_8.
6. Let $D_4 = \langle a, b : a^2 = b^2 = (ab)^4 = 1\rangle$. Draw $\mathrm{Cay}(D_4, \{a, b\})$. Why is this graph undirected?

"Algebraic Graph Theory" is a branch of mathematics in which algebraic methods are applied to problems about graphs. There are three main branches of algebraic graph theory, involving the use of linear algebra, the use of group theory, and the study of graph invariants. The first branch of algebraic graph theory involves the study of graphs in connection with linear algebra. Mainly, it studies the spectrum of the adjacency matrix, or the Laplacian matrix of a graph (this part of algebraic graph theory is also called spectral graph theory). Secondly, algebraic graph theory involves the study of graphs in connection to group theory, particularly automorphism groups and geometric group theory. The focus is placed on various families of graphs based on symmetry such as symmetric graphs, vertex-transitive graphs, edge-transitive graphs, distance-transitive graphs, distance-regular graphs, and strongly regular graphs. Finally, the third branch of algebraic graph theory concerns algebraic properties of invariants of graphs, and especially the

chromatic polynomial, the Tutte polynomial and knot invariants. The references and related work can be found in Godsil and Royle's celebrated book [14].

There are many graphs arising from number theory and finite field theory such as Cayley graphs, symplectic graphs (Subsection 4.7.2), Paley graphs (see Project 25) and functional digraphs (e.g., [36, 37]). Moreover, these graphs can be extended to general results in abstract integral domains, e.g., function fields, PID or even UFD (see, [27] and the following project).

Project 20 (gcd-graph)**.** We consider a unique factorization domain D. Let $c \in D$ be a nonzero nonunit element. Assume that the commutative ring $D/(c)$ is finite. Let $\mathcal{C}$ be a set of proper divisor of c. Define the **gcd-graph**, $D_c(\mathcal{C})$, to be a graph whose vertex set is the quotient ring $D/(c)$ and edge set is $\{\{x+(c), y+(c)\} : x, y \in D \text{ and } \gcd(x-y, c) \in \mathcal{C}\}$. The gcd considered here is unique up to associate. Prove that that $D_c(\{1\}) = G_{D/(c)} = \mathrm{Cay}(D/(c), D/(c)^\times)$, the unitary Cayley graph in Example 3.5.10. This gcd-graph on a quotient ring of a unique factorization domain (UFD) introduced in [27] generalizes a gcd-graph or an integral circulant graph (i.e., its adjacency matrix is circulant and all eigenvalues are integers) defined over $\mathbb{Z}_n, n \geq 2$, (see [29, 39]). An integral circulant graph can also be considered as an extension of a unitary Cayley graph.

Project 21 (Energy of a graph)**.** Let G be a graph with vertex set $\{v_1, v_2, \ldots, v_n\}$. The **adjacency matrix of** G, denoted by $A(G)$, is the $n \times n$ matrix given by

$$a_{ij} = \begin{cases} 1 & \text{if there is an edge joining } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases}$$

The **eigenvalues and eigenvectors of a graph** G are defined to be the eigenvalues and eigenvectors of its adjacency matrix $A(G)$. The sum of absolute values of all eigenvalues of a graph G is called the **energy** of G and denoted by $E(G)$.

Let R be a finite local ring with unique maximal ideal M of size m. For $k, l \in \mathbb{N}$, we write $\mathbf{0}_{k\times l}$ and $J_{k\times l}$ for the $k \times l$ matrix whose all entries are 0 and 1, respectively. We also use $\vec{0}_k = \mathbf{0}_{k\times 1}$ and $\vec{1}_k = J_{k\times 1}$.

(a) Prove that the adjacency matrix of the unitary Cayley graph $\mathrm{Cay}(R, R^\times)$ is given by

$$A(\mathrm{Cay}(R, R^\times)) = \begin{bmatrix} 0_{m\times m} & J_{m\times m} & J_{m\times m} & \cdots & J_{m\times m} \\ J_{m\times m} & 0_{m\times m} & J_{m\times m} & \cdots & J_{m\times m} \\ J_{m\times m} & J_{m\times m} & 0_{m\times m} & \cdots & J_{m\times m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ J_{m\times m} & J_{m\times m} & J_{m\times m} & \cdots & 0_{m\times m} \end{bmatrix}.$$

(b) Compute the eigenvalues and eigenvectors of the unitary Cayley graph $\mathrm{Cay}(R, R^\times)$ and determine its energy.

The energy is a graph parameter introduced by Gutman (see [25] and [21] for a good survey) arising from the Hückel molecular orbital approximation for the total π-electron energy. Nowadays, the energy of graph is studied for purely mathematical interest.

---

4 | Modules and Noetherian Rings

Modules can be considered as a generalization of vector spaces. It is like we study linear algebra over a ring. In this chapter, we first cover basic concepts of modules. Next, we work on free modules. Projective and injective modules are introduced. We also present the proof of the structure theorems for modules over a PID. Finally, we talk about Noetherian and Artinion rings. Noetherian rings have a lot of applications in algebraic geometry and algebraic number theory.

Each ring R that we consider will be assumed to contain a multiplicative identity element, which will be denoted by 1. We shall therefore regard the possession of such an identity as one of the defining conditions of the ring concept and also assume $1 \neq 0$.

4.1 Modules

The definition of a module is similar to a vector space. However, now our scalars are in a ring.

Let R be a ring. We say that M is a **(left) R-module** provided:

1. $(M, +)$ is an additive abelian group
2. there is a multiplication $R \times M \to M$ which satisfies for all $\alpha, \beta \in R$ and $u, v \in M$,
   (a) $\alpha(u + v) = \alpha u + \alpha v$,
   (b) $(\alpha + \beta)u = \alpha u + \beta u$ and
   (c) $\alpha(\beta u) = (\alpha\beta)u$
3. if 1 is the unity of R, then $1 \cdot u = u$ for all $u \in M$.

Remark. Note that we abuse notations by not distinguishing between the addition in M or in R and the multiplication in R or the multiplication $R \times M \to M$. A right R-module can be defined analogously.

Examples 4.1.1.

1. If $R = F$, a field, an F-module is just a **vector space over** F.
2. Any abelian group A is a $\mathbb{Z}$-module, where the action of $\mathbb{Z}$ is given by for $a \in A$,

$$0 \cdot a = 0_A, n \cdot a = \underbrace{a + a + \cdots + a}_{n} \text{ if } n > 0 \quad \text{and}$$

$$n \cdot a = \underbrace{(-a) + (-a) + \cdots + (-a)}_{-n} \text{ if } n < 0.$$

3. Let F be a field and $R = M_n(F)$ the ring of $n \times n$ matrices over F. Let $V = F^n$ be n-dimensional vector space of $n \times 1$ column vectors over F. Then V is an R-module where the multiplication $R \times V \to V$ is given by $A \cdot \vec{v} = A\vec{v}$ (matrix multiplication).
4. Let R be a ring. Then R is an R-module with the usual multiplication $R \times R \to R$. More generally, any left ideal A of R is a left R-module. In fact, a subset A of R is a left ideal in R if and only if the left multiplication $R \times A \to A$ makes A into a left R-module. That is, the set of left ideals of R is the set left R-modules of R. Hence, if R is a ring, then R can be viewed as an R-module, called a **regular left [right] R-module**, and is denoted by ${}_RR$ $[R_R]$.

We collect basic terminologies about modules in the following definitions.

Let R be a ring. We say that N is an R**-submodule or submodule** of an R-module M if N is a subgroup of M as an additive group and the multiplications $R \times M \to M$ and $R \times N \to N$ agree on N.

Let R be a ring. The **direct sum** of R-modules M and N is the abelian group direct sum of M and N

$$M \oplus N = \{(m, n) : m \in M, n \in N\}$$

with the action of R on $M \oplus N$ given by

$$r(m, n) = (rm, rn).$$

One often writes $m + n$ in place of (m, n).

Let R be a ring and let M and N be R-modules.

1. A map $f : M \to N$ is an R**-module homomorphism** provided
   (a) $f : M \to N$ is a homomorphism of abelian groups and
   (b) if $r \in R$ and $m \in M$, then $f(rm) = rf(m)$.
2. We call a diagram of R-module homomorphisms

   $$M \xrightarrow{f} N \xrightarrow{g} P$$

   exact if $\operatorname{im} f = \ker g$. More generally, a sequence of R-modules and homomorphisms

   $$\cdots \longrightarrow M_1 \xrightarrow{f_1} M_2 \xrightarrow{f_2} M_3 \longrightarrow \cdots$$

   that may be finite or run to infinity in either direction is called **exact** if for any three consecutive terms the subsequence $M_i \longrightarrow M_{i+1} \longrightarrow M_{i+2}$ is exact. An exact sequence of the form

   $$0 \longrightarrow M' \xrightarrow{f} M \xrightarrow{g} M'' \longrightarrow 0$$

   is called a **short exact sequence**. This means that f is a monomorphism (1-1), g is an epimorphism (onto) and $\ker g = \operatorname{im} f$.
3. If N is a submodule of M, then the quotient group $(M/N, +)$ can be made into an R-module by defining $r(x + N) = rx + N$. It is called a **factor module of** M **by** N.
4. Let $f : M \to N$ be a homomorphism of R-modules. The **kernel** of f is

   $$\ker f = \{m \in M : f(m) = 0_N\}$$

   and the **cokernel** of f is $N/\operatorname{im} f$. They are clear that $\ker f$ and $\operatorname{im} f$ are R-submodules of M and N, respectively. Evidently, f is surjective if and only if $\operatorname{coker} f = 0$. In any case, we have

   $$0 \longrightarrow \ker f \hookrightarrow M \xrightarrow{f} N \longrightarrow \operatorname{coker} f \longrightarrow 0$$

   is exact.

Remark. The isomorphism theorems also hold for R-modules and their homomorphisms. Note however that the first isomorphism theorem will say a bit more, because $\operatorname{coker} f = N/\operatorname{im} f$ is an R-module. This is not the case with homomorphisms of groups or rings: If $f : G \to H$ is a group homomorphism, then $f(G) = \operatorname{im} f$ is not in general a normal subgroup of H, hence $H/\operatorname{im} f$ is not in general a group. And if $f : R \to S$ is a ring homomorphism, then $f(R) = \operatorname{im} f$ is never an ideal in S (unless it is all of S), so $S/\operatorname{im} f$ is not a ring.

The isomorphism theorems can be stated as theorems about commutative diagrams and exact sequences. The use of diagrams to describe module homomorphisms is very common, we now give the isomorphism theorems in their diagram theoretic versions. Note that many homomorphisms are projections or injections implicitly defined by the diagram. The proofs of the isomorphism theorems are left as exercises.

Theorem 4.1.1. [First Isomorphism Theorem] *Let M and N be R-modules. Then the following diagram of R-modules has an exact row and a commutative square.*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \ker f & \hookrightarrow & M & \xrightarrow{f} & N & \longrightarrow & \operatorname{coker} f \longrightarrow 0 \\ & & & & \downarrow{\scriptstyle \pi} & & \uparrow{\scriptstyle i} & & \\ & & & & M/\ker f & \xrightarrow[\bar{f}]{\cong} & \operatorname{im} f & & \end{array}$$

Theorem 4.1.2. [Second Isomorphism Theorem] *Let N_1 and N_2 be submodules of an R-module N. Then there is a commutative diagram with exact rows in which the vertical map of the right is an isomorphism.*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & N_1 \cap N_2 & \longrightarrow & N_2 & \longrightarrow & N_2/(N_1 \cap N_2) & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow{\scriptstyle \cong} & & \\ 0 & \longrightarrow & N_1 & \longrightarrow & N_1 + N_2 & \longrightarrow & (N_1 + N_2)/N_1 & \longrightarrow & 0 \end{array}$$

Theorem 4.1.3. [Third Isomorphism Theorem] *If $N_2 \leq N_1 \leq N$ are R-modules, then the following diagram is commutative and has exact rows:*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & N_1 & \longrightarrow & N & \longrightarrow & N/N_1 & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow{\scriptstyle \mathrm{id}} & & \\ 0 & \longrightarrow & N_1/N_2 & \longrightarrow & N/N_2 & \longrightarrow & N/N_1 & \longrightarrow & 0 \end{array}$$

That is, $N/N_1 \cong (N/N_2)/(N_1/N_2)$.

Theorem 4.1.4. *Let N_1 and N_2 be submodules of an R-module N. Then the following diagram is commutative and has exact rows and columns.*

$$\begin{array}{ccccccccc} & & 0 & & 0 & & 0 & & \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & N_1 \cap N_2 & \longrightarrow & N_2 & \longrightarrow & N_2/(N_1 \cap N_2) & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & N_1 & \longrightarrow & N & \longrightarrow & N/N_1 & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & N_1/(N_1 \cap N_2) & \longrightarrow & N/N_2 & \longrightarrow & N/(N_1 + N_2) & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ & & 0 & & 0 & & 0 & & \end{array}$$

Proof. The commutativity and the exactness of the top two rows and the two left columns are clear. The exactness of the third row and right column come, respectively, from the isomorphisms $(N_1 + N_2)/N_2 \cong N_1/(N_1 \cap N_2)$ and $(N_1 + N_2)/N_1 \cong N_2/(N_1 \cap N_2)$. □

The next theorem is widely used in mathematics. It is proved by the technique called “diagram chasing”.

Theorem 4.1.5. [5-Lemma] *Suppose the following diagram is commutative and has exact rows.*

$$\begin{array}{ccccccccc} A_1 & \xrightarrow{\alpha_1} & A_2 & \xrightarrow{\alpha_2} & A_3 & \xrightarrow{\alpha_3} & A_4 & \xrightarrow{\alpha_4} & A_5 \\ \downarrow{\scriptstyle f_1} & & \downarrow{\scriptstyle f_2} & & \downarrow{\scriptstyle f_3} & & \downarrow{\scriptstyle f_4} & & \downarrow{\scriptstyle f_5} \\ B_1 & \xrightarrow[\beta_1]{} & B_2 & \xrightarrow[\beta_2]{} & B_3 & \xrightarrow[\beta_3]{} & B_4 & \xrightarrow[\beta_4]{} & B_5 \end{array}$$

If f_1, f_2, f_4 and f_5 are isomorphisms, so is f_3. More precisely,
1. *if f_1 is onto and f_2 and f_4 are 1-1, then f_3 is 1-1, and*
2. *if f_5 is 1-1 and f_2 and f_4 are onto, then f_3 is onto.*

Proof. (1) Assume f_1 is onto and f_2 and f_4 are 1-1. Suppose $x \in A_3$ and $f_3(x) = 0$. We shall show that $x = 0$. Since $f_4(\alpha_3(x) = \beta_3(f_3(x)) = \beta_3(0) = 0$ and f_4 is 1-1, $\alpha_3(x) = 0$, so $x \in \ker \alpha_3 = \operatorname{im} \alpha_2$ from the exactness of the top row. Thus, $x = \alpha_2(y)$ for some $y \in A_2$. Then $0 = f_3(x) = f_3(\alpha_2(y)) = \beta_2(f_2(y))$, so $f_2(y) \in \ker \beta_2 = \operatorname{im} \beta_1$ from the exactness of the bottom row. Thus, $f_2(y) = \beta_1(z)$ for some $z \in B_1$. Since f_1 is onto, there is a $u \in A_1$ with $f_1(u) = z$. Then $f_2(y) = \beta_1(z) = \beta_1(f_1(u)) = f_2(\alpha_1(u))$, so $y = \alpha_1(u)$ since f_2 is 1-1. Hence, $x = \alpha_2(y) = \alpha_2(\alpha_1(u)) = 0$ since $\alpha_2\alpha_1 = 0$ by the exactness of the top row.
(2) Assume f_5 is 1-1 and f_2 and f_4 are onto. Let $x \in B_3$. We must find $w \in A_3$ with $f_3(w) = x$. Since f_4 is onto, we can choose $y \in A_4$ with $f_4(y) = \beta_3(x)$. Then $f_5(\alpha_4(y)) = \beta_4(f_4(y)) = \beta_4(\beta_3(x)) = 0$ from the bottom row is exact. But f_5 is 1-1, so $\alpha_4(y) = 0$. Since the top row is exact, $y = \alpha_3(z)$ for some $z \in A_3$. Then $\beta_3(x) = f_4(y) = f_4(\alpha_3(z)) = \beta_3(f_3(z))$, so $\beta_3(x - f_3(z)) = 0$. Thus, there is a $u \in B_2$ with $\beta_2(u) = x - f_3(z)$ from the bottom row is exact. Since f_2 is onto, there is a $v \in A_2$ with $f_2(v) = u$. Hence, $x - f_3(z) = \beta_2(u) = \beta_2(f_2(v)) = f_3(\alpha_2(v))$, so $x = f_3(z + \alpha_2(v)) = f_3(w)$ where $w = z + \alpha_2(v)$. That is, f_3 is onto. □

Theorem 4.1.6. [Split Exact Sequence] *Let $0 \longrightarrow L \xrightarrow{\alpha} M \xrightarrow{\beta} N \longrightarrow 0$ be a short exact sequence of R-modules. Then the following three conditions are equivalent.*
(i) There exists an isomorphism $M \cong L \oplus N$ in which $\alpha : l \mapsto (l, 0)$ and $\beta : (l, n) \mapsto n$.
(ii) There exists a **section** *of β, that is, a homomorphism $s : N \to M$ such that $\beta \circ s = \mathrm{id}_N$.*
(iiii) There exists a **retraction** *of α, that is, a homomorphism $r : M \to L$ such that $r \circ \alpha = \mathrm{id}_L$.*
If this happens, the sequence is a **split exact sequence**.

Proof. (i) ⇒ (ii) or (iii) is easy.
(ii) ⇒ (i). The given section s is clearly injective because it has a left inverse; we claim that $M = \alpha(L) \oplus s(N)$. To see this, any $m \in M$ is of the form

$$m = (m - s(\beta(m))) + s(\beta(m)),$$

where the second term is obviously in $s(N)$; since $\beta \circ s = \mathrm{id}_N$, the first term is clearly in $\ker \beta$, and by exactness this is $\alpha(L)$. Furthermore, $\alpha(L) \cap s(N) = \{0\}$, since if $n \in N$ is such that $s(n) \in \alpha(L) = \ker \beta$ then $n = \beta(s(n)) = 0$.
(iii) ⇒ (i) is similar to (ii) ⇒ (i) and left as an exercise. □

For finite dimensional vector spaces over a field, every subspace has complement, so every short exact sequence splits. Whether an exact sequence splits or not depends on what ring it is considered over. For example,

$$0 \longrightarrow k[x]x \longrightarrow k[x] \longrightarrow k[x]/k[x]x \longrightarrow 0$$

is split over k but not over $k[x]$.

Proof. It is easy to see that $k[x] \cong k \oplus k[x]x$ as k-vector spaces. Note that k is a $k[x]$-module, where the scalar multiplication is given by $(a_0+a_1x+\cdots+a_nx^n)c = a_0c$ for all $c, a_i \in k$. Assume that $k[x] \overset{\varphi}{\cong} k\oplus k[x]x$ as $k[x]$-modules and $\varphi : 1 \mapsto (a_0, a_1x + \cdots + a_nx^n)$. Then for all $m \in \mathbb{N} \cup \{0\}$ and $b_0, b_1, \ldots, b_m \in k$,

$$(b_0 + b_1x + \cdots + b_mx^m) \overset{\varphi}{\mapsto} (b_0a_0, (b_0 + b_1x + \cdots + b_mx^m)(a_1x + \cdots + a_nx^n)).$$

Since φ is 1-1, $a_0 \neq 0$. Since φ is onto, $a_2, \ldots, a_n = 0$ and $a_1 \neq 0$. Thus, we reduce the above map to

$$(b_0 + b_1x + \cdots + b_mx^m) \overset{\varphi}{\mapsto} (b_0a_0, a_1(b_0x + b_1x^2 + \cdots + b_mx^{m+1})).$$

Consider $(0, x)$ in $k + k[x]x$. If $\varphi(b_0 + b_1x + \cdots + b_mx^m) = (0, x)$, then $b_0a_0 = 0$ and $a_1b_0 = 1$ which is impossible because a_0 and a_1 are nonzero. Hence, φ is not onto which is a contradiction. □

Exercises 4.1.

1. Let M_1 and M_2 be R-submodules of an R-module M. Define $\phi : M_1 \times M_2 \to M_1 + M_2$ by $\phi(m_1, m_2) = m_1 + m_2$. Prove that ϕ is an isomorphism if and only if $M_1 \cap M_2 = \{0\}$.
   Let R be a ring, I an ideal of R and M an R-module. Prove that
   $$IM := \left\{ \sum_{i=1}^{n} r_ix_i : n \geq 1, r_i \in I, x_i \in M \right\}$$
   is an R-submodule of M and M/IM is an R/I-module by the scalar multiplication defined by $(r + I)(x + IM) := rx + IM$.
2. Complete the proof of Theorem 4.1.6.
3. (a) If $\phi : M \to M$ be an R-module homomorphism such that $\phi \circ \phi = \phi$, prove that $M = \ker \phi \oplus \operatorname{im} \phi$.
   (b) If $\alpha : M \to N$ and $\beta : N \to M$ are R-module homomorphisms such that $\beta \circ \alpha = \mathrm{id}_M$, prove that $N = \operatorname{im} \alpha \oplus \ker \beta$.

4.2 Free Modules and Matrices

Like a finite dimensional vector space over a field, we shall see in this section that a free module (i.e., a module with basis) over a commutative ring behaves in a similar way.

Let $M_1, \ldots, M_k$ be R-modules. The **direct sum** of $M_1, \ldots, M_k$ is the set of k-tuples

$$\{(m_1, \ldots, m_k) : m_i \in M_i\}$$

with the following operations:

$$(m_1, \ldots, m_k) + (n_1, \ldots, n_k) = (m_1 + n_1, \ldots, m_k + n_k)$$
$$r(m_1, \ldots, m_k) = (rm_1, \ldots, rm_k), r \in R.$$

The direct sum of $M_1, \ldots, M_k$ is denoted by $M_1 \oplus \cdots \oplus M_k$ or $\bigoplus_{i=1}^{k} M_i$.

Let $M_1, \ldots, M_k$ be submodules of an R-module M. The **sum** of $M_1, \ldots, M_k$ is the set

$$\{m_1 + \cdots + m_k : m_i \in M_i \text{ for all } i\},$$

denoted by $M_1 + \cdots + M_k$ or $\sum_{i=1}^{n} M_i$. It is a submodule of M. We say that $M_1, \ldots, M_k$ are **independent** if for any $m_i \in M_i$ with $m_1 + \cdots + m_k = 0$, we have $m_1 = \cdots = m_k = 0_M$. This condition is equivalent to $M_i \cap \sum_{j \neq i} M_i = \{0_M\}$ for all i.

Theorem 4.2.1. *Let $M_1, \ldots, M_k$ be submodules of an R-module M. Then*

1. *The map $\phi : M_1 \oplus \cdots \oplus M_k \to M$ defined by $\phi(m_1, \ldots, m_k) = m_1 + \cdots + m_k$ is an R-module homomorphism whose image is $M_1 + \cdots + M_k$.*
2. *ϕ is one-to-one if and only if $M_1, \ldots, M_k$ are independent submodules of M. In case ϕ is an isomorphism, we say that M is the* **internal direct sum** *of the submodules $M_1, \ldots, M_k$.*

Let M be an R-module and X a subset of M. The submodule of M **generated by** X, denoted by RX, is the set of all finite sums

$$\{r_1x_1 + \cdots + r_kx_k : r_i \in R \text{ and } x_i \in X\}.$$

If $RX = M$, we say that X **generates or spans** M. If some finite subset $\{x_1, \ldots, x_k\}$ of M generates M, we say that M is **finitely generated** M and we write

$$M = Rx_1 + \cdots + Rx_k.$$

If M is generated by a single element, i.e., if $M = Rx$ for some $x \in M$, M is said to be **cyclic**.

We say that $x_1, \ldots, x_k \in M$ are **linearly independent over** R if for any $r_1, \ldots, r_k \in R$ with $r_1x_1 + \cdots + r_kx_k = 0_M$, we have $r_1 = \cdots = r_k = 0$. A subset X (possibly infinite) of M is **linearly independent** if every finite subset of X is linearly independent. We say that a set X is **linearly dependent** if it is not linearly independent.

Remarks.

1. By convention, the empty set is linearly independent and $R\emptyset = \{0_M\}$.
2. If $x \in M$, $\{x\}$ is a linearly independent set if and only if $Rx \cong R$ as left R-modules. In particular, if we take $M = R$, a left R-module, then $\{x\}$ is a linearly independent set (where $x \in R$) if and only if x is not a right zero divisor, i.e., $a \neq 0 \Rightarrow ax \neq 0$.
3. If $\{x_1, \ldots, x_k\}$ is a linearly independent set, then $Rx_1 + \cdots + Rx_k \cong \underbrace{R \oplus \cdots \oplus R}_{k}$ as left R-modules.
4. Any subset of a linearly independent set is a linearly independent set.

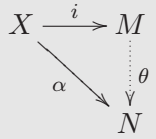
If an R-module M is generated by a linearly independent set X, we say that M is the **free R-module** on the set X and that X is a **basis** for M. If $X = \{x_1, \ldots, x_k\}$ is a finite set, we say that M is the **finitely generated free module spanned by** $x_1, \ldots, x_k$.

Now let $M = Rx_1 + \cdots + Rx_n$ be the free R-module on the set $X = \{x_1, \ldots, x_n\}$. Suppose N is *any* left R-module and $y_1, \ldots, y_n$ are *any* elements of N. Let us define a map $\phi : M \to N$ by

$$\phi(r_1x_1 + \cdots + r_nx_n) = r_1y_1 + \cdots + r_ny_n.$$

Then ϕ is a homomorphism of left R-modules such that $\phi(x_i) = y_i$ for all i. In fact, we could also define a homomorphism even if X were infinite. The point is that any set map $X \to N$ gives rise to an R-module homomorphism $M \to N$. More precisely,

Theorem 4.2.2. [Universal Mapping Property of a Free Module] *Let R be a ring, X a set and $M = M(X)$ the free R-module on the set X. Let $i : X \to M$ be defined by $i(x) = 1 \cdot x$ for all $x \in X$. (i may be thought of as an inclusion map.) Suppose N is an R-module and $\alpha : X \to N$ is a set map. Then there exists a unique R-module homomorphism $\theta : M \to N$ such that $\theta \circ i = \alpha$.*



Hence, any module is a homomorphic image of a free module.

Next, let us consider homomorphism of finitely generated free R-modules. Suppose M and N are free R-modules with bases $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ where

$$M = Rx_1 + \cdots + Rx_m \text{ and } N = Ry_1 + \cdots + Ry_n.$$

Let $\phi : M \to N$ be an R-module homomorphism. Then ϕ will be completely defined as soon as we specify $\phi(x_1), \ldots, \phi(x_m)$. Moreover, by the above theorem, any choice of $\phi(x_1), \ldots, \phi(x_m)$ is possible. Hence,

$$\phi \leftrightarrow (\phi(x_1), \ldots, \phi(x_m))$$

is a 1-1 correspondence between the set of R-module homomorphisms $\phi : M \to N$ and $N \times \cdots \times N$ (m-copies). We have not written $N \oplus \cdots \oplus N$ because so far this correspondence is only a 1-1 correspondence of sets. We do not know if any structure is preserved.

Let M and N be R-modules. The set of R-module homomorphisms from M to N is denoted by $\hom_R(M, N)$.

Remarks. 1. $\hom_R(M, N)$ is an abelian group with the addition given by

$$(\phi + \theta)(m) = \phi(m) + \theta(m).$$

2. If R is *commutative*, then we can make $\hom_R(M, N)$ into a left R-module by defining $(r\phi)(m) = r\phi(m)$. Note that $r\phi : M \to N$ is really an R-module homomorphism, for if $m \in M, s \in R$, then

$$(r\phi)(sm) = r(\phi(sm)) = r(s\phi(m)) = (rs)\phi(m) = (sr)\phi(m) = s(r\phi(m)) = s[(r\phi)(m)].$$

However, this computation makes it clear that the commutativity of R is essential. If R is not commutative, there is no natural way to make $\hom_R(M, N)$ into a left R-module.

Let us restate the remarks above in the next theorem.

Theorem 4.2.3. *Let M and N be left R-modules.*

1. *$\hom_R(M, N)$ is an abelian group (or $\mathbb{Z}$-module) with addition $(\phi + \theta)(m) = \phi(m) + \theta(m)$.*
2. *If R is commutative, $\hom_R(M, N)$ is a left R-module, where $(r\phi)(m) = r\phi(m)$.*
3. *If $M = Rx_1 + \ldots Rx_m$ is the free R-module with basis $x_1, \ldots, x_m$, then*

$$\begin{aligned} \hom_R(M, N) &\longrightarrow N \oplus \cdots \oplus N \\ \phi &\longmapsto (\phi(x_1), \ldots, \phi(x_m)) \end{aligned}$$

is an isomorphism of abelian groups. If R is commutative, it is an isomorphism of R-modules.

For $k \geq 1$ and a ring R, let R^k denote the R-module of $k \times 1$ column vectors over R. Now let us return to free R-modules $M = Rx_1 + \cdots + Rx_m$ and $N = Ry_1 + \cdots + Ry_n$. As noted earlier, if $\phi : M \to N$ is an R-module homomorphism, then ϕ is completely determined by $\phi(x_1), \ldots, \phi(x_m)$, and $\phi \mapsto (\phi(x_1), \ldots, \phi(x_m))$ is an isomorphism of abelian groups, and it is an isomorphism of R-modules if R is commutative. Since $N = Ry_1 + \cdots + Ry_n$ is free on $y_1, \ldots, y_n$ every element of N can be uniquely expressed in the form

$$y = r_1 y_1 + \ldots r_n y_n.$$

In particular, we can write

$$\begin{aligned} \phi(x_1) &= a_{11}y_1 + a_{21}y_2 + \cdots + a_{n1}y_n \\ \phi(x_2) &= a_{12}y_1 + a_{22}y_2 + \cdots + a_{n2}y_n \\ &\vdots \\ \phi(x_m) &= a_{1m}y_1 + a_{2m}y_2 + \cdots + a_{nm}y_n. \end{aligned}$$

In this way, we have abelian group isomorphisms

$$\begin{array}{ccccc} \hom_R(M,N) & \longrightarrow & N\oplus\cdots\oplus N & \longrightarrow & n\times m \text{ matrices over } R \\ \phi & \longmapsto & (\phi(x_1),\ldots,\phi(x_m)) & \longmapsto & \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix}. \end{array}$$

Moreover, in case R is commutative, this is an isomorphism of left R-modules.

Next, let R be a commutative ring and $M = Rx_1 + \cdots + Rx_m$, $N = Ry_1 + \cdots + Ry_n$ and $P = Rz_1 + \cdots + Rz_p$ be finitely generated free modules over R with the indicated free generators. Let $\alpha : M \to R^m$, $\beta : N \to R^n$ and $\gamma : P \to R^p$ be the R-module isomorphisms

$$\alpha(r_1x_1 + \cdots + r_mx_m) = \begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix}, \beta(s_1y_1 + \cdots + s_ny_n) = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}, \gamma(t_1z_1 + \cdots + t_pz_p) = \begin{bmatrix} t_1 \\ \vdots \\ t_p \end{bmatrix}.$$

Write R_{uv} for the R-module of $u \times v$ matrices over R. For each R-module homomorphism $\phi : M \to N$, we define

$$[\phi] = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix} \in R_{nm}$$

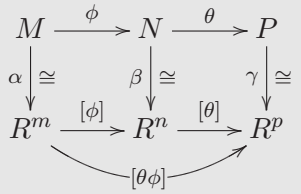
implicitly from the equations

$$\begin{aligned} \phi(x_1) &= a_{11}y_1 + a_{21}y_2 + \cdots + a_{n1}y_n \\ \phi(x_2) &= a_{12}y_1 + a_{22}y_2 + \cdots + a_{n2}y_n \\ &\vdots \\ \phi(x_m) &= a_{1m}y_1 + a_{2m}y_2 + \cdots + a_{nm}y_n. \end{aligned}$$

Similarly, for R-module homomorphisms $\theta : N \to P$ and $\tau : M \to P$ define $[\theta] \in R_{pn}$ and $[\tau] \in R_{pm}$, respectively. Then $\phi \mapsto [\phi]$, $\theta \mapsto [\theta]$ and $\tau \mapsto [\tau]$ are isomorphisms of R- modules $\hom_R(M,N) \cong R_{nm}$, $\hom_R(N,P) \cong R_{pn}$ and $\hom_R(M,P) \cong R_{pm}$, respectively. Moreover, we obtain the following theorem.

Theorem 4.2.4. *Let R be a commutative ring. Under the above set-up we have:*

1. *Each matrix $[\phi] \in R_{nm}$ defines a homomorphism $[\phi] : R^m \to R^n$ by left multiplication of an $n \times m$ matrix by an $m \times 1$ matrix. The same is true for $[\theta] : R^n \to R^p$ and $[\tau] : R^m \to R^p$.*
2. *The following diagram is commutative*



In particular, $[\theta][\phi] = [\theta\phi]$ where the left product is multiplication of matrices.

Recall the following fact about matrices: Let R be a commutative ring and suppose $A \in M_n(R)$. Then

$$A \text{ is invertible} \Leftrightarrow A \text{ is left invertible} \Leftrightarrow A \text{ is right invertible} \Leftrightarrow \det A \text{ is a unit in } R.$$

In particular, if $AC = I$, then $CA = I$. Moreover, we have

Theorem 4.2.5. *Let R be a commutative ring and let $[\phi] \in R_{mn}$ and $[\theta] \in R_{nm}$. Suppose $[\phi][\theta] = I_m$ and $[\theta][\phi] = I_n$ are identity matrices of sizes $m \times m$ and $n \times n$, respectively. Then $m = n$.*

Proof. Assume that $m > n$. Then $m = n + r$ for some $r \in \mathbb{N}$. Write

$$[\phi] = \begin{bmatrix} A_{n\times n} \\ B_{r\times n} \end{bmatrix} \quad \text{and} \quad [\theta] = \begin{bmatrix} C_{n\times n} & D_{n\times r} \end{bmatrix},$$

so

$$[\phi][\theta] = \begin{bmatrix} AC & AD \\ BC & BD \end{bmatrix} = \begin{bmatrix} I_n & \mathbf{0} \\ \mathbf{0} & I_r \end{bmatrix}.$$

Thus, $\mathbf{0} = C(AD) = (CA)D = I_nD = D$ which contradicts $BD = I_r$. Hence, $m \leq n$. Similarly, we obtain a contradiction if $m < n$. Therefore, $m = n$. □

Theorem 4.2.6. *Let R be a commutative ring and suppose that $M = Rx_1 + \cdots + Rx_m$ and $N = Ry_1 + \cdots + Ry_n$ are free R-modules with indicated generators. If M and N are isomorphic R-modules, then $m = n$.*

Proof. Let $\phi : M \to N$ be an isomorphism with inverse $\theta : N \to M$. By Theorem 4.2.4, we can identify M with $m \times 1$ column vectors and N with $n \times 1$ column vectors and obtain a commutative diagram

$$\begin{array}{ccccc} M & \xrightarrow{\phi} & N & \xrightarrow{\theta} & M \\ \downarrow\cong & & \downarrow\cong & & \downarrow\cong \\ R^m & \xrightarrow{[\phi]} & R^n & \xrightarrow{[\theta]} & R^m \end{array} \qquad [\theta][\phi] : R^m \to R^m$$

In other words, $[\phi]$ is an $n \times m$ matrix and $[\theta]$ is an $m \times n$ matrix with $[\phi][\theta] = I_n$ and $[\theta][\phi] = I_m$. Hence, $m = n$ by Theorem 4.2.5. □

If $M = Rx_1 + \cdots + Rx_m$ is a free R-module on $x_1, \ldots, x_m$ over a *commutative* ring R, m is called the **rank** of M. In particular, if $R = \mathbb{Z}$, then M is a free abelian group on $x_1, \ldots, x_m$, and hence we have shown:

Corollary 4.2.7. *If F is a finitely generated free abelian group, then any two bases of F have the same number of elements.*

Using this corollary, we can verify that if a group G is free on A and also on B, which are finite sets, then the sets A and B have the same number of elements. It is Theorem 3.5.4 for the finite basis case.

Proof of Theorem 3.5.4 for the finite basis case. Assume that G is a free group on A and also on B, where A and B are finite sets. By Exercise 1, G/G' is a free abelian group of rank $|A|$ and $|B|$, respectively. By Corollary 4.2.7, $|A| = |B|$. □

Remarks.

1. As we have seen that subgroups of a free (abelian) group are free. This is not true for general R-modules. For example, let $R = \mathbb{Z}_6$. Then ${}_RR$ is a free R-module generated by $\{1\}$. $N = \{0, 2, 4\}$ is an R-submodule of ${}_RR$. Since $\emptyset$ does not span N, $\emptyset$ is not a basis. If $B \neq \emptyset$ is a basis of N, then $0 \notin B$, so 2 or 4 are in B. Since $3 \cdot 2 = 0$ and $3 \cdot 4 = 0$ where $3 \neq 0$, where B is not linearly independent. *Hence, submodules of a free module may not be free.*
2. In the case of free abelian groups and vector spaces, it is true that any two bases of have the same cardinality. This is not true in general as shown in the following example.

Example 4.2.1. Let S be a ring and F a free S-module with infinite denumerable basis $\{e_1, e_2, e_3, \dots\}$. Let $R = \hom_S(F, F)$. Then R is a ring with identity 1_R, so $\{1_R\}$ is a basis for ${}_RR$. Next, we define $f_1, f_2 \in R$ as follows: $f_1(e_{2n}) = e_n, f_1(e_{2n-1}) = 0$ and $f_2(e_{2n}) = 0, f_2(e_{2n-1}) = e_n$. To show that $\{f_1, f_2\}$ spans ${}_RR$, let $g \in R$. Define $g_1, g_2 \in R$ by $g_1(e_n) = g(e_{2n})$ and $g_2(e_n) = g(e_{2n-1})$. Then $(g_1f_1 + g_2f_2)(e_{2n-1}) = g_1f_1(e_{2n-1}) + g_2f_2(e_{2n-1}) = g_2(e_n) = g(e_{2n-1})$ and $(g_1f_1 + g_2f_2)(e_{2n}) = g_1f_1(e_{2n}) + g_2f_2(e_{2n}) = g_1(e_n) = g(e_{2n})$. Thus, $g = g_1f_1 + g_2f_2$. Next we shall prove that $\{f_1, f_2\}$ is linearly independent over R. Let $h_1, h_2 \in R$ such that $h_1f_1 + h_2f_2 = 0$. Then for any $n \geq 1$, $h_1(e_n) = h_1(e_n) + 0 = h_1f_1(e_{2n}) + h_2f_2(e_{2n}) = (h_1f_1 + h_2f_2)(e_{2n}) = 0$ and $h_2(e_n) = 0 + h_2(e_n) = h_1f_1(e_{2n-1}) + h_2f_2(e_{2n-1}) = (h_1f_1 + h_2f_2)(e_{2n-1}) = 0$, so $h_1 = h_2 = 0$. Hence, $\{f_1, f_2\}$ is linearly independent and so it is a basis of ${}_RR$.

Exercises 4.2.

1. Show that $\mathbb{Q}$ is not a free $\mathbb{Z}$-module.
2. Show that M is a cyclic left R-module if and only if it is isomorphic to R/I (considered as a left R-module) for some left ideal I of R.
3. Show that $\{e_i\}_{i\in I}$ is a basis of a left R-module M if and only if $(r_i)_{i\in I} \mapsto \sum_{i\in I} r_ie_i$ is an isomorphism of $\bigoplus_{i\in I} {}_RR$ onto M.
4. Prove that the module ${}_RR$ in Example 4.2.1 has a basis with m elements for every positive integers m.
5. Let R be a ring and M, N and N' R-modules. Then $\hom_R(M, N)$ and $\hom_R(M, N')$ are $\mathbb{Z}$-modules. For an R-module homomorphism $f : N \to N'$, we define $\hom(M, -)(f) : \hom_R(M, N) \to \hom_R(M, N')$ by
$$\hom(M, -)(f)(h) = f \circ h$$
for all $h \in \hom_R(M, N)$. Show that
(a) $\hom(M, -)(f)$ is a $\mathbb{Z}$-module homomorphism from $\hom_R(M, N)$ to $\hom_R(M, N')$.
(b) If $0 \to N \xrightarrow{f} N' \xrightarrow{g} N''$ is exact, then
$$0 \longrightarrow \hom_R(M, N) \xrightarrow{\hom(M,-)(f)} \hom_R(M, N') \xrightarrow{\hom(M,-)(g)} \hom_R(M, N'')$$
is exact.
In a similar manner, one can prove that exactness of $N \xrightarrow{f} N' \xrightarrow{g} N'' \to 0$ implies exactness of
$$0 \longrightarrow \hom_R(N'', M) \xrightarrow{\hom(-,M)(g)} \hom_R(N', M) \xrightarrow{\hom(-,M)(f)} \hom_R(N, M)$$
where $\hom(-, M)(f)(h) = h \circ f$ for all $h \in \hom_R(N', M)$ and $\hom(-, M)(g)(h) = h \circ g$ for all $h \in \hom_R(N'', M)$.
6. Let R be a ring, I a proper ideal of R and F a free R-module with a basis X. Then F/IF is a free R/I-module with a basis of cardinality $|X|$.

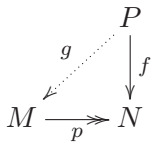
4.3 Projective and Injective Modules

The concept of projective modules is a generalization of the idea of a free module. Injective modules, introduced by Baer, are dual to that of projective modules. We follow [6] for this section.

Let R be a ring. An R-module P is called **projective** if given any diagram

$$\begin{array}{ccc} & & P \\ & & \downarrow f \\ M & \xrightarrow[p]{} & N \end{array}$$

there exists a homomorphism $g : P \to M$ such that



is commutative. In other words, given an epimorphism $p : M \to N$, then any homomorphism $f : P \to N$ can be factored as $f = pg$ for some $g : P \to M$.

We recall that for any module M, if $0 \longrightarrow N' \xrightarrow{i} N \xrightarrow{p} N'' \longrightarrow 0$ is exact, then

$$0 \longrightarrow \hom_R(M, N') \xrightarrow{\hom(M,i)} \hom_R(M, N) \xrightarrow{\hom(M,p)} \hom_R(M, N'')$$

is exact. Now suppose $M = P$ is projective. Then given $f \in \hom(P, N'')$ there exists a $g \in \hom_R(P, N)$ such that $\hom_R(P, p)(g) = pg = f$. Thus, in this case, $\hom(P, p)$ is surjective and so we actually have the exactness of

$$0 \longrightarrow \hom_R(M, N') \xrightarrow{\hom(M,i)} \hom_R(M, N) \xrightarrow{\hom(M,p)} \hom_R(M, N'') \longrightarrow 0$$

as a consequence of the exactness of $0 \longrightarrow N' \xrightarrow{i} N \xrightarrow{p} N'' \longrightarrow 0$.

The converse holds also. Suppose $\hom(P, -)$ is exact and suppose $M \xrightarrow{p} N$. Let $K = \ker p$. Then we have the exact sequence $0 \longrightarrow K \xrightarrow{i} M \xrightarrow{p} N \longrightarrow 0$ where i is the inclusion map. Applying the exactness of $\hom(P, -)$, we obtain the property of a projective module. Therefore,

Theorem 4.3.1. *Let P be an R-module. Then P is projective if and only if for any R-modules N, N' and N'', if $0 \longrightarrow N' \xrightarrow{i} N \xrightarrow{p} N'' \longrightarrow 0$ is a short exact sequence, then*

$$0 \longrightarrow \hom_R(P, N') \xrightarrow{\hom(P,i)} \hom_R(P, N) \xrightarrow{\hom(P,p)} \hom_R(P, N'') \longrightarrow 0$$

is also a short exact sequence of $\mathbb{Z}$-modules.

By Theorem 4.2.2, we have:

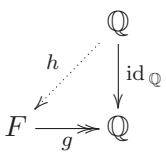
Theorem 4.3.2. *Every free module is projective.*

Example 4.3.1. $\mathbb{Q}$ is not a projective $\mathbb{Z}$-module.

Proof. Let F be a free $\mathbb{Z}$-module with countable basis $X = \{x_1, x_2, \dots\}$. Define $g : X \to \mathbb{Q}$ by

$$g : x_n \mapsto \frac{1}{n} \quad \text{for all } n \in \mathbb{N}.$$

Then g induces a $\mathbb{Z}$-module homomorphism from F to $\mathbb{Q}$. Since $g(mx_n) = \frac{m}{n}$ for all $m \in \mathbb{Z}$ and $n \in \mathbb{N}$, g is onto. Assume that $\mathbb{Q}$ is projective.



Then there exists an $h : \mathbb{Q} \to F$ such that $gh = \mathrm{id}_{\mathbb{Q}}$. Suppose $h(1) = \sum_i a_i x_i$ (with all but finite $a_i = 0$). Let $k = 1 + \prod_{i, a_i \neq 0} |a_i|$ and assume that $h(k^{-1}) = \sum_i b_i x_i$ (again, with all but finite $a_i = 0$). Then

$$\sum_i k b_i x_i = k \sum_i b_i x_i = k h(k^{-1}) = h(1) = \sum_i a_i x_i,$$

so $\sum_i (a_i - k b_i) x_i = 0$. Since X is linearly independent, $a_i = k b_i$ for all i which implies $k \mid a_i$ for all i. This forces $k = 1$ and $a_i = 0$ for all i. Thus, h is the zero map which contradicts $gh = \mathrm{id}_{\mathbb{Q}}$. Hence, $\mathbb{Q}$ is not projective. □

How close are projective modules to being free? We shall give two important characterizations of projective modules as follows.

Theorem 4.3.3. *The following properties of a module P are equivalent:*
(i) P is projective.
(ii) Any short exact sequence $0 \longrightarrow M \longrightarrow N \longrightarrow P \longrightarrow 0$ splits.
(iii) P is a direct summand of a free module (that is, there exists a free module F isomorphic to $P \oplus P'$ for some P').

Proof. (i) ⇒ (ii). Let $0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$ be exact and consider the diagram

$$\begin{array}{ccc} & & P \\ & & \downarrow{\scriptstyle \mathrm{id}_P} \\ N & \xrightarrow[g]{} & P \end{array}$$

By hypothesis we can fill this in with $g' : P \to N$ to obtain a commutative diagram. Then $gg' = \mathrm{id}_P$ and the given short exact sequence splits.
(ii) ⇒ (iii). Since any module is a homomorphic image of a free module (Theorem 4.2.2), we have a short exact sequence $0 \longrightarrow P' \xrightarrow{i} F \xrightarrow{p} P \longrightarrow 0$ where F is a free module. If P satisfies property (ii), then this exact sequence splits and hence $F \cong P \oplus P'$.
(iii) ⇒ (i). We are given that there exists a sequence $0 \longrightarrow P' \xrightarrow{i} F \xrightarrow{p} P \longrightarrow 0$ with F is free. Now suppose we have a diagram

$$\begin{array}{ccc} & & P \\ & & \downarrow{\scriptstyle f} \\ M & \xrightarrow[q]{} & N \end{array}$$

Combining the two diagrams, we obtain

$$\begin{array}{ccccccccc} 0 & \longrightarrow & P' & \xrightarrow{i} & F & \underset{i'}{\overset{p}{\rightleftarrows}} & P & \longrightarrow & 0 \\ & & & & & \searrow^{fp} & \downarrow{\scriptstyle f} & & \\ & & & & M & \xrightarrow[q]{} & N & & \end{array}$$

where $pi' = \mathrm{id}_P$ (since the top line splits). Since F is free, hence projective, we can fill in $g : F \to M$ to obtain $fp = qg$. Then $f = f \mathrm{id}_P = fpi' = qgi'$ and $gi' : P \to M$ make

$$\begin{array}{ccc} & & P \\ & \swarrow^{gi'} & \downarrow{\scriptstyle f} \\ M & \xrightarrow[q]{} & N \end{array}$$

commutative. Hence, P is projective. □

Of particular interest are the modules that are finitely generated and projective. The theorem gives the following characterization of these modules.

Corollary 4.3.4. *A module P is finitely generated and projective if and only if P is a direct summand of a free module with a finite base.*

Proof. If P is a direct summand of a free module F with finite base, then P is projective. Moreover, P is a homomorphic image of F, so P has a finite set of generators (the images of the base under an epimorphism of F onto P). Conversely, suppose P is finitely generated and projective. Then the first condition implies that we have an exact sequence $0 \to P' \to F \to P \to 0$ where F is free with finite base. The proof of the theorem shows that if P is projective, then $F \cong P \oplus P'$, so P is a direct summand of a free module with finite base. □

The concept of a projective module has a dual obtained by reversing the arrows in the definition as follows.

An R-module Q is called **injective** if given any diagram of homomorphisms

$$\begin{array}{ccccc} 0 & \longrightarrow & N & \xrightarrow{\ i\ } & M \\ & & \Big\downarrow{\scriptstyle f} & & \\ & & Q & & \end{array}$$

there exists a homomorphism $g : M \to Q$ such that the diagram obtained by filling in g is commutative. In other words, given $f : N \to Q$ and a monomorphism $i : N \to M$ there exists a $g : M \to Q$ such that $f = gi$.

With a slight change of notation, the definition amounts to this: Given an exact sequence $0 \to N' \xrightarrow{i} N$, the sequence

$$\hom_R(N, Q) \xrightarrow{\hom(i,Q)} \hom_R(N', Q) \longrightarrow 0$$

is exact. Since we know that exactness of $0 \to N' \xrightarrow{i} N \xrightarrow{p} N'' \to 0$ implies exactness of

$$0 \longrightarrow \hom_R(N'', M) \xrightarrow{\hom(p,M)} \hom_R(N, M) \xrightarrow{\hom(i,M)} \hom_R(N', M)\ ,$$

it is clear that Q is injective if and only if $\hom(-, Q)$ is exact in the sense that it maps any short exact sequence $0 \to N' \to N \to N'' \to 0$ into a short exact sequence of $\mathbb{Z}$-module

$$0 \to \hom_R(N'', Q) \to \hom_R(N, Q) \to \hom_R(N', Q) \to 0.$$

It is easily seen also that the definition of injective is equivalent to the following: If N is a submodule of a module M, then any homomorphism of N into Q can be extended to a homomorphism of M into Q. Another result, which is easily established by dualizing the proof of the analogous result on projective (Theorem 4.3.3), is that if Q is injective, then any short exact sequence $0 \to Q \to M \to N \to 0$ splits. The converse of this holds also. However, the proof requires the dual of the easy result that any module is a homomorphic image of a projective module (in fact, a free module). The dual statement is that any module can be embedded in an injective one. We shall see that this is the case, but the proof will turn out to be fairly difficult.

Theorem 4.3.5. [Baer] *A right module Q is injective if and only if any homomorphism of a right ideal I of R into Q can be extended to a homomorphism of R into Q.*

Proof. Obviously, the condition is necessary. Now suppose it holds and suppose M is a module and f is a homomorphism of a submodule N of M into Q. Consider the set $\{(g, M')\}$ where M' is a submodule of M containing N and g is a homomorphism of M' into Q such that $g|N = f$. We define a partial order in the set $\{(g, M')\}$ by declaring that $(g_1, M_1') \geq (g_2, M_2')$ if $M_1' \supset M_2'$ and $g_1|M_2' = g_2$. It is clear that any totally ordered subset has an upper bound in this set. Hence, by Zorn's lemma, there exists a maximal (g, M'); that is, we have an extension of f to a homomorphism g of $M' \supset N$ which is maximal in the sense that if g_1 is a homomorphism of an $M_1' \supset M'$ such that $g_1|M' = g$, then necessarily $M_1' = M'$. We claim that $M' = M$. Otherwise, there is an $x \in M, \notin M'$ and so $xR + M'$ is a submodule of M properly containing M'. Now let

$$I = \{s \in R : xs \in M'\}.$$

Then $I = \text{ann}(x + M')$ in M/M', so I is a right ideal of R. If $s \in I$, then $xs \in M'$, so $g(xs) \in Q$. It is immediate that the map $h : s \mapsto g(xs)$ is a module homomorphism of I into Q. Hence, by hypothesis, h can be extended to a homomorphism k of R into Q. We shall use this to obtain an extension of g to a homomorphism of $xR + M'$ to Q. The elements of $xR + M'$ have the form $xr + y, r \in R, y \in M'$. If we have a relation $xs + y' = 0, s \in R, y' \in M'$, then $s \in I$. Then

$$k(s) = h(s) = g(xs) = -g(y').$$

Thus, $xs + y' = 0$ for $s \in R, y' \in M'$, implies that $k(s) + g(y') = 0$. It follows that

$$xr + y \mapsto k(r) + g(y),$$

$r \in R, y \in M'$, is a well defined map. For, if $xr_1 + y_1 = xr_2 + y_2, r_i \in R, y_i \in M'$, then $xs + y' = 0$ for $s = r_1 - r_2$, $y' = y_1 - y_2$. Then $k(s) + g(y') = 0$ and $k(r_1 - r_2) + g(y_1 - y_2) = 0$. Since k and g are homomorphisms, this implies that $k(r_1) + g(y_1) = k(r_2) + g(y_2)$. It is immediate that the map $rx+y \mapsto k(r)+g(y)$ is a module homomorphism of $xR+M'$ into Q extending the homomorphism g of M'. This contradicts the maximality of (g, M'). Hence, $M' = M$ and we have proved that if f is a homomorphism of a submodule N of M into Q, then f can be extended to a homomorphism of M into Q. Hence, Q is injective. □

For certain "nice" rings, the concept of injectivity of modules is closely related to the simpler notion of divisibility, which we proceed to define.

If $a \in R$, then the module M is said to be **divisible by** a if the map $x \mapsto xa$ of M into M is surjective. A module is called **divisible** if it is divisible by every $a \neq 0$. It is clear that if M is divisible by a or if M is divisible, then any homomorphic image of M has the same property. In some sense injectivity is generalization of divisibility, for we have

Theorem 4.3.6. *1. If R has no zero divisors $\neq 0$, then any injective R-module is divisible.*
2. If R is a ring such that every right ideal of R is principal ($= aR$ for some $a \in R$), then any divisible R-module is injective.

Proof. (1) Suppose R has no zero-divisors $\neq 0$ and let Q be an injective R-module. Let $x \in Q, r \in R, r \neq 0$. If $a, b \in R$ and $ra = rb$, then $a = b$. Hence, we have a well defined map $ra \mapsto xa$, $a \in R$, of the right ideal rR into Q. Clearly this is a module homomorphism. Since Q is injective, the map $ra \mapsto xa$ can be extended to a homomorphism of R into Q. If $1 \mapsto y$ under this extension, then $r = 1r \mapsto yr$. Since $r = r1 \mapsto x1 = x$, we have $x = yr$. Since x was arbitrary in Q and r was any non-zero element of R, this shows that Q is divisible.
(2) Suppose R is a ring in which every right ideal is principal. Let M be a divisible R-module and let f be a homomorphism of the right ideal rR into M. If $r = 0$, then f is the zero map and this can be extended to the zero map of R. If $r \neq 0$ and $f(r) = x \in M$, then there exists a y in M such that $x = yr$. Then $a \mapsto ya$ is a module homomorphism of R into M and since

$rb \mapsto yrb = xb = f(r)b = f(rb)$, $a \mapsto ya$ is an extension of f. Thus, any module homomorphism of a right ideal of R into M can be extended to a homomorphism of R. Hence, M is injective by Baer's criterion. □

If R satisfies both conditions stated in the theorem, then an R-module is injective if and only if it is divisible. In particular, this holds if R is a PID. We can use this to construct some examples of injective modules.

Examples 4.3.2.
1. Let R be a subring of a field F and regard F as an R-module in the natural way. Evidently F is a divisible R-module. Hence, if K is any R-submodule of F, then F/K is a divisible R-module. In particular, $\mathbb{Q}$ is an injective $\mathbb{Z}$-module which is not projective.
2. Let D be a PID, F its field of fractions. If $r \in D$, then the D-module F/rD is divisible and hence is injective by Theorem 4.3.6.

Our next objective is to prove that any module can be embedded in an injective module, that is, given any M there exists an exact sequence $0 \to M \xrightarrow{i} Q$ with Q is injective. The first step in the proof we shall give is as follows.

Lemma 4.3.7. *Any abelian group can be embedded in a divisible group (= a divisible $\mathbb{Z}$-module).*

Proof. First let F be a free abelian group with base $\{x_\alpha\}$ and F' the vector space over $\mathbb{Q}$ with $\{x_\alpha\}$ as base. Then F is embedded in F' and it is clear that F' is divisible. Now let M be an arbitrary abelian group. Then M is isomorphic to a factor group F/K of a free abelian group F. Hence, F'/K is a divisible group and $F/K \cong M$ is a subgroup. □

An immediate consequence of this and Theorem 4.3.6 is the next corollary.

Corollary 4.3.8. *Any $\mathbb{Z}$-module can be embedded in an injective $\mathbb{Z}$-module.*

Now for an arbitrary R-module M, we have the isomorphism of M onto $\hom_R(R, M)$ which maps an element $x \in M$ into the homomorphism f_x such that $1 \mapsto x$. This is an R-isomorphism if we make $\hom_R(R, M)$ into a right R-module by defining $fa, a \in R$, by $(fa)(b) = f(ab)$. Also $\hom_{\mathbb{Z}}(R, M)$ is a right R-module using this definition of fa. Clearly $\hom_R(R, M)$ is a submodule of $\hom_{\mathbb{Z}}(R, M)$. Since M is isomorphic to $\hom_R(R, M)$, we have an embedding of M in $\hom_{\mathbb{Z}}(R, M)$. Now embed M in an injective $\mathbb{Z}$-module Q, which can be done by the foregoing corollary. Then we have an embedding of $\hom_{\mathbb{Z}}(R, Q)$ as R-modules. This gives an embedding of M in an injective R-module, since we have the following lemma.

Lemma 4.3.9. *If Q is an injective $\mathbb{Z}$-module, then $\hom_{\mathbb{Z}}(R, Q)$ is an injective R-module.*

Proof. We must show that if $0 \to N' \xrightarrow{f} N$ is an exact sequence of R-modules, then

$$\hom_R(N, \hom_{\mathbb{Z}}(R, Q)) \xrightarrow{f^*} \hom_R(N', \hom_{\mathbb{Z}}(R, Q)) \to 0$$

is exact, where $f^* = \hom_R(f, \hom_{\mathbb{Z}}(R, Q))$. We have an isomorphism

$$\varphi_N : \hom_{\mathbb{Z}}(N \otimes_R R, Q) \to \hom_R(N, \hom_{\mathbb{Z}}(R, Q))$$

and the definition shows that this is "natural" in N. Since the isomorphism of $N \otimes_R R$ onto N such that $y \otimes 1 \mapsto y$ is natural in N, we have an isomorphism

$$\psi_N : \hom_{\mathbb{Z}}(N, Q) \to \hom_R(N, \hom_{\mathbb{Z}}(R, Q))$$

which is natural in N, that is we have the commutativity of

$$\begin{array}{ccc} \hom_{\mathbb{Z}}(N,Q) & \xrightarrow{\psi_N} & \hom_R(N,\hom_{\mathbb{Z}}(R,Q)) \\ \downarrow{\scriptstyle \bar{f}} & & \downarrow{\scriptstyle f^*} \\ \hom_{\mathbb{Z}}(N',Q) & \xrightarrow[\psi_{N'}]{} & \hom_R(N',\hom_{\mathbb{Z}}(R,Q)) \end{array}$$

where $\bar{f} = \hom(f,Q)$. Now $\bar{f}$ is surjective since $\mathbb{Q}$ is $\mathbb{Z}$-injective. Since ψ_N and $\psi_{N'}$ are isomorphisms, this implies that f^* is surjective. □

The foregoing lemma completes the proof of the embedding theorem.

Theorem 4.3.10. *Any module can be embedded in an injective module.*

The proof we have given is due to B. Eckmann and A. Schöpf. We can apply the theorem to complete the following characterization of injectives, which we idicated earlier.

Theorem 4.3.11. *The following properties of a module Q are equivalent:*
(i) Q is injective.
(ii) Any short exact sequence $0 \to Q \to M \to N \to 0$ splits.
(iii) Q is a direct summand of every module containing it as a submodule.

Proof. We leave the proof of (i) ⇒ (ii) as an exercise. Conversely, suppose any short exact sequence $0 \to Q \to M \to N \to 0$ splits. By the embedding theorem we have an exact sequence $0 \to Q \xrightarrow{i} M$ where M is injective. Then we have the short exact sequence $0 \to Q \xrightarrow{i} M \xrightarrow{p} M/Q \to 0$ where p is the canonical homomorphism of M onto M/Q. By hypothesis, we can find a $p' : M \to Q$ such that $p'i = \mathrm{id}_Q$. Now suppose we have a diagram

$$\begin{array}{ccccc} 0 & \longrightarrow & N' & \xrightarrow{j} & N \\ & & \downarrow{\scriptstyle f} & & \\ & & Q & & \end{array}$$

Since M is injective, we can enlarge this to a commutative diagram

$$\begin{array}{ccccc} 0 & \longrightarrow & N' & \xrightarrow{j} & N \\ & & \downarrow{\scriptstyle f} & \searrow^{p'g} & \\ & & Q & \swarrow_{g} & \\ & & {\scriptstyle i}\downarrow\uparrow{\scriptstyle p'} & & \\ & & M & & \end{array}$$

This means that by the injectivity of M we have $g : N \to M$ such that $if = gj$. Then $f = \mathrm{id}_Q f = p'if = (p'g)j$. Hence, Q is injective. □

Exercises 4.3. 1. Let $R = \mathbb{Z}_6$. Define $\mathbb{Z}_6 \times \mathbb{Z}_2 \to \mathbb{Z}_2$ by $[r]_6[x]_2 := [rx]_2$ and $\mathbb{Z}_6 \times \mathbb{Z}_3 \to \mathbb{Z}_3$ by $[r]_6[x]_3 := [rx]_3$. Prove that $\mathbb{Z}_2$ and $\mathbb{Z}_3$ are $\mathbb{Z}_6$-modules and $\mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3$ as $\mathbb{Z}_6$-modules.
$\mathbb{Z}_2$ and $\mathbb{Z}_3$ are not free $\mathbb{Z}_6$-modules. $\mathbb{Z}_6$ is a free $\mathbb{Z}_6$-module. Since $\mathbb{Z}_2$ and $\mathbb{Z}_3$ are direct summands of $\mathbb{Z}_6$-module, they are projective.

2. Show that if e is an idempotent ($e^2 = e$) in a ring R, the eR is a projective right module and Re is an projective left module.
3. Show that
   (a) $\bigoplus_\alpha P_\alpha$ is projective if and only if every P_α is projective.
   (b) $\bigoplus_\alpha Q_\alpha$ is injective if and only if every Q_α is injective.
4. Prove that
   (a) A direct sum of abelian groups is divisible if and only if each summand is divisible.
   (b) A homomorphic image of a divisible module is divisible.
5. Let R be an integral domain that is not a field. If M is an R-module such that M is both injective and projective, prove that $M = \{0\}$.
6. Prove (i) $\Rightarrow$ (ii) in Theorem 4.3.11 by dualizing the proof of Theorem 4.3.3.
7. Consider the polynomial ring $\mathbb{Z}[x]$ as a $\mathbb{Z}$-module.
   (a) Is $\mathbb{Z}[x]$ free? (b) Is $\mathbb{Z}[x]$ projective?
   (c) Is $\mathbb{Z}[x]$ injective? (d) Is $\mathbb{Z}[x]$ divisible?

Project 22 (Injective hull)**.** It is possible to prove a sharper result than Theorem 4.3.10, namely that there is a *minimal* injective R-module H containing M in the sense that any injective map of M into an injective R-module Q factor through H. More precisely, show that if $M \subseteq Q$ for an injective R-module Q then there is an injection $i : H \to Q$ that restricts to the identity map on M; using i to identify H as a subset of Q we have $M \subseteq H \subseteq Q$. This module H is called the **injective hull** or **injective envelope of** M. For example, the injective hull of $\mathbb{Z}$ is $\mathbb{Q}$, and the injective hull of any field is itself. Furthermore, prove that:

(a) The injective hull of an injective module is itself.
(b) The injective hull of an integral domain is its field of fractions.

4.4 Modules over a PID

Our main goal of this section is to prove the structure theorem for modules over a PID.

Theorem 4.4.1. *Let R be a PID and suppose that M is a finitely generated R-module. Then there is an integer $r \geq 0$ and nonzero elements $d_1, \ldots, d_k \in R$ with $d_1 | d_2, \ldots, d_{k-1} | d_k$ such that*

$$M \cong \underbrace{R \oplus \cdots \oplus R}_{r \text{ copies}} \oplus R/Rd_1 \oplus \cdots \oplus R/Rd_k.$$

Moreover, if N is another finitely generated R-module and

$$N \cong \underbrace{R \oplus \cdots \oplus R}_{\bar{r} \text{ copies}} \oplus R/R\bar{d}_1 \oplus \cdots \oplus R/R\bar{d}_{\bar{k}},$$

where $\bar{d}_i | \bar{d}_{i+1}$, then M and N are isomorphic as R-modules if and only if $r = \bar{r}$, $k = \bar{k}$ and d_i and $\bar{d}_i$ are associates for $i = 1, \ldots, k$.

Note that we cannot assert more than that d_i and $\bar{d}_i$ are associates, for if d and $\bar{d}$ are associates, then $R/Rd \cong R/R\bar{d}$.

Since abelian groups are equivalent to $\mathbb{Z}$-modules, this theorem can be stated as "A finitely generated $\mathbb{Z}$-module is a direct sum of cyclic modules". Actually, the theorem was more precise in that it actually classified all finitely generated $\mathbb{Z}$-modules up to isomorphism. That is, one has

Theorem 4.4.2. *Let M be a finitely generated $\mathbb{Z}$-module. Then there are nonnegative integers $r \geq 0$, $d_1, \ldots, d_k > 0$ where $d_1|d_2, \ldots, d_{k-1}|d_k$ such that*

$$M \cong \underbrace{\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{r \text{ copies}} \oplus \mathbb{Z}/d_1\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/d_k\mathbb{Z}.$$

Moreover, if N is another finitely generated $\mathbb{Z}$-module and

$$N \cong \underbrace{\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{\bar{r} \text{ copies}} \oplus \mathbb{Z}/\bar{d}_1\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/\bar{d}_{\bar{k}}\mathbb{Z},$$

where $\bar{d}_i|\bar{d}_{i+1}$, then M and N are isomorphic if and only if $r = \bar{r}$, $k = \bar{k}$ and $d_i = \bar{d}_i$ for $i = 1, \ldots, k$. r is called the **rank** *or* **torsion-free rank of** *M and $d_1, \ldots, d_k$ are called the* **invariant factors of** *M.*

Therefore, we have a major theorem on abelian groups:

Corollary 4.4.3. *A finitely generated abelian group is a direct product of cyclic groups.*

The strategy of our proof is the following. First we observe that even with no hypothesis on the ring R, the following statements are equivalent:

(i) M is a finitely generated R-module and can be generated by s elements.

(ii) Let $F = Rx_1 + \cdots + Rx_s$ be a free R-module with s free generators. Then there is an exact sequence of R-modules $0 \longrightarrow K \longrightarrow F \xrightarrow{\phi} M \longrightarrow 0$ where $K = \ker \phi$.

(iii) $M \cong F/K$ where F is a free R-module on s free generators and K is an R-submodule of F.

Now let us suppose that R is commutative and we have a free R-module F and a submodule K where $F = Rx_1 + \cdots + Rx_r + Ry_1 + \cdots + Ry_k$ and $K = d_1Ry_1 + \cdots + d_kRy_k$. Then it is easy to see that

$$F/K \cong \underbrace{R \oplus \cdots \oplus R}_{r \text{ copies}} \oplus R/Rd_1 \oplus \cdots \oplus R/Rd_k$$

since $Rx_i \cong R$ and $Ry_i/d_iRy_i \cong R/Rd_i$.

If we have an arbitrary commutative ring R, K may not have an appropriate form. Moreover, no change of basis may be possible which changes K to the appropriate form. However, in case R is a PID, it is always possible to choose a basis for F and a basis for K (which will also be free) so that the above situation exists. In addition, it will be possible to choose $d_1, \ldots, d_k$ so that $d_1|d_2, \ldots, d_{k-1}|d_k$. This will yield the desired structure theorem for finitely generated modules over R.

The proof will consist of two stages.

Stage I. We prove an appropriate theorem about $m \times n$ matrices over a PID R.

Stage II. We show that theorem about $m \times n$ matrices proved in Stage I can be translated into a theorem about modules—namely the structure theorem for modules over a PID.

We shall now prove a theorem which says that any $m \times n$ matrix $[A]$ over a PID R can be transformed to a diagonal matrix by a transformation

$$[A] \to [P][A][Q]$$

where $[P]$ and $[Q]$ are appropriate invertible matrices over R.

Let R be a ring and $\mathrm{GL}_n(R)$ the group of invertible $n \times n$ matrices over R. It is called the **general linear group over** R. Moreover, if R is commutative, then

$$\mathrm{GL}_n(R) = \{A \in M_n(R) : \det A \text{ is a unit in } R\}.$$

Theorem 4.4.4. *Let R be a commutative ring.*

1. *If $Ra + Rb = R$, then $\mathrm{GL}_2(R)$ contains a matrix of the form $\begin{bmatrix} a & b \\ * & * \end{bmatrix}$.*
2. *If $A = [a_{ij}]$ is an $m \times n$ matrix over R, $P \in \mathrm{GL}_m(R)$ and $PA = [b_{ij}]$, then*
$$\sum_{i,j} Ra_{ij} = \sum_{i,j} Rb_{ij}.$$
In other words, the entries of A and of PA generate the same ideal in R.
3. *If E is the elementary matrix obtained by interchanging the i-th and j-th rows of the identity matrix I_m and A is an $m \times n$ matrix, then $E \in \mathrm{GL}_m(R)$ and EA is the matrix obtained by interchanging the i-th and j-th rows of A.*
4. *If E is the elementary matrix obtained by multiplying the i-th row of the identity matrix I_m by a unit $c \in R$ and A is an $m \times n$ matrix, then $E \in \mathrm{GL}_m(R)$ and EA is the matrix obtained by multiplying the i-th row of A by c.*
5. *If E is the elementary matrix obtained by adding c times the j-th row of I_m to the i-th row of I_m and A is an $m \times n$ matrix, then $E \in \mathrm{GL}_m(R)$ and EA is the matrix obtained by adding c times the j-th row of A to the i-th row of A.*
6. *The analogues of (1)–(5) hold for right multiplications and column transformations.*

Proof. (1) If $Ra + Rb = R$, let $ra + sb = 1$. Then
$$\begin{bmatrix} a & b \\ -s & r \end{bmatrix} \begin{bmatrix} r & -b \\ s & a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$
(2) The entries b_{ij} of PA are R-linear combinations of the a_{ij}, so $\sum Rb_{ij} \subseteq \sum Ra_{ij}$. But since $P^{-1}(PA) = A$ the entries a_{ij} of A are R-linear combinations of b_{ij}, so $\sum Ra_{ij} \subseteq \sum Rb_{ij}$.
(3), (4), (5) and (6) are clear. □

Remark. Passing from A to EA as in (3), (4) and (5) of the above theorem are called **elementary row transformations of** A. **Elementary column transformations of** A are defined similarly.

Recall that if R is a PID and $a, b \in R$, then $Ra + Rb = Rc$ where $c = \gcd(a, b)$. Moreover, if we let $a = cx$ and $b = cy$, then $\gcd(x, y) = 1$, or $Rx + Ry = R$. More generally, $Ra_1 + \cdots + Ra_n = Rd$ where $d = \gcd(a_1, \ldots, a_n)$ and if $a_i = db_i$, then $Rb_1 + \cdots + Rb_n = R$. In UFD, $\gcd(a, b) = 1$ does not imply $Ra + Rb = R$. For example, $R = F[x, y]$, where F is a field, and $\gcd(x, y) = 1$.

Theorem 4.4.5. *Let R be a PID and $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in M_2(R)$. Then there exist $P, Q \in \mathrm{GL}_2(R)$ such that*
$$PAQ = \begin{bmatrix} e & 0 \\ 0 & f \end{bmatrix}$$
where $e = \gcd(a, b, c, d)$ and $e \mid f$.

Proof. We first claim
$(*)$ if $\begin{bmatrix} p & q \\ r & s \end{bmatrix} \in M_2(R)$, either $p = \gcd(p, q, r, s)$, or there exist $P_1, Q_1 \in \mathrm{GL}_2(R)$ such that $Rp_1 \supset Rp$ and
$$P_1 \begin{bmatrix} p & q \\ r & s \end{bmatrix} Q_1 = \begin{bmatrix} p_1 & q_1 \\ r_1 & s_1 \end{bmatrix}.$$

Case I. $q, r, s \in Rp$. Then $p = \gcd(p, q, r, s)$ and we are done.
Case II. $q \notin Rp$. Then $Rp + Rq = Rp_1$ where $p_1 = \gcd(p, q)$ and so $Rp_1 \supset Rp$. Let $p = p_1 x$

and $q = p_1 y$. Then $Rx + Ry = R$, so we can choose u, v with $xu + yv = 1$. Then $pu + qv = (p_1 x)u + (p_1 y)v = p_1$ and

$$\begin{bmatrix} p & q \\ r & s \end{bmatrix} \begin{bmatrix} u & -y \\ v & x \end{bmatrix} = \begin{bmatrix} p_1 & * \\ * & * \end{bmatrix}$$

where $Q_1 = \begin{bmatrix} u & -y \\ v & x \end{bmatrix} \in \mathrm{GL}_2(R)$.

Case III. $r \notin R_p$. Then we do the analogue of Case II with a transformation on the first column.

Case IV. $q, r \in Rp$ and $s \notin Rp$. Then we perform a succession of elementary row and column transformations followed by the manoeuvre of Case II: Let $q = \alpha p$ and $r = \beta p$:

$$\begin{aligned} \begin{bmatrix} p & q \\ r & s \end{bmatrix} &\sim \begin{bmatrix} p & q - \alpha p \\ r & s - \alpha r \end{bmatrix} = \begin{bmatrix} p & 0 \\ r & s - \alpha\beta p \end{bmatrix} \\ &\sim \begin{bmatrix} p & 0 \\ r - \beta p & s - \alpha\beta p \end{bmatrix} = \begin{bmatrix} p & 0 \\ 0 & s - \alpha\beta p \end{bmatrix} \\ &\sim \begin{bmatrix} p & s - \alpha\beta p \\ 0 & s - \alpha\beta p \end{bmatrix} \sim \begin{bmatrix} p & s \\ 0 & s - \alpha\beta p \end{bmatrix} \sim \begin{bmatrix} \gcd(p,s) & * \\ * & * \end{bmatrix} \text{ (by Case II).} \end{aligned}$$

Since this succession of operators corresponds to a transformation $\begin{bmatrix} p & q \\ r & s \end{bmatrix} \sim P_1 \begin{bmatrix} p & q \\ r & s \end{bmatrix} Q_1$ where $P_1, Q_1 \in \mathrm{GL}_2(R)$, we are done in Case IV also. This proves $(*)$.

Next we claim

$(**)$ there exist $\bar{P}, \bar{Q} \in \mathrm{GL}_2(R)$ such that $\bar{P}A\bar{Q} = \begin{bmatrix} e & * \\ * & * \end{bmatrix}$ where $e = \gcd(a, b, c, d)$.

If $a = \gcd(a, b, c, d)$ we are done.

If not, use $(*)$ to choose $P_1, Q_1 \in \mathrm{GL}_2(R)$ such that $Ra_1 \supset Ra$ and

$$P_1 \begin{bmatrix} a & b \\ c & d \end{bmatrix} Q_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}.$$

If $a_1 = \gcd(a_1, b_1, c_1, d_1)$, take $\bar{P} = P, \bar{Q} = Q$ and end the process. If not, use $(*)$ again to choose $P_2, Q_2 \in \mathrm{GL}_2(R)$ such that $Ra_2 \supset Ra_1$ and

$$P_2 \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} Q_2 = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}.$$

Continue inductively to get

$$P_{m+1} \begin{bmatrix} a_m & b_m \\ c_m & d_m \end{bmatrix} Q_{m+1} = \begin{bmatrix} a_{m+1} & b_{m+1} \\ c_{m+1} & d_{m+1} \end{bmatrix}$$

where $Ra_{m+1} \supset Ra_m$ as long as $a_m \neq \gcd(a_m, b_m, c_m, d_m)$. Since $Ra \subset Ra_1 \subset Ra_2 \subset \ldots$ is a strictly increasing chain of ideals in a PID R, the process must terminate. That is, $a_m = \gcd(a_m, b_m, c_m, d_m)$ for some m. Now let $\bar{P} = P_m P_{m-1} \ldots P_1$ and $\bar{Q} = Q_1 \ldots Q_{m-1} Q_m$. Then $\bar{P}A\bar{Q} = \bar{P} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \bar{Q} = \begin{bmatrix} a_m & b_m \\ c_m & d_m \end{bmatrix}$ where $a_m = \gcd(a_m, b_m, c_m, d_m) = \gcd(a, b, c, d)$ because $\bar{P}, \bar{Q} \in \mathrm{GL}_2(R)$, as required. This proves $(**)$.

Finally, to prove the theorem we follow the transformation $A \sim \bar{P}A\bar{Q}$ by two elementary transformations:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \sim \bar{P}A\bar{Q} = \begin{bmatrix} e & \alpha e \\ \beta e & \gamma e \end{bmatrix} \sim \begin{bmatrix} e & 0 \\ \beta e & (\gamma - \alpha\beta)e \end{bmatrix} \sim \begin{bmatrix} e & 0 \\ 0 & (\gamma - \alpha\beta)e \end{bmatrix} = \begin{bmatrix} e & 0 \\ 0 & f \end{bmatrix}.$$

Hence, we have the desired theorem. □

Theorem 4.4.6. *Let R be a PID and A an $m \times n$ matrix over R. Then there exist $P \in \mathrm{GL}_m(R)$ and $Q \in \mathrm{GL}_n(R)$ such that*

$$PAQ = \begin{bmatrix} d_1 & & & & & \\ & d_2 & & 0 & & \\ & & \ddots & & & \\ & & & d_r & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

where the $d_i \neq 0$ and $d_1 \mid d_2, \ldots, d_{r-1} \mid d_r$.

Proof. We shall prove this theorem by induction on (m, n).
Case I. $A = \begin{bmatrix} a_1 & \ldots & a_n \end{bmatrix}$ is a $1 \times n$ matrix. The proof of Theorem 4.4.5 shows that for any $a, b \in R$ there is a 2×2 matrix $Q \in \mathrm{GL}_2(R)$ such that $\begin{bmatrix} a & b \end{bmatrix} Q = \begin{bmatrix} \gcd(a,b) & 0 \end{bmatrix}$. Hence, for an appropriate $Q_1 \in \mathrm{GL}_2(R)$

$$\begin{bmatrix} a_1 & \ldots & a_n \end{bmatrix} \begin{bmatrix} Q_1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}_{n \times n} = \begin{bmatrix} \gcd(a_1, a_2) & 0 & a_3 & \ldots & a_n \end{bmatrix}.$$

Then a succession of such right multiplications together with elementary transformations (or a suitable induction) show that we can obtain

$$AQ = \begin{bmatrix} \gcd(a_1, \ldots, a_n) & 0 & \ldots & 0 \end{bmatrix} \quad \text{for some } Q \in \mathrm{GL}_n(R)$$

as follows:

$$\begin{aligned} \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix} &\sim \begin{bmatrix} \gcd(a_1, a_2) & 0 & a_3 & \ldots & a_n \end{bmatrix} \\ &\sim \begin{bmatrix} \gcd(a_1, a_2) & a_3 & 0 & a_4 & \ldots & a_n \end{bmatrix} \\ &\sim \begin{bmatrix} \gcd(a_1, a_2, a_3) & 0 & 0 & a_4 & \ldots & a_n \end{bmatrix} \\ &\sim \cdots \sim \begin{bmatrix} \gcd(a_1, \ldots, a_n) & 0 & 0 & \ldots & 0 \end{bmatrix}. \end{aligned}$$

Case II. A is an $m \times 1$ matrix. This is similar to Case I.
Case III. The general case. We already know the result for $m \times 1$ and $1 \times n$ matrices and to proceed the induction, we shall assume that $m, n \geq 2$ and that we know the result for an $(m-1) \times (n-1)$ matrix over R. Let $A = [a_{ij}]_{m \times n}$ and let $Q_1 \in \mathrm{GL}_n(R)$ be such that $\begin{bmatrix} a_{11} & \ldots & a_{1n} \end{bmatrix} Q_1 = \begin{bmatrix} a & 0 & \ldots & 0 \end{bmatrix}$ where $a = \gcd(a_{11}, \ldots, a_{1n})$ by Case I. Then

$$AQ_1 = \begin{bmatrix} a & 0 & \ldots & 0 \\ X & & Y & \\ & & & \end{bmatrix}_{m \times n}$$

where X is an $(m-1) \times 1$ columns matrix and Y is some $(m-1) \times (n-1)$ matrix. By the inductive hypothesis there are $P_2 \in \mathrm{GL}_{m-1}(R)$ and $Q_2 \in \mathrm{GL}_{n-1}(R)$ such that

$$P_2 Y Q_2 = \begin{bmatrix} e_1 & & \\ & \ddots & \\ & & e_s \\ & & \end{bmatrix}$$

where $e_1 \mid e_2, \ldots, e_{s-1} \mid e_s$. Then

$$\begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} AQ_1 \begin{bmatrix} 1 & \\ & Q_2 \end{bmatrix} = \begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} \begin{bmatrix} a & 0 \\ X & Y \end{bmatrix} \begin{bmatrix} 1 & \\ & Q_2 \end{bmatrix} = \begin{bmatrix} a & 0 \\ P_2X & P_2YQ_2 \end{bmatrix}$$

$$= \begin{bmatrix} a & 0 & \cdots & 0 \\ a_2 & e_1 & & \\ a_3 & & \ddots & \\ \vdots & & & e_s \\ a_m & & & \end{bmatrix}.$$

We next use Case II to find a $P_2' \in \mathrm{GL}_{m-1}(R)$ such that $P_2' \begin{bmatrix} a \\ a_3 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} b \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ where $b = \gcd(a, a_3, \ldots a_m)$.

We now perform a pair of transformations

$$\begin{bmatrix} a & 0 & 0 & \cdots & \cdots & 0 \\ a_2 & e_1 & & & & \\ a_3 & & e_2 & & & \\ \vdots & & & \ddots & & \\ \vdots & & & & e_s & \\ a_m & & & & & \end{bmatrix} \sim \begin{bmatrix} a_2 & e_1 & 0 & \cdots & \cdots & 0 \\ a & 0 & 0 & \cdots & \cdots & 0 \\ a_3 & 0 & e_2 & & & \\ \vdots & \vdots & & \ddots & & \\ \vdots & \vdots & & & e_s & \\ a_m & 0 & & & & \end{bmatrix} \sim \begin{bmatrix} a_2 & e_1 & 0 & \cdots & \cdots & 0 \\ b & 0 & * & \cdots & \cdots & * \\ 0 & 0 & & & & \\ \vdots & \vdots & & Z & & \\ \vdots & \vdots & & & & \\ 0 & 0 & & & & \end{bmatrix}$$

where $Z = \begin{bmatrix} e_2 & & \\ & \ddots & \\ & & e_s \end{bmatrix}$ is a matrix all of whose entries are divisible by e_1.

Now by Theorem 4.4.5, there are $P_3, Q_3 \in \mathrm{GL}_2(R)$ such that

$$P_3 \begin{bmatrix} a_2 & e_1 \\ b & 0 \end{bmatrix} Q_3 = \begin{bmatrix} e & 0 \\ 0 & f \end{bmatrix}$$

where $e = \gcd(a_2, e_1, b) = \gcd(a_2, e_1, b, \text{ entries of } Z) = \gcd(\text{entries of } A)$. Then

$$\begin{bmatrix} P_3 & & & \\ & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \begin{bmatrix} a_2 & e_1 & 0 & \cdots & 0 \\ b & 0 & * & \cdots & * \\ 0 & & & & \\ \vdots & & & Z & \\ 0 & & & & \end{bmatrix} \begin{bmatrix} Q_3 & & & \\ & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} e & 0 & * & \cdots & * \\ 0 & f & * & \cdots & * \\ 0 & & & & \\ \vdots & & Z & & \\ 0 & & & & \end{bmatrix} = \begin{bmatrix} e & * & \cdots & * \\ 0 & & & \\ \vdots & & W & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} e & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & W' & \\ 0 & & & \end{bmatrix}$$

where W' is an $(m-1) \times (n-1)$ matrix over R and e divides every entry of W'. Now we use the inductive hypothesis again to choose $P_4 \in \mathrm{GL}_{m-1}(R), Q_4 \in \mathrm{GL}_{n-1}(R)$ such that

$$P_4 W' Q_4 = \begin{bmatrix} d_2 & & \\ & \ddots & \\ & & d_r \\ & & \end{bmatrix}$$

where $d_2 \mid d_3, \ldots, d_{r-1} \mid d_r$. We note that since e divides all the entries of W', $e \mid d_2$. Hence, setting $e = d_1$, we have

$$\begin{bmatrix} 1 & \\ & P_4 \end{bmatrix} \begin{bmatrix} e & \\ & W' \end{bmatrix} \begin{bmatrix} 1 & \\ & Q_4 \end{bmatrix} = \begin{bmatrix} e & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_r \\ & & & \end{bmatrix} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_r \\ & & & \end{bmatrix}$$

where $d_1 \mid d_2, \ldots, d_{r-1} \mid d_r$. The whole series of transformations on A amount to a transformation $A \sim PAQ$ where $P \in \mathrm{GL}_m(R)$ and $Q \in \mathrm{GL}_n(R)$. Therefore, the theorem is proved. □

Before we can use our result on matrices to show that a finitely generated module over a PID R is a direct sum of cyclic modules, we need to show that every submodule of $R^n = \underbrace{R \oplus \cdots \oplus R}_{n}$ is finitely generated. In fact we shall show that every submodule of R^n is free of rank $\leq n$.

Theorem 4.4.7. *Let R be a ring and let $M = Rx_1 + \cdots + Rx_m$ and $N = Ry_1 + \cdots + Ry_n$ be free R-modules of rank m and n, respectively. Suppose $0 \to M \xrightarrow{j} P \xrightarrow{\pi} N \to 0$ is an exact sequence of R-modules. Then P is a free R-module of rank $m + n$.*

Proof. It follows from Theorem 4.3.3. □

Theorem 4.4.8. *If R is a PID and P is a submodule of $R^n = \underbrace{R \oplus \cdots \oplus R}_{n}$, then P is free of rank $\leq n$.*

Proof. We shall use induction on n. For $n = 1$, we have P is a submodule of R, i.e. P is an ideal of R, so $P = Rx$ for some $x \in R$. If $x = 0$, $P = 0$, so P is free of rank 0; if $x \neq 0$, $Rx \cong R$ as a left R-module, so P is free of rank 1. Next suppose $n > 1$ and the theorem is true for free R-submodules of rank $< n$. Let

$$R^n = Rx_1 \oplus \cdots \oplus Rx_n = (Rx_1 \oplus \cdots \oplus Rx_{n-1}) \oplus Rx_n.$$

Then we have an exact sequence

$$0 \to Rx_1 \oplus \cdots \oplus Rx_{n-1} \xrightarrow{i} R^n \xrightarrow{\pi} Rx_n \to 0$$

where i is the inclusion map and π is the projection onto the last factor. Let $M = (Rx_1 \oplus \cdots \oplus Rx_{n-1}) \cap P \subseteq Rx_1 \oplus \cdots \oplus Rx_{n-1}$ and $N = \pi(P) \subseteq Rx_n$. Then

$$0 \to M \xrightarrow{i} P \xrightarrow{\pi|_P} N \to 0$$

is an exact sequence of R-modules. M is a submodule of R^{n-1} and N is a submodule of $Rx_n \cong R$, so both are free of ranks $\leq n - 1$ and 1, respectively. Hence, P is free of rank $\leq n = (n - 1) + 1$ by Theorem 4.4.7. □

Theorem 4.4.9. *Let R be a PID and A a finitely generated R-module. Then A is a direct sum of cyclic R-modules. More precisely,*

$$A \cong \underbrace{R \oplus \cdots \oplus R}_{r} \oplus R/Rd_1 \oplus \cdots \oplus R/Rd_k$$

where $r \geq 0$ and $d_1, \ldots, d_k$ are nonzero elements of R and $d_1 \mid d_2, \ldots, d_{k-1} \mid d_k$.

Proof. Since A is finitely generated, there is an exact sequence $0 \to N \to M \to A \to 0$ where $M = Rx_1 + \dots Rx_n$ is free of finite rank n and N is a submodule of M. By Theorem 4.4.8, N is finitely generated, say by

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{n1}x_n \\ y_2 &= a_{12}x_1 + a_{22}x_2 + \cdots + a_{n2}x_n \\ &\vdots \\ y_m &= a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{nm}x_n. \end{aligned}$$

Let $\bar{M} = R^n$ be the space of $n \times 1$ column vector over R and let $\bar{N}$ be the R-submodule of $\bar{M}$ generated by the columns of the $n \times m$ matrix

$$[\bar{N}] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}.$$

There is an obvious R-module isomorphism $\alpha : M \to \bar{M}$ defined by

$$\alpha(r_1x_1 + \cdots + r_nx_n) = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}$$

and it is clear that $\alpha(N) = \bar{N}$. Hence, we have R-module isomorphisms

$$A \cong M/N \cong \alpha(M)/\alpha(N) = \bar{M}/\bar{N}.$$

By Theorem 4.4.6, there are matrices $[P] \in \mathrm{GL}_n(R)$ and $[Q] \in \mathrm{GL}_m(R)$ such that

$$[P][\bar{N}][Q] = \begin{bmatrix} d_1 & & & & & \\ & \ddots & & & 0 & \\ & & d_k & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{bmatrix}_{n \times m}$$

where the $d_i \neq 0$ and $d_1 \mid d_2, \dots, d_{k-1} \mid d_k$.

$[\bar{N}][Q]$ is an $n \times m$ matrix whose columns generate $\bar{N}$ as an R-module. Further, left multiplication by $[P]$ is an isomorphism of $\bar{M}$ which carries $\bar{N}$ to the R-submodule $\bar{U}$ of $\bar{M}$ generated by the columns of $[P][\bar{N}][Q]$. Then

$$A \cong \bar{M}/\bar{N} \cong [P]\bar{M}/[P]\bar{N} = \bar{M}/\bar{U}.$$

However, we can see by inspection that

$$\bar{M}/\bar{U} \cong R/Rd_1 \oplus R/Rd_2 \oplus \cdots \oplus R/Rd_k \oplus \underbrace{R \oplus \cdots \oplus R}_{n-k}.$$

Hence, $A \cong \underbrace{R \oplus \cdots \oplus R}_{n-k} \oplus R/Rd_1 \oplus R/Rd_2 \oplus \cdots \oplus R/Rd_k$ where $d_1 \mid d_2, \dots, d_{k-1} \mid d_k$. □

Example 4.4.1. Let A be the $\mathbb{Z}$-module generated by x, y and z with the relations

$$x + y = 0 \quad \text{and} \quad x - y + 2z = 0.$$

Express A as a direct sum of cyclic modules.

Solution. Observe that $A = \{(x, y, z) \in \mathbb{Z}^3 : x + y = 0 \text{ and } x - y + 2z = 0\}$. Consider the exact sequence

$$0 \longrightarrow N = \mathbb{Z}(x + y) + \mathbb{Z}(x - y + 2z) \longrightarrow \mathbb{Z}x \oplus \mathbb{Z}y \oplus \mathbb{Z}z \longrightarrow A \longrightarrow 0.$$

Then $y_1 = x + y$ and $y_2 = x - y + 2z$, so

$$[\bar{N}] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} = [\bar{U}].$$

Thus, $U = \mathbb{Z}x \oplus \mathbb{Z}(2y)$. Hence,

$$A \cong M/U = (\mathbb{Z}x \oplus \mathbb{Z}y \oplus \mathbb{Z}z)/(\mathbb{Z}x \oplus \mathbb{Z}(2y)) \cong \{0\} \oplus \mathbb{Z}_2 \oplus \mathbb{Z}z \cong \mathbb{Z}_2 \oplus \mathbb{Z}$$

as desired. □

Our next goal is to show that the direct summands which occur in the above decomposition are unique up to isomorphism. This does NOT mean that the actual summands which occur are unique. For example, suppose $R = \mathbb{Z}$ and

$$A = \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3 \cong \mathbb{Z}x_1 \oplus \mathbb{Z}x_2 \oplus \mathbb{Z}y_1 \oplus \mathbb{Z}y_2$$

where $\mathbb{Z}x_1$ and $\mathbb{Z}x_2$ are free summands and $3y_1 = 3y_2 = 0$. Then we can also write

$$A = \mathbb{Z}(x_1 + 2x_2 + y_2) \oplus \mathbb{Z}x_2 \oplus \mathbb{Z}(2y_1 + y_2) \oplus \mathbb{Z}(y_1 + y_2) \cong \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3$$

In the first case, the direct summands are

$$\mathbb{Z}x_1 \cong \mathbb{Z},\ \mathbb{Z}x_2 \cong \mathbb{Z},\ \mathbb{Z}y_1 \cong \mathbb{Z}_3,\ \mathbb{Z}y_2 \cong \mathbb{Z}_3.$$

In the second case, the direct summands are

$$\mathbb{Z}(x_1 + 2x_2 + y_2) \cong \mathbb{Z},\ \mathbb{Z}x_2 \cong \mathbb{Z},\ \mathbb{Z}(2y_1 + y_2) \cong \mathbb{Z}_3,\ \mathbb{Z}(y_1 + y_2) \cong \mathbb{Z}_3.$$

Then the summands which occurs are distinct submodules in the two cases, but the isomorphism classes of summands are the same, namely $\mathbb{Z}, \mathbb{Z}, \mathbb{Z}_3$ and $\mathbb{Z}_3$.

As a preparation for proving uniqueness, we shall need the concept of a *torsion element*.

Let R be an integral domain and M an R-module. An $m \in M$ is called a **torsion element** of M if there is a nonzero $r \in R$ such that $rm = 0$. Let $\tau(M)$ denote the set of torsion elements of R, called the **torsion submodule** of M. If $\tau(M) = 0$, M is said to be a **torsion free** R-module.

Theorem 4.4.10. *Let R be an integral domain and M an R-module. Then*

1. *$\tau(M)$ is a submodule of M.*
2. *$\tau(M/\tau(M)) = 0$.*

Proof. (1) The only problem in showing that $\tau(M)$ is a submodule of M is in showing that $\tau(M)$ is closed under addition. Suppose $x, y \in \tau(M)$. Then there exist nonzero elements $r, s \in R$ such that $rx = 0$ and $sy = 0$. Since R is an integral domain, $rs \neq 0$. But

$$rs(x + y) = s(rx) + r(sy) = 0 + 0 = 0.$$

Hence, $x + y \in \tau(M)$.
(2) Suppose $x + \tau(M) \in \tau(M/\tau(M))$. Then there is a nonzero $r \in R$ such that

$$r(x + \tau(M)) = rx + \tau(M) = 0 + \tau(M),$$

i.e., $rx \in \tau(M)$. Thus, there is a nonzero $s \in R$ such that $s(rx) = 0$, so $sr \neq 0$ and $(sr)x = 0$. Hence, $x \in \tau(M)$, so $\tau(M/\tau(M)) = 0$. □

Remarks.
1. If R is not an integral domain, then the torsion elements of an R-module M may not form a submodule, even if R is commutative. For example, let $R = F \times F$ where F is a field and let $M = R = F \times F$. Then the torsion elements of M are all elements of the form $(a, 0)$ or $(0, b)$. But if $a, b \neq 0$, $(a, b) = (a, 0) + (0, b)$ is not a torsion element.
2. If R is not commutative, then the torsion elements of an R-module M may not form a submodule, even if R has no zero divisors. For example, there exists a non-commutative domain R (such as the polynomial rings over the quaternion ring) such that for some nonzero $x, y \in R$, $Rx \cap Ry = 0$. In other words, x and y have no common left multiple except 0. For such an R, x and y, let $M = R/Rx$ as a left R-module. Then
   (a) $y + Rx$ is not a torsion element of M, for $0 = r(y + Rx) = ry + Rx$, so $ry \in Rx \cap Ry = 0$. Thus, $ry = 0$, so $r = 0$.
   (b) $1 + Rx$ is a torsion element of M since $x(1 + Rx) = x + Rx = 0$. Since $1 + Rx$ generates $M = R/Rx$ as a left R-module, it follows that the torsion elements of M do not form a submodule.

Theorem 4.4.11. *Let R be a PID and let p be an irreducible element of R.*
1. *Rp is a maximal ideal of R, i.e., R/Rp is a field.*
2. *If $d \in R$ and $p \nmid d$, then $p(R/Rd) = R/Rd$.*
3. *If $d \in R$ and $p \mid d$, then $p(R/Rd) = Rp/Rd \cong R/R(d/p)$.*

Proof. (1) Suppose Rp is not a maximal ideal and let $Rp \subset Rx \subset R$. Then $p = rx$ where neither r nor x is a unit of R, which contradicts the hypothesis that p is irreducible.
(2) Since Rp is a maximal ideal and $p \nmid d$, $Rp + Rd = R$. Thus, we can choose $r, s \in R$ with $rp + sd = 1$. Then for any $x \in R$, $x + Rd = (rp + sd)x + Rd = prx + Rd = p(rx + Rd)$. Hence, $p(R/Rd) = R/Rd$.
(3) Since $p \mid d$, $Rd \subset Rp$. The multiplication by p defines an onto R-module homomorphism $\varphi_p : R \to Rp/Rd$ where $\varphi_p(x) = xp + Rd$. It is easy to verify that $\ker \varphi_p = R(d/p)$. Hence, we have the theorem. □

Theorem 4.4.12. *Let R be a PID and suppose $d_1, \ldots, d_k, e_1, \ldots, e_m$ are nonzero nonunits of R, where $d_1 \mid d_2, \ldots, d_{k-1} \mid d_k, e_1 \mid e_2, \ldots, e_{m-1} \mid e_m$. Suppose*

$$A = R/Rd_1 \oplus \cdots \oplus R/Rd_k \cong R/Re_1 \oplus \cdots \oplus R/Re_m = B.$$

Then $k = m$ and $R/Rd_i \cong R/Re_i$ for $i = 1, \ldots, m$. In particular, d_i and e_i are associate for $i = 1, \ldots, m$.

Proof. Let p be a prime of R which divides d_1. Then $p \mid d_i$ for all $i = 1, \ldots, k$, so

$$(R/Rd_i)/p(R/Rd_i) = (R/Rd_i)/(Rp/Rd_i) \cong R/Rp$$

for all $i = 1, \ldots, k$. Thus, $A/pA \cong \underbrace{R/Rp \oplus \cdots \oplus R/Rp}_{k}$. In other words, A/pA is a vector space over the field R/Rp of dimension k. Note that since $p(M/pM) = pM/pM = 0$ for any R-module M, M/pM may be considered as an R/Rp module, i.e., as a vector space over R/Rp.

Since $A \cong B$, $A/pA \cong B/pB$ since any isomorphism $\phi : A \to B$ carries pA onto pB. But since B is generated as an R/Rp-module by $\leq m$ elements. Thus,

$$m \geq \dim_{R/Rp}(B/pB) = \dim_{R/Rp}(A/pA) = k.$$

By symmetry, $k \geq m$. Hence, $m = k$.

We now show that $R/Rd_i \cong R/Re_i$ by induction on the number n of prime divisors of $d_1 \cdots d_k$. E.g., for $d_1 \ldots d_k = p_1^{\alpha_1} \ldots p_r^{\alpha_r}$, we have $n = \alpha_1 + \cdots + \alpha_r$. If $n = 1$, then $k = 1$ and $A = R/Rd_1 \cong R/Re_1 = B$. For inductive step, let p be a prime divisor of d_1 and hence of $d_2, \ldots, d_k$. Then $A/pA \cong \underbrace{R/Rp \oplus \cdots \oplus R/Rp}_{k}$ as above. Suppose $p \nmid e_1$. Then $p(R/Re_1) = R/Re_1$, by Theorem 4.4.11, so $(R/Re_1)/p(R/Re_1) = 0$. Thus,

$$\begin{aligned} B/pB &= (R/Re_1)/p(R/Re_1) \oplus \cdots \oplus (R/Re_k)/p(R/Re_k) \\ &\cong (R/Re_2)/p(R/Re_2) \oplus \cdots \oplus (R/Re_k)/p(R/Re_k) \end{aligned}$$

is generated by $\leq k-1$ elements. Hence, $\dim_{R/Rp}(B/pB) \leq k-1$ and $\dim_{R/Rp}(A/pA) = k$, a contradiction, since $A/pA \cong B/pB$ as above. Then $p \mid e_1$, so $p \mid e_i$ for all $i = 1, \ldots, k$. By Theorem 4.4.11, we have isomorphisms

$$\begin{aligned} R/R(d_1/p) \oplus \cdots \oplus R/R(d_k/p) &\cong p(R/Rd_1) \oplus \cdots \oplus p(R/Rd_k) \\ &= pA \cong pB \cong p(R/Re_1) \oplus \cdots \oplus p(R/Re_k) \\ &\cong R/R(e_1/p) \oplus \cdots \oplus R/R(e_k/p). \end{aligned}$$

Now the number of prime factors of $(d_1/p) \cdots (d_k/p)$ is strictly less than the number of prime factors of $d_1 \cdots d_k$. Hence, the inductive hypothesis applies to the isomorphism

$$R/R(d_1/p) \oplus \cdots \oplus R/R(d_k/p) \cong R/R(e_1/p) \oplus \cdots \oplus R/R(e_k/p).$$

Thus, we may conclude that $R/R(d_i/p) \cong R/R(e_i/p)$ for $i = 1, \ldots, k$.

Note that for any ideal I of R, $I = \operatorname{ann}(R/I) = \{r \in R : r(R/I) = 0\}$. Hence, $R/I \cong R/J$ if and only if $I = \operatorname{ann}(R/I) \cong \operatorname{ann}(R/J) = J$ (as submodules of R), and so

$$R/R(d_i/p) \cong R/R(e_i/p) \Leftrightarrow R(d_i/p) \cong R(e_i/p) \Leftrightarrow Rd_i \cong Re_i \Leftrightarrow R/Rd_i \cong R/Re_i.$$

Therefore, $R/Rd_i \cong R/Re_i$ for $i = 1, \ldots, k$ and the theorem is proved. □

Theorem 4.4.13. *Let R be a PID. Suppose that*

$$A \cong \underbrace{R \oplus \cdots \oplus R}_{r} \oplus R/Rd_1 \oplus \cdots \oplus R/Rd_k$$

and

$$B \cong \underbrace{R \oplus \cdots \oplus R}_{s} \oplus R/Re_1 \oplus \cdots \oplus R/Re_m$$

are isomorphic R-modules where the d_i and e_i are nonzero nonunits, $d_1 \mid d_2, \ldots, d_{k-1} \mid d_k$ and $e_1 \mid e_2, \ldots, e_{m-1} \mid e_m$. Then $r = s$, $k = m$ and $R/Rd_i \cong R/Re_i$ for all $i = 1, \ldots, k$.

Proof. We first observe that the torsion submodules of A and B are

$$\tau(A) = R/Rd_1 \oplus \cdots \oplus R/Rd_k \text{ and } \tau(B) = R/Re_1 \oplus \cdots \oplus R/Re_m.$$

Also,

$$A/\tau(A) = \underbrace{R \oplus \cdots \oplus R}_{r} \text{ and } B/\tau(B) = \underbrace{R \oplus \cdots \oplus R}_{s}.$$

Now if $\phi : A \to B$ is an isomorphism, it is easy to see that $\phi(\tau(A)) = \tau(B)$, so ϕ induces isomorphisms

$$\phi|_{\tau(A)} : \tau(A) \to \tau(B) \text{ and } \hat{\phi} : A/\tau(A) \to B/\tau(B).$$

In particular, $\hat{\phi}$ is an isomorphism between a free R-module of rank r and one of rank s. Hence, $r = s$ by Theorem 4.2.6. Finally, Theorem 4.4.12 applies to the isomorphism between $\tau(A)$ and $\tau(B)$ and shows that $k = m$ and $R/Rd_i \cong R/Re_i$ for all $i = 1, \ldots, k$. □

Exercises 4.4.

1. Let R be a commutative ring such that every submodule of a free R-module is free. Prove that R is a PID.
2. Prove that every finitely generated subgroup of the additive group $(\mathbb{Q}, +)$ is cyclic.
3. Let $R = \mathbb{Z}[x]$ and let $M = (2, x)$ be the ideal generated by 2 and x, considered as a submodule of R. Show that $\{2, x\}$ is not a basis of M. Show that the rank of M is 1 but that M is not free of rank 1.
4. Let R be a PID. Prove that
   (a) For any $a, b \in R$, if $\gcd(a, b) = 1$, then $R/Rab \cong R/Ra \oplus R/Rb$.
   (b) If $d = p_1^{n_1} \ldots p_k^{n_k}$ where $p_1, \ldots, p_k$ are distinct primes and $n_1, \ldots, n_k > 0$, then
   $$R/Rd \cong R/Rp_1^{n_1} \oplus \cdots \oplus R/Rp_k^{n_k}.$$
5. Let M be the $\mathbb{Z}$-module generated by a, b and c with the relations
   $$4a + 3b + 3c = 0 \quad \text{and} \quad 2a - b + 3c = 0.$$
   Express M as a direct sum of cyclic modules. What are the orders of these modules?
6. Let D be the ring of Gaussian integers $\mathbb{Z}[i]$ and $M = D^3$ the free D-module of rank 3. Take K to be the submodule generated by $(1, 2, 1), (0, 0, 5)$ and $(1, -i, 6)$. Prove that M/K is finite and determine its order.
7. Let D be the ring of Gaussian integers $\mathbb{Z}[i]$. Determine the structure of D^3/K where K is generated by $f_1 = (1, 3, 6), f_2 = (2 + 3i, -3i, 12 - 18i)$ and $f_3 = (2 - 3i, 6 + 9i, -18i)$. Show that $M = D^3/K$ is finite (of order 352512). (The order of the ring $\mathbb{Z}[i]/(a + bi)$ is $a^2 + b^2$.)
8. Let $D = \mathbb{Q}[x]$ be the polynomial ring in one variable over the field $\mathbb{Q}$ of rational numbers. Let K be the submodule of D^3 generated by $(2x - 1, x, x^2 + 3)$ and (x, x, x^2). Find polynomials $g_1, \ldots, g_r$ such that $D^3/K \cong D/(g_1) \oplus \cdots \oplus D/(g_r)$.

4.5 Noetherian Rings

In the proof of Theorem 2.4.11 (every PID is a UFD), the fact that "every ideal of R is principal" is used to argue that there is no infinite strictly increasing chain of ideals in R. A ring with this property is called a *Noetherian ring*, in honor of Emmy Noether, who inaugurated the use of chain condition in algebra. Noetherian rings are of the utmost importance in algebraic geometry and algebraic number theory. One reason for this is that for any field F, $F[x_1, \ldots, x_n]$, $n \geq 2$, is Noetherian domain but not a PID. We shall study Noetherian rings in this section.

A partially ordered set Σ has the **ascending chain condition (a.c.c.)** if every chain

$$s_1 \leq s_2 \leq \ldots$$

eventually breaks off, that is, $s_k = s_{k+1} = \dots$ for some k. This is a finiteness condition in logic that allows arguments by induction, even when the partially ordered set Σ is infinite. It is easy to see that a partially ordered set Σ has the a.c.c. if and only if every nonempty subset $S \subset \Sigma$ has a maximal element: If $\emptyset \neq S \subset \Sigma$ does not have a maximal element, then choose $s_1 \in S$, and for each s_k, an element s_{k+1} with $s_k < s_{k+1}$, thus contradicting the a.c.c..

Theorem 4.5.1. *Let R be a ring. The following three conditions are equivalent.*
(i) The set Σ of left ideals of R has the a.c.c.; in other words, every increasing chain of left ideals $I_1 \subset I_2 \subset \dots$ eventually stops, that is $I_k = I_{k+1} = \dots$ for some k.
(ii) Every nonempty set $\mathscr{S}$ of left ideals has a maximal element.
(iii) Every left ideal $I \subset R$ is finitely generated.
If one of these conditions hold, then R is **Noetherian** *(named after E. Noether).*

Proof. Here (i) $\Leftrightarrow$ (ii) is the purely logical statement about partially ordered sets already discussed, whereas (i) or (ii) $\Leftrightarrow$ (iii) is directly concerned with rings and ideals.

(i) $\Rightarrow$ (iii). Pick $f_1 \in I$, then if possible $f_2 \in I \smallsetminus (f_1)$, and so on. At each step, if $I \neq (f_1, \dots, f_k)$, pick $f_{k+1} \in I \smallsetminus (f_1, \dots, f_k)$. Then by the a.c.c. (i), the chain of ideals

$$(f_1) \subset (f_1, f_2) \subset \cdots \subset (f_1, \dots, f_k) \subset \cdots$$

must break off at some stage, and this can only happen if $(f_1, \dots, f_k) = I$ for some k. This proof involves an implicit appeal to the axiom of choice. It is perhaps cleaner to do (i) $\Rightarrow$ (ii) purely in set theory, then argue as follows.

(ii) $\Rightarrow$ (iii). Let I be a left ideal of R and consider the set $\mathscr{S}$ of finitely generated left ideals contained in I. Then $\{0\} \in \mathscr{S}$, so that $\mathscr{S}$ has a maximal element J by (ii). But then $J = I$, since any element $f \in I \smallsetminus J$ would give rise to a strictly bigger finitely generated left ideal $J \subset (J, f) \subseteq I$.

(iii) $\Rightarrow$ (i). Let $I_1 \subset I_2 \subset \dots$ be an increasing chain of left ideals. Then $J = \bigcup_k I_k$ is again an ideal. If J is finitely generated then $J = (f_1, \dots, f_n)$ and each $f_i \in I_{k_i}$, so that setting $k = \max k_i$ gives $J = I_k$ and the chain stops. □

Remarks.
1. Every PID is Noetherian. Hence, we may consider a Noetherian ring as a generalization of a PID.
2. Most rings of interest are Noetherian this is a very convenient condition to work with. At first sight, more concrete conditions (such as R finitely generated over k or over $\mathbb{Z}$) might seem more attractive, but as a rule, the Noetherian condition is both more general and more practical to work with.
3. The descending chain condition (d.c.c.) on a partially ordered set is defined in a similar way. A ring whose ideals satisfy the d.c.c. is called an **Artinian ring**. This is also a very important notion, but is more special: the d.c.c. for rings turns out to be very much stronger than the a.c.c. (and implies it). We shall discuss this kind of rings in the next section.

Example 4.5.1. $\mathbb{Z}$ is Noetherian but not Artinian since $\mathbb{Z} \supset p\mathbb{Z} \supset p^2\mathbb{Z} \supset \dots$, p prime, is a decreasing chain which does not stop.

Examples 4.5.2. Here are three examples of non-Noetherian rings. Let k be a field.
1. The polynomial ring $k[x_1, \dots, x_n, \dots]$ in an infinite number of indeterminates is obviously non-Noetherian.
2. Consider the ring A_1 of polynomials in x, y of the form $f(x, y) = a + xg(x, y)$ with a a constant and $g \in k[x, y]$; that is, f involves no pure power y^j of y with $j > 0$. In other words,

$$\begin{aligned} A_1 &= \left\{ f(x,y) = \sum a_{ij} x^i y^j : i, j \geq 0 \text{ and } i > 0 \text{ if } j \neq 0 \right\} \\ &= k[x, xy, xy^2, \dots, xy^n, \dots] \subset k[x, y]. \end{aligned}$$

It is clear that $(x, xy, xy^2, \dots)$ is a maximal ideal of A_1, and is not finitely generated. (It looks as if it should be generated by x, but, of course, $y, y^2, \dots$ are not elements of the ring A_1.) Thus, A_1 is not Noetherian.

3. A rather similar example is the ring A_2 of polynomials in x, y, y^{-1} of the form $g(x, y) + xh(x, y, y^{-1})$; that is,

$$A_2 = \left\{ f(x,y) = \sum a_{ij} x^i y^j : i \geq 0, \text{ and } j \geq 0 \text{ if } i = 0 \right\} = k[x, y, x/y, x/y^2, \dots, x/y^n, \dots].$$

In this ring $x = (x/y) \cdot y$, and $x/y = (x/y^2) \cdot y$, etc., so that the element x does not have a factorization into irreducibles and

$$(x) \subset (x/y) \subset (x/y^2) \subset \cdots$$

is an infinite ascending chain.

Let R be a ring. An R-module M is **Noetherian** if the submodules of M have the a.c.c., that is, any increasing chain $M_1 \subset M_2 \subset \ldots \subset M_k \subset \ldots$ of submodules eventually stops. Just as before, it is equivalent to say that any nonempty set of submodules of M has a maximal element, or that every submodule of M is finitely generated.

Theorem 4.5.2. *Let $0 \longrightarrow L \xrightarrow{\alpha} M \xrightarrow{\beta} N \longrightarrow 0$ be an exact sequence of R-modules. Then M is Noetherian if and only if L and N are.*

Proof. Obviously, the condition is necessary. Suppose $M_1 \subset M_2 \subset \ldots$ is an increasing chain of submodules of M; then identifying $\alpha(L)$ with L and taking intersection gives a chain

$$L \cap M_1 \subset L \cap M_2 \subset \ldots$$

of submodules of L and applying β gives a chain $\beta(M_1) \subset \beta(M_2) \subset \ldots$ of submodules of N. Each of these two chains eventually stops, by the assumption on L and N, so that we need to prove the following statement: □

Lemma 4.5.3. *For submodules, $M_1 \subset M_2 \subset M$, if $L \cap M_1 = L \cap M_2$ and $\beta(M_1) = \beta(M_2)$, then $M_1 = M_2$.*

Proof. Indeed, if $m \in M_2$, then $\beta(m) \in \beta(M_2) = \beta(M_1)$, so that there is an $n \in M_1$ such that $\beta(m) = \beta(n)$. Then $\beta(m - n) = 0$, so that $m - n \in M_2 \cap \ker \beta = M_1 \cap \ker \beta$. Hence, $m \in M_1$. □

We record consequences of this theorem in:

Corollary 4.5.4. *Let M be an R-modules and N an R-submodule of M. Then M is Noetherian if and only if N and M/N are.*

Corollary 4.5.5.

1. *If M_i are Noetherian modules, $i = 1, \dots, r$, then $\bigoplus_{i=1}^{r} M_i$ is Noetherian.*
2. *If R is a Noetherian ring, then an R-module M is Noetherian if and only if it is finitely generated over R.*
3. *If R is a Noetherian ring and M is a finitely generated R-module, then any submodule $N \subset M$ is again finitely generated.*
4. *If R is a Noetherian ring and $\varphi : R \to B$ is a ring homomorphism such that B is a finitely generated R-module, then B is a Noetherian ring. In particular, a homomorphic image of a Noetherian ring is a Noetherian ring.*

Proof. (1) A direct sum $M_1 \oplus M_2$ is a particular case of an exact sequence, so that the previous proves (1) when $r = 2$. The case $r > 2$ follows by an easy induction.
(2) If M is finitely generated then there is a surjective homomorphism $R^r \to M \to 0$ for some r, so that M is a quotient $M \cong R^r/N$ for some submodule $N \subset R^r$; now R^r is a Noetherian module by (1), so M Noetherian follows by the implication $\Rightarrow$ of the above theorem. Conversely, M Noetherian obviously implies M is finitely generated.
(3) This just uses the previous implication: M finitely generated and R Noetherian implies that M is Noetherian, so that N is Noetherian, which implies that N is a finitely generated R-module.
(4) B is Noetherian as an R-module; but left ideals of B are submodules of B as an R-submodule, so that B is a Noetherian ring. □

The following result provides many examples of Noetherian rings, and is the main motivation behind the use of the a.c.c. in commutative algebra. Note that in Hilbert's day, a "basis" of a module meant simply a family of generators.

Theorem 4.5.6. [Hilbert Basis Theorem] *If R is a commutative Noetherian ring, then so is the polynomial ring $R[x]$.*

Proof. We shall prove that any ideal $I \subset R[x]$ is finitely generated. For this, define auxiliary sets $J_n \subset R$ by

$$J_n = \{a \in R : \text{there exists } f \in I \text{ such that } f(x) = ax^n + b_{n-1}x^{n-1} + \cdots + b_1x + b_0\}.$$

In other words, J_n is the set of leading coefficients of elements of I of degree $n \geq 0$. Then it is easy to check that J_n is an ideal (using the fact that I is an ideal), and that $J_n \subset J_{n+1}$ (because for $f \in I$, also $xf \in I$), and therefore

$$J_0 \subseteq J_1 \subseteq J_2 \subseteq \dots$$

is an increasing chain of ideals. Using the assumption that R is Noetherian, we deduce that $J_n = J_{n+1} = \dots$ for some n.

For each $m \leq n$, the ideal $J_m \subset R$ is finitely generated, say $J_m = (a_{m,1}, \dots, a_{m,r_m})$; and by definition of J_m, for each $a_{m,j}$ with $1 \leq j \leq r_m$ there is a polynomial $f_{m,j} \in I$ of degree m having the leading coefficient $a_{m,j}$. This allows us to write down a finite set

$$S = \{f_{m,j}\}_{0 \leq m \leq n, 1 \leq j \leq r_m}$$

of elements of I.

We now claim that S generates I. Indeed, for any polynomial $f(x) \in I$, if $f(x)$ has degree m then its leading coefficient a is in J_m, hence if $m \geq n$, then $a \in J_m = J_n$, so that $a = \sum b_i a_{n,i}$ with $b_i \in R$ and $f(x) - \sum b_i x^{m-n} f_{n,i}(x)$ has degree $< m$; similarly, if $m \leq n$, then $a \in J_m$, so that $a = \sum b_i a_{m,i}$ with $b_i \in R$ and $f(x) - \sum b_i f_{m,i}(x)$ has degree $< m$. By induction on m, it follows that f can be written as a linear combination of the finitely many elements in S. This proves that any ideal of $R[x]$ is finitely generated. □

Corollary 4.5.7. *If R is a commutative Noetherian ring and $\varphi : R \to B$ is a ring homomorphism such that B is a commutative finitely generated extension ring of $\varphi(R)$, then B is Noetherian.*

Proof. The assumption is that B is a quotient of a polynomial ring, $B \cong R[x_1, \dots, x_n]/I$ for some ideal I. Now by Hilbert Basis Theorem and an obvious induction, R Noetherian implies that so is $R[x_1, \dots, x_n]$, and by Corollary 4.5.5, (4), $R[x_1, \dots, x_n]$ is Noetherian implies that so is $R[x_1, \dots, x_n]/I$. □

Exercises 4.5.
1. Let M be a finitely generated R-module where R is Noetherian. Suppose I is an ideal of R such that for each element $a \in I$, there exists a nonzero element $m \in M$ such that $am = 0_M$. Show that $Ix = \{0_M\}$ for some nonzero element $x \in M$.
2. Let M be a Noetherian R-module. Prove that $I = \{r \in R : rm = 0_M \text{ for all } m \in M\}$ is an ideal of R and R/I is Noetherian.
3. Let M be a Noetherian R-module and $\varphi : M \to M$ be a surjective module homomorphism. Prove that φ is an isomorphism. [*Hint*. consider the chain of submodules $\ker \varphi \subset \ker \varphi^2 \subset \cdots$.]

4.6 Artinian Rings

In this section, we study deeper commutative ring theory. Our main goal is to show that any finite commutative ring is a direct product of a finite number of local rings. However, we present results on more a general ring, called an "Artinian ring".

The **Jacobson radical of a ring** R is the intersection of all maximal ideals of R and is denoted by $\operatorname{Jac} R$. Note that if R is a local ring with unique maximal ideal M, then $\operatorname{Jac} R = M$. Let R be a ring. An element $a \in R$ is **nilpotent** if $a^n = 0$ for some $n \in \mathbb{N}$. The set of all nilpotent elements in a commutative ring R is an ideal, called the **nilradical of** R. It is also clear that every prime ideal in a commutative ring contains the nilradical.

Theorem 4.6.1. *Let J be the Jacobson radical of a commutative ring R.*
1. *If I is a proper ideal of R, then so is the ideal generated by I and J.*
2. *The Jacobson radical contains the nilradical of R.*
3. *For $x \in R$, $x \in J$ if and only if $1 - rx$ is a unit for all $r \in R$. In partucular, if R is a local ring with unique maximal ideal M, then $1 - m$ is a unit in R for all $m \in M$.*
4. [Nakayama's lemma] *If M is any finitely generated R-module and $JM = M$, then $M = \{0\}$.*
5. *If M is finitely generated and $M = N + IM$ for some ideal $I \subseteq J$ and submodule N of M, then $M = N$.*
6. *Let I be an ideal in the Jacobson radical of R, and suppose that M is finitely generated. If $m_1, \ldots, m_n$ have images in M/IM that generate it as an R-module, then $m_1, \ldots, m_n$ also generate M as an R-module.*

Proof. (1) If I is a proper ideal of R, then I is contained in some maximal ideal M of R. Since $J \subseteq M$, $I \cup J \subseteq M$.
(2) Let $a \in R$ be nilpotent. Then $a^n = 0$ for some $n \in \mathbb{N}$. Since maximal ideals are prime and $a^n \in M$, so $a \in M$.
(3) Suppose $1 - rx$ is not a unit for some $r \in R$ and let M be a maximal ideal containing $1 - rx$. Since $1 \notin M$, $rx \notin M$, so $x \notin M$. But $J \subseteq M$, it follows that $x \notin J$. Conversely, assume that $x \notin J$. Then there is a maximal ideal M such that $x \notin M$. Thus, $R = (x, M)$, so $1 = rx + m$ for some $r \in R$ and $m \in M$. Hence, $1 - rx = m \in M$ which implies that $1 - rx$ is not a unit in R.
(4) Assume that $M \neq \{0\}$ and let n be the smallest positive integer such that M is generated by n elements, say $m_1, \ldots, m_n$. Since $M = JM$, we have

$$m_n = r_1 m_1 + \cdots + r_n m_n \quad \text{for some } r_1, \ldots, r_n \in J.$$

Thus, $(1 - r_n)m_n = r_1 m_1 + \cdots + r_{n-1} m_{n-1}$. By (3), $1 - r_n$ is a unit, so m_n lies in the module generated by $m_1, \ldots, m_{n-1}$ which contradicts the minimality of n. Hence, $M = \{0\}$.
(5) Apply (4) to M/N.
(6) Apply (5) to $N = \sum_i Rm_i$. □

Remark. In the special case of a finitely generated module M over a local ring R with unique maximal ideal J, the quotient M/JM is a vector space over the field R/J. Statement (6) implies that a basis of M/JM lifts to a minimal set of generators of M. Conversely, every minimal set of generators of M is obtained in this way, and any two such sets of generators are related by an invertible matrix with entries in the ring.

A ring whose ideals satisfy the descending chain condition (d.c.c.), i.e., whenever $I_1 \supseteq I_2 \supseteq \dots$ is a decreasing chain of ideals of R, then there is a positive integer m such that $I_k = I_m$ for all $k \geq m$, is called an **Artinian ring** (named after E. Artin). Clearly, every finite ring is Artinian. Also, it is immediate that every quotient ring of an Artinian ring is Artinian. Similar to Theorem 4.5.1, we have the following theorem.

Theorem 4.6.2. *R is an Artinian ring if and only if every nonempty set $\mathscr{S}$ of ideals has a minimal element.*

An R-module M is said to be **Artinian** if it satisfies d.c.c. on submodules. Similar to Theorem 4.5.2, we have:

Theorem 4.6.3. *Let $0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$ be an exact sequence of R-modules. Then M is an Artinian R-module if and only if L and N are.*

Lemma 4.6.4. *Let M be a maximal ideal of the commutative ring R and suppose that $M^m = \{0\}$ for some $m \in \mathbb{N}$. Then R is Noetherian if and only if R is Artinian.*

Proof. Observe that each successive quotient M^i/M^{i+1}, $i = 0, 1, \dots, m-1$, in the filtration $R \supseteq M \supseteq M^2 \supseteq \cdots \supseteq M^{m-1} \supseteq M^m = \{0\}$ is a module over the field $F = R/M$. Consider the exact sequence $0 \longrightarrow M \longrightarrow R \longrightarrow R/M \longrightarrow 0$ of R-modules. Assume that R is Noetherian. By Theorem 4.5.1, M and R/M is Noetherian. Thus, R/M and M are Artinian by Exercise 4.6 (2). Hence, it follows from Theorem 4.6.3 that R is Artinian. The converse is proved in the same way. □

Lemma 4.6.5. *Let R be a commutative ring and P a prime ideal of R. If I and J are ideals of R such that $P \supseteq I \cap J$, then $I \subseteq P$ or $J \subseteq P$.*

Proof. Assume that $I \nsubseteq P$ and $J \nsubseteq P$. Let $x \in I$, $\notin P$ and $y \in J$, $\notin P$. Then $xy \in I \cap J$. Since x and y are not in P and P is a prime ideal, $xy \notin P$ which contradicts $P \supseteq I \cap J$. □

Now, we are ready to prove our main results.

Theorem 4.6.6. *Let R be a commutative Artinian ring.*

1. *There are only finitely many maximal ideals in R.*
2. *The quotient $R/(\operatorname{Jac} R)$ is a direct product of a finite number of fields. More precisely, if $M_1, \dots, M_n$ are finitely many maximal ideals in R, then*

$$R/(\operatorname{Jac} R) \cong k_1 \times \cdots \times k_n,$$

   where k_i is the field R/M_i for all $i \in \{1, \dots, n\}$.
3. *Every prime ideal of R is maximal. The Jacobson radical of R equals the nilradical of R and $(\operatorname{Jac} R)^m = \{0\}$ for some $m \in \mathbb{N}$.*
4. *The ring R is isomorphic to the direct product of a finite number of Artinian local rings.*
5. *Every Artinian ring is Noetherian.*

Proof. (1) Let $\mathscr{S}$ be the set of all ideals of R that are the intersection of a finite number of maximal ideals. By Theorem 4.6.2, $\mathscr{S}$ has a minimal element, say $M_1 \cap \cdots \cap M_n$. Then for any maximal ideal M, we have

$$M \cap M_1 \cap \cdots \cap M_n = M_1 \cap \cdots \cap M_n,$$

so $M \supseteq M_1 \cap \cdots \cap M_n$. By Lemma 4.6.5, $M_i \subseteq M$ for some i. Since M_i and M are maximal, $M_i = M$ and hence $M_1, \ldots, M_n$ are all maximal ideals of R.
(2) Since $M_i + M_j = R$ for all $i \neq j$ and $\operatorname{Jac} R = M_1 \cap \cdots \cap M_n$, the statement follows from the Chinese remainder theorem applied to $M_1, \ldots, M_n$.
(3) We first show that $J = \operatorname{Jac} R$ is nilpotent. By d.c.c., there is some $m \in \mathbb{N}$ such that $J^m = J^{m+i}$ for all $i \in \mathbb{N}$. Assume that $J^m \neq \{0\}$. Let $\mathscr{S}$ be the set of proper ideals I such that $IJ^m \neq \{0\}$. Then $J \in \mathscr{S}$. Let I_0 be a minimal element of $\mathscr{S}$. Thus, there is some $x \in I_0$ such that $xJ^m \neq \{0\}$. By minimality of I_0, we have $I_0 = (x)$. Since $((x)J)J^m = xJ^{m+1} = xJ^m$, it follows that $(x) = (x)J$ by minimality of (x). By Nakayama's lemma, $(x) = \{0\}$, a contradiction. Hence, $J^m = \{0\}$.

Since $a^m \in J^m = \{0\}$ for all $a \in J$, every element of J is nilpotent. But J contains the nilradical of R, so these two ideals are equal.

Let P be a prime ideal P of R. Then P contains the nilradical of R, so it contains J. Thus, P/J is a prime ideal of R/J. By (2), $R/J \cong k_1 \times \cdots \times k_n$ and thus a prime ideal of R/J consists of the elements that are 0 in one of the components. In particular, such a prime ideal is also a maximal ideal. Hence, P is maximal as desired.
(4) Let $M_1, M_2, \ldots, M_n$ be all the distinct maximal ideals of R and let $J = \operatorname{Jac} R$ and $J^m = \{0\}$ as in (3). Then

$$\bigcap_{i=1}^{n} M_i^m \subseteq \left(\bigcap_{i=1}^{n} M_i\right)^m \subseteq J^m = \{0\}.$$

It follows from the Chinese remainder theorem that

$$R \cong R/M_1^m \times R/M_2^m \times \cdots \times R/M_n^m,$$

and each R/M_i^m is an Artinian ring (because R is) with unique maximal ideal M_i/M_i^m.
(5) From (4), it suffices to prove that an Artinian local ring is Noetherian. Assume that R is an Artinian with unique maximal ideal M. Then $\operatorname{Jac} R = M$ and $M^m = \{0\}$ for some $m \in \mathbb{N}$. Thus, the desired result follows from Lemma 4.6.4. □

Corollary 4.6.7. *Every finite commutative ring is a direct product of a finite number of local rings.*

Example 4.6.1. Let $n > 1$. If $n = p_1^{a_1} p_2^{a_2} \ldots p_r^{a_r}$, then

$$\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/p_1^{a_1}\mathbb{Z} \times \mathbb{Z}/p_2^{a_2}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_r^{a_r}\mathbb{Z}.$$

Each $\mathbb{Z}/p_i^{a_i}\mathbb{Z}$ is a local ring with unique maximal ideal $p_i\mathbb{Z}/p_i^{a_i}\mathbb{Z}$ for all $i \in \{1, 2, \ldots, r\}$.

Exercises 4.6.

1. Prove that an Artinain integral domain is a field. Hence, $\mathbb{Z}$ is not Artinian.
2. Suppose $R = F$ is a field. Prove that an R-module M is Artinian if and only if it is Noetherian if and only if M is a finite dimensional vector space over F.
3. Let F be a field and let $f(x)$ be a polynomial in $F[x]$ of degree at least one. Decompose the quotient ring $F[x]/(f(x))$ as a direct product of a finite number of local rings.
4. Let R and S be commutative rings. Prove that $(R \times S)^\times = R^\times \times S^\times$.

4.7 Symplectic Geometry

In this section, we see some applications of module theory, especially a free R-module over commutative rings (§4.2), to obtain a structure theorem for finite dimensional symplectic spaces over a local ring. In addition, we study the symplectic graphs over a commutative ring. This is the work of the author published in Discrete Mathematics [32] and European Journal of Combinatorics [33, 34]. However, we present a new combinatorial approach for obtaining the main theorem (Theorem 4.7.5). In addition, we suggest a nice parallel project on studying the orthogonal graphs over a finite commutative ring.

4.7.1 Symplectic Spaces

A bilinear form and a symplectic form for an R-module are defined analogously as for a vector space.

Let R be a commutative ring and V a left R-module. A **bilinear form** on V is a map $\beta : V \times V \to R$ which is R-linear in both variables. That is,

$$\beta(a\vec{x} + b\vec{y}, \vec{z}) = a\beta(\vec{x}, \vec{z}) + b\beta(\vec{y}, \vec{z}) \text{ and } \beta(\vec{x}, a\vec{y} + b\vec{z}) = a\beta(\vec{x}, \vec{y}) + b\beta(\vec{x}, \vec{z})$$

for all $\vec{x}, \vec{y}, \vec{z} \in V$ and $a, b \in R$. A bilinear form β is called **symplectic** if $\forall \vec{x} \in V, \beta(\vec{x}, \vec{x}) = 0$.

Remark. If a bilinear form β is symplectic then

$$0 = \beta(\vec{x} + \vec{y}, \vec{x} + \vec{y}) = \beta(\vec{x}, \vec{x}) + \beta(\vec{y}, \vec{x}) + \beta(\vec{x}, \vec{y}) + \beta(\vec{y}, \vec{y}) = \beta(\vec{y}, \vec{x}) + \beta(\vec{x}, \vec{y})$$

for all $\vec{x}, \vec{y} \in V$. That is, any symplectic bilinear form is also skew-symmetric.

For an R-submodule W of V, we write $W^{\perp}$ (read "W perp") for the submodule $\{\vec{x} \in V : \forall \vec{w} \in W, \beta(\vec{x}, \vec{w}) = 0\}$, called the **orthogonal complement of** W.

Example 4.7.1. Consider $V = \mathbb{Z}_2$ as a vector space over $\mathbb{Z}_2$. The bilinear form $\beta(x, y) = xy$ is skew-symmetric but not symplectic because $\beta(1, 1) = 1 \neq 0$.

A bilinear form $\beta : V \times V \to R$ is called **regular** or **non-degenerate** or **non-singular** if

1. $\forall f \in \mathrm{Hom}_R(V, R), \exists \vec{x}_0, \vec{y}_0 \in V, f(\vec{x}) = \beta(\vec{x}_0, \vec{x})$ and $f(\vec{x}) = \beta(\vec{x}, \vec{y}_0)$.
2. If $\vec{y} \in V$ and $\forall \vec{x} \in V, \beta(\vec{x}, \vec{y}) = 0$, then $\vec{y} = \vec{0}$. Similarly, if $\vec{y} \in V$ and $\forall \vec{x} \in V, \beta(\vec{y}, \vec{x}) = 0$, then $\vec{y} = \vec{0}$.

Example 4.7.2. Let p be a prime number and let R be the ring of integers modulo p^n, $\mathbb{Z}_{p^n}$, or the field of p^n elements, $\mathbb{F}_{p^n}$, where $n \in \mathbb{N}$. For $\nu \geq 1$, let V denote the set of 2ν-tuples $(a_1, \ldots, a_{2\nu})$ of elements in R. Define $\beta : V \times V \to R$ by the product

$$\beta\big((a_1, \ldots, a_{2\nu}), (b_1, \ldots, b_{2\nu})\big) = (a_1, \ldots, a_{2\nu})\, K\, (b_1, \ldots, b_{2\nu})^t,$$

where $K = \begin{bmatrix} 0 & I_\nu \\ -I_\nu & 0 \end{bmatrix}_{2\nu \times 2\nu}$ and I_ν is the $\nu \times \nu$ identity matrix, for all vectors $(a_1, \ldots, a_{2\nu})$, $(b_1, \ldots, b_{2\nu}) \in V$. Then β is a non-degenerate symplectic bilinear form.

Let R be a commutative ring and V a free R-module of rank n where $n \geq 2$. Let β be a non-degenerate symplectic bilinear form. We call the pair (V, β) a **symplectic space**. An R-module automorphism σ on V is an **isometry** on V if $\beta(\sigma(\vec{x}), \sigma(\vec{y})) = \beta(\vec{x}, \vec{y})$ for all $\vec{x}, \vec{y} \in V$. The group of isometries on V is called the **symplectic group of** (V, β) **over** R and denoted by $\mathrm{Sp}_R(V)$.

Let R be a commutative ring and (V, β) a symplectic space, where V is a free R-module of rank $n \geq 2$. A vector $\vec{x}$ in V is said to be **unimodular** if there is an f in $\mathrm{Hom}_R(V, R)$ with $f(\vec{x}) = 1$; equivalently, if $\vec{x} = \alpha_1\vec{b}_1 + \cdots + \alpha_n\vec{b}_n$, where $\{\vec{b}_1, \ldots, \vec{b}_n\}$ is a basis for V, then the ideal $(\alpha_1, \ldots, \alpha_n) = R$. If $\vec{x}$ is unimodular, then the **line** Rx is a free R-direct summand of rank one.

A **hyperbolic pair** $\{\vec{x}, \vec{y}\}$ is a pair of unimodular vectors in V with the property that $\beta(\vec{x}, \vec{y}) = 1$. The module $H = R\vec{x} \oplus R\vec{y}$ is called a **hyperbolic plane**.

Note that when R is a field, unimodular vectors coincide with nonzero vectors. When R is a local ring, we have a criterion to determine whether a vector in V is unimodular as follows.

Theorem 4.7.1. *Let R be a local ring and (V, β) a symplectic space, where V is a free R-module of rank $n \geq 2$ with basis $\{\vec{e}_1, \ldots, \vec{e}_n\}$. A vector $\vec{x} = a_1\vec{e}_1 + \cdots + a_n\vec{e}_n$ in V is unimodular if and only if a_i is a unit of R for some $i \in \{1, \ldots, n\}$.*

Proof. If some a_i is a unit in R, then $(a_1, \ldots, a_n) = R$, so $\vec{x}$ is unimodular. Conversely, assume that $\vec{x}$ is unimodular. Then there exists an $f \in V^*$ such that $1 = f(\vec{x}) = a_1 f(\vec{e}_1) + \cdots + a_n f(\vec{e}_n)$. Suppose that a_i is not a unit in R for all i. Since R is a local ring, $a_i \in M$ for all i, and thus $a_1 f(\vec{e}_1) + \cdots + a_n f(\vec{e}_n) \in M$. By Theorem 4.6.1 (3), $0 = 1 - (a_1 f(\vec{e}_1) + \cdots + a_n f(\vec{e}_n))$ is a unit in R, which is a contradiction. Therefore, a_i is a unit of R for some $i \in \{1, \ldots, n\}$. □

In addition, if R is a local ring, we show that the rank of symplectic space (V, β) must be even. Let $\{\vec{x}, \vec{y}\}$ be a hyperbolic pair of unimodular vectors in V and $H = R\vec{x} \oplus R\vec{y}$ the corresponding hyperbolic plane. Then for $\vec{z} \in V$, it is easy to see that the vector $\vec{w} = \vec{z} - \beta(\vec{z}, \vec{y})\vec{x} + \beta(\vec{z}, \vec{x})\vec{y}$ is in $H^\perp$, and so $\vec{z}$ can be decomposed as the sum

$$\vec{z} = \vec{w} + (\beta(\vec{z}, \vec{y})\vec{x} - \beta(\vec{z}, \vec{x})\vec{y})$$

of vectors in $H^\perp$ and H, respectively. Since β is non-degenerate, $H \cap H^\perp = \{\vec{0}\}$. Thus, $V = H \oplus H^\perp$.

Notation. If W_1 and W_2 are R-submodule of an R-module V and $\beta(\vec{w}_1, \vec{w}_2) = 0$ for all $\vec{w}_1 \in W_1$ and $\vec{w}_2 \in W_2$, we write $W_1 \perp W_2$.

Moreover, any unimodular vector $\vec{u}$ may be complemented to a hyperbolic pair as follows. First note that there is an $f \in \mathrm{Hom}_R(V, R)$ such that $f(\vec{u}) = 1$. Since β is non-degenerate, there is a $\vec{v} \in V$ with $1 = f(\vec{u}) = \beta(\vec{u}, \vec{v})$. Then $\{\vec{u}, \vec{v}\}$ is a hyperbolic pair. Combining this with the previous observation, we have the first part of the following results.

Theorem 4.7.2. *Let R be a local ring. Let (V, β) be a symplectic space over R of rank ≥ 2. Then V splits as an orthogonal direct sum $V = H \perp H^\perp$ for some hyperbolic plane H. Moreover, $H^\perp$ is a free R-module. Therefore, V is an orthogonal direct sum $V = H_1 \perp H_2 \perp \ldots \perp H_m$ of hyperbolic planes $H_1, H_2, \ldots, H_m$. In particular, the rank of V is even.*

Note that $H^\perp$ is a direct summand of the free module V. By Corollary 4.3.4, it is finitely generated and projective. Then this theorem follows directly from the next lemma.

Lemma 4.7.3. *A finitely generated projective module V over a local ring R is free.*

Proof. Let M be the unique maximal ideal in R. Choose $\vec{v}_1, \ldots, \vec{v}_t \in V$ so that the cosets $\{\vec{v}_1 + MV, \ldots, \vec{v}_t + MV\}$ is a basis for the vector space V/MV over the field R/M. Here, the scalar action is given by $(c + M)(\vec{x} + MV) = c\vec{x} + MV$. Let $\varphi : R^t \to V$ be defined by $\varphi(r_1, \ldots, r_t) = \sum_{i=1}^t r_i \vec{v}_i$. By Remark after Theorem 4.6.1, φ is onto. Since V is projective, $R^t = K \oplus L$ where $K = \ker \varphi$ and $L \cong M$. Then K is finitely generated. Since φ induces an isomorphism from R^t/MR^t to V/MV, it follows that $K/MK \oplus L/ML \cong V/MV$. These are finite dimensional vector spaces over the field R/M. Comparing dimensions yields $K/MK = 0$. Thus, $K = \{\vec{0}\}$ by Nakayama's lemma. Hence, φ is an isomorphism. □

4.7.2 Symplectic Graphs

The general symplectic graph associated with nonsingular alternate matrices over a field is studied by Tang and Wan [42] as a new family of strongly regular graphs. Meemark and Prinyasart [32] introduced the symplectic graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$ for a symplectic space V over a commutative ring R. They showed that their symplectic graph is vertex transitive and arc transitive when $R = \mathbb{Z}_{p^n}$, p is an odd prime and $n \geq 1$. There are many articles influenced by this definition such as [30], [31], [24]. Mostly, the work was on strong regularity, automorphism groups, vertex and arc transitivities, chromatic numbers and subconstituents of symplectic graphs over a finite field, modulo p^n, and modulo pq, where p and q are primes and $n \geq 1$. Recently, Meemark and Puirod [33] studied those topics over finite local rings and obtained results parallel to [42], [32], [30], [31] and [24]. Following [32], we recall the definition of the symplectic graph.

Let R be a commutative ring and (V, β) a symplectic space, where V is a free R-module of rank 2ν, $\nu \geq 1$. Define the graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$ with vertex set is the set of lines $\{R\vec{x} : \vec{x}$ is a unimodular vector in $V\}$ and with adjacency given by

$$R\vec{x} \text{ is adjacent to } R\vec{y} \Leftrightarrow \beta(\vec{x}, \vec{y}) \in R^{\times}.$$

We call $\mathcal{G}_{\mathrm{Sp}_R(V)}$, the **symplectic graph of** (V, β) **over** R.

A **strongly regular graph** with parameters (v, k, λ, μ) is a k-regular graph on v vertices such that for every pair of adjacent vertices there are λ vertices adjacent to both, and for every pair of non-adjacent vertices there are μ vertices adjacent to both.

Let k be a finite field of odd characteristic and let (V', β) be a symplectic space of dimension 2ν where $\nu \geq 1$. Tang and Wan [42] showed that the symplectic graph $\mathcal{G}_{\mathrm{Sp}_k(V')}$ is a strongly regular graph with parameters

$$\left(\frac{|k|^{2\nu} - 1}{|k| - 1}, |k|^{2\nu-1}, |k|^{2\nu-2}(|k| - 1), |k|^{2\nu-2}(|k| - 1)\right).$$

Their proof used orthogonal complements and matrix theory over finite fields.

Let R be a finite local ring with unique maximal ideal M and residue $k = R/M$. Let V be a free R-module of rank 2ν, $\nu \geq 1$, and let V' be the 2ν-dimensional vector space over k induced from V via the canonical map $\pi : R \to k$ given by $\pi : r \mapsto r + M$. Moreover, if (V, β) is a symplectic space, then (V', β') is a symplectic space, where β' is given by

$$\beta'(\pi(\vec{a}), \pi(\vec{b})) = \pi(\beta(\vec{a}, \vec{b}))$$

for all $\vec{a}, \vec{b} \in V$. Here, we write $\pi(\vec{a}) = (\pi(a_1), \pi(a_2), \ldots, \pi(a_{2\nu}))$ for all $\vec{a} = (a_1, a_2, \ldots, a_{2\nu}) \in V$. Note that the relation

$$R\vec{x} \sim R\vec{y} \Leftrightarrow k\pi(\vec{x}) = k\pi(\vec{y}) \tag{4.7.1}$$

is an equivalence relation on the vertex set of the graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$. Since R is a local ring, it follows that

$$\beta(\vec{a}, \vec{b}) \in R^{\times} \Leftrightarrow \pi(\beta(\vec{a}, \vec{b})) \neq M \Leftrightarrow \beta'(\pi(\vec{a}), \pi(\vec{b})) \in k^{\times}.$$

This gives (3) of the next theorem.

Theorem 4.7.4. [Lifting Theorem] *Let R be a finite local ring with unique maximal ideal M and residue $k = R/M$. Let $\kappa = \frac{|k|^{2\nu}-1}{|k|-1}$ and $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\kappa$ be unimodular vectors in V such that the vertex set*

$$\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_k(V')}) = \{k\pi(\vec{x}_i) : i = 1, 2, \ldots, \kappa\}.$$

1. *The set $\Pi = \{R(\vec{x}_1 + M^{2\nu}), R(\vec{x}_2 + M^{2\nu}), \ldots, R(\vec{x}_\kappa + M^{2\nu})\}$ is a partition of $\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_R(V)})$, where $R(\vec{x}_i + M^{2\nu}) = \{R(\vec{x}_i + \vec{m}) : \vec{m} \in M^{2\nu}\}$ for all $\in \{1, 2, \ldots, \kappa\}$. Moreover, for each $i \in \{1, 2, \ldots, \kappa\}$, any two distinct vertices in $R(\vec{x}_i + M^{2\nu})$ are non-adjacent vertices. For each i, the lifting of the vertices corresponding with elements in $k\pi(\vec{x}_i)$ to vertices in $R(\vec{x}_i + M^{2\nu})$ is demonstrated below.*

$\bullet$ $k\pi(\vec{x}_1)$	$\longrightarrow$	$\bullet \quad \bullet \quad \cdots \quad \bullet$ $R(\vec{x}_1 + \vec{m}),\ \vec{m} \in M^{2\nu}$
$\bullet$ $k\pi(\vec{x}_2)$	$\longrightarrow$	$\bullet \quad \bullet \quad \cdots \quad \bullet$ $R(\vec{x}_2 + \vec{m}),\ \vec{m} \in M^{2\nu}$
$\vdots$	$\vdots$	$\vdots$
$\bullet$ $k\pi(\vec{x}_\kappa)$	$\longrightarrow$	$\bullet \quad \bullet \quad \cdots \quad \bullet$ $R(\vec{x}_\kappa + \vec{m}),\ \vec{m} \in M^{2\nu}$

2. *$|R(\vec{x}_i + M^{2\nu})| = |M|^{2\nu-1}$ for all $i \in \{1, \ldots, \kappa\}$.*
3. *For unimodular vectors $\vec{a}, \vec{b} \in V$, we have $R\vec{a}$ and $R\vec{b}$ are adjacent vertices in $\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_R(V)})$ if and only if $k\pi(\vec{a})$ and $k\pi(\vec{b})$ are adjacent vertices in $\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_k(V')})$.*
4. *For $i, j \in \{1, 2, \ldots, \kappa\}$, if $k\pi(\vec{x}_i)$ and $k\pi(\vec{x}_j)$ are adjacent vertices, then $R(\vec{x}_i + \vec{m}_1)$ and $R(\vec{x}_j + \vec{m}_2)$ are adjacent vertices in $\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_R(V)})$ for all $\vec{m}_1, \vec{m}_2 \in M^{2\nu}$.*

Proof. The first part of (1) follows from the relation (4.7.1) and (4) is an immediate consequence of (3). Note that

$$\beta(\vec{x}_i + \vec{m}_1, \vec{x}_i + \vec{m}_2) = \beta(\vec{x}_i, \vec{m}_1) + \beta(\vec{m}_2, \vec{x}_i) + \beta(\vec{m}_1, \vec{m}_2) \in M$$

for all $i \in \{1, 2, \ldots, \kappa\}$ and $\vec{m}_1, \vec{m}_2 \in M^{2\nu}$. This proves the second part of (1).

Next, let $\vec{m}_1, \vec{m}_2 \in M$ and assume that $R(\vec{x}_i + \vec{m}_1) = R(\vec{x}_i + \vec{m}_2)$. Then $\vec{x}_i + \vec{m}_1 = \lambda(\vec{x}_i + \vec{m}_2)$ for some $\lambda \in R^\times$. Thus, $(1 - \lambda)\vec{x}_i = \lambda\vec{m}_2 - \vec{m}_1 \in M^{2\nu}$. Since $\vec{x}_i$ is unimodular, $1 - \lambda \in M$, so $\lambda = 1 + \mu$ for some $\mu \in M$. Hence, $\vec{x}_i + \vec{m}_1 = (1 + \mu)(\vec{x}_i + \vec{m}_2)$. Finally, we show that $R(1 + \mu)(\vec{x} + \vec{m}) = R(\vec{x} + \vec{m})$ for all $\mu \in M$, $\vec{x} \in V$ unimodular, and $\vec{m} \in M^{2\nu}$ and we therefore have (2). Clearly, $R(1 + \mu)(\vec{x} + \vec{m}) \subseteq R(\vec{x} + \vec{m})$. Since $\mu \in M$, $1 + \mu \in R^\times$. Then $r(\vec{x} + \vec{m}) = (r(1 + \mu)^{-1})(1 + \mu)(\vec{x} + \vec{m})$ for all $r \in R$ which gives another inclusion. □

The results for a symplectic graph over a finite local ring are presented in the next theorem. Our proof here is an application of the lifting theorem with some combinatorial arguments. This approach is clean and much difference from the one given in [33]. It explains the reason why we do not have strong regularity clearer and does not involve counting the number of solutions of messy equations like in [32, 33].

Theorem 4.7.5. *Let R be a finite local ring and let (V, β) be a symplectic space of dimension 2ν, where $\nu \geq 1$.*

1. *The symplectic graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$ is $|R|^{2\nu-1}$-regular on $\dfrac{|R|^{2\nu} - |M|^{2\nu}}{|R^\times|}$ many vertices.*
2. *Every two adjacent vertices of $\mathcal{G}_{\mathrm{Sp}_R(V)}$ has $|R|^{2\nu-2}\left|R^\times\right|$ common neighbors.*
3. *Every two non-adjacent vertices of $\mathcal{G}_{\mathrm{Sp}_R(V)}$ has $|R|^{2\nu-2}|R^\times|$ or $|R|^{2\nu-1}$ common neighbors.*

Proof. By Theorem 4.7.4 (1) and (2), the number of vertices of $\mathscr{G}_{\mathrm{Sp}_R(V)}$ is given by

$$\kappa|M|^{2\nu-1} = \frac{|k|^{2\nu}-1}{|k|-1}|M|^{2\nu-1} = \frac{(|k|^{2\nu}-1)|M|^{2\nu}}{(|k|-1)|M|} = \frac{|R|^{2\nu}-|M|^{2\nu}}{|R^\times|}$$

Since the graph $\mathscr{G}_{\mathrm{Sp}_k(V')}$ is $|k|^{2\nu-1}$-regular, Theorem 4.7.4 (3) implies that the graph $\mathscr{G}_{\mathrm{Sp}_R(V)}$ is also regular of degree

$$|k|^{2\nu-1}|M|^{2\nu-1} = |R|^{2\nu-1}.$$

For each pair of adjacent vertices $R(\vec{x}_i+\vec{m}_1)$ and $R(\vec{x}_j+\vec{m}_2)$ in the graph $\mathscr{G}_{\mathrm{Sp}_R(V)}$, the number of common neighbors is given by the product of the common neighbors of vertices $k\pi(\vec{x}_i)$ and $k\pi(\vec{x}_j)$ and $|M|^{2\nu-1}$ by Theorem 4.7.4 (4). Thus,

$$\lambda = |k|^{2\nu-2}(|k|-1)|M|^{2\nu-1} = |R|^{2\nu-2}|R^\times|.$$

Assume that $R(\vec{x}_i+\vec{m}_1)$ and $R(\vec{x}_j+\vec{m}_1)$ are non-adjacent vertices in $\mathscr{G}_{\mathrm{Sp}_R(V)}$. If $i \neq j$, then $k\pi(\vec{x}_i)$ and $k\pi(\vec{x}_j)$ are non-adjacent vertices in $\mathscr{G}_{\mathrm{Sp}_k(V')}$, so the number of common neighbors of $R(\vec{x}_i+\vec{m}_1)$ and $R(\vec{x}_j+\vec{m}_2)$ is the product of common neighbors of $k\pi(\vec{x}_i)$ and $k\pi(\vec{x}_j)$ and $|M|^{2\nu-1}$ which equals $|R|^{2\nu-2}|R^\times|$ by Theorem 4.7.4 (3) and (4). For $i=j$, it is easy to see that the number of common neighbors is the degree of regularity of $\mathscr{G}_{\mathrm{Sp}_R(V)}$. This proves the theorem. □

Let R be a local ring with unique maximal ideal M and let (V,β) be a symplectic space of rank 2ν, where $\nu \geq 1$. By Theorem 4.7.2, V possesses a canonical basis $\{\vec{e}_1,\ldots,\vec{e}_{2\nu}\}$ such that $\{\vec{e}_j,\vec{e}_{\nu+j}\}$ is a hyperbolic pair for all $1 \leq j \leq \nu$ and V is an orthogonal direct sum $V = H_1 \perp H_2 \perp \ldots \perp H_\nu$, where $H_j = R\vec{e}_j \oplus R\vec{e}_{\nu+j}$ is a hyperbolic plane for all $1 \leq j \leq \nu$.

Write unimodular vectors $\vec{a} = a_1\vec{e}_1+\cdots+a_{2\nu}\vec{e}_{2\nu}$ and $\vec{b} = b_1\vec{e}_1+\cdots+b_{2\nu}\vec{e}_{2\nu}$ for some $a_i, b_i \in R$. Then

$$\begin{aligned}\beta(\vec{a},\vec{b}) &= \beta(a_1\vec{e}_1+\ldots+a_{2\nu}\vec{e}_{2\nu}, b_1\vec{e}_1+\ldots+b_{2\nu}\vec{e}_{2\nu})\\ &= \sum_{i=1}^{2\nu}\sum_{j=1}^{2\nu} a_ib_j\beta(\vec{e}_i,\vec{e}_j)\\ &= \sum_{i=1}^{\nu}(a_ib_{\nu+i}-a_{\nu+i}b_i)\end{aligned}$$

because $\beta(\vec{e}_i,\vec{e}_i)=0, \beta(\vec{e}_i,\vec{e}_{\nu+i})=1$ and $\beta(\vec{e}_i,\vec{e}_j) = -\beta(\vec{e}_j,\vec{e}_i)$ for all $i,j \in \{1,\ldots,2\nu\}$. Hence, the adjacency condition becomes

$$R\vec{a} \text{ is adjacent to } R\vec{b} \quad \text{if and only if} \quad \sum_{i=1}^{\nu}(a_ib_{\nu+i}-a_{\nu+i}b_i) \in R^\times.$$

Next, let R be a finite commutative ring. By Corollary 4.6.7, R is a product of finite local rings and we have completely studied our graphs over a finite local ring. Write

$$R = R_1 \times R_2 \times \cdots \times R_t$$

as a direct product of finite local rings R_i, $i = 1,2,\ldots,t$. Consider $V = R^{2\nu}$, a free R-module of rank 2ν, where $\nu \geq 1$. We have the canonical 1-1 correspondence

$$\vec{x} = (x_1,x_2,\ldots,x_{2\nu}) \overset{\varphi}{\mapsto} ((x_1^{(j)})_{j=1}^t,(x_2^{(j)})_{j=1}^t,\ldots,(x_{2\nu}^{(j)})_{j=1}^t).$$

Note that if $\vec{x}, \vec{y} \in V$, then this correspondence induces the symplectic map β on V by

$$\begin{aligned}\beta(\vec{x},\vec{y}) &= \beta\Big(((x_1^{(j)})_{j=1}^t,(x_2^{(j)})_{j=1}^t,\ldots,(x_{2\nu}^{(j)})_{j=1}^t),((y_1^{(j)})_{j=1}^t,(y_2^{(j)})_{j=1}^t,\ldots,(y_{2\nu}^{(j)})_{j=1}^t)\Big)\\ &= (\beta_1(\vec{x}^{(1)},\vec{y}^{(1)}),\beta_2(\vec{x}^{(2)},\vec{y}^{(2)}),\ldots,\beta_t(\vec{x}^{(t)},\vec{y}^{(t)}))\\ &= \Big(\sum_{i=1}^{\nu}(x_i^{(1)}y_{\nu+i}^{(1)}-x_{\nu+i}^{(1)}y_i^{(1)}),\sum_{i=1}^{\nu}(x_i^{(2)}y_{\nu+i}^{(2)}-x_{\nu+i}^{(2)}y_i^{(2)}),\ldots,\sum_{i=1}^{\nu}(x_i^{(t)}y_{\nu+i}^{(t)}-x_{\nu+i}^{(t)}y_i^{(t)})\Big),\end{aligned}$$

where $\vec{x}^{(j)} = (x_1^{(j)}, \vec{x}_2^{(j)}, \ldots, \vec{x}_{2\nu}^{(j)}) \in V^{(j)} := R_j^{2\nu}$ and $(V^{(j)}, \beta_j)$ is a symplectic space of R_j of rank 2ν, for all $j = 1, 2, \ldots, t$. Since $R^\times = R_1^\times \times R_2^\times \times \cdots \times R_t^\times$, we have

$$\beta(\vec{x},\vec{y}) \in R^\times \Leftrightarrow \sum_{i=1}^{\nu}(x_i^{(j)}y_{\nu+i}^{(j)}-x_{\nu+i}^{(j)}y_i^{(j)}) \in R_j^\times \text{ for all } j \in \{1,2,\ldots,t\}. \tag{4.7.2}$$

This shows that the adjacency condition does not depend on the bilinear map β. Recall from the previous paragraph that when R_j is a local ring, the adjacency condition becomes

$$R_j\vec{a} \text{ is adjacent to } R_j\vec{b} \quad \text{if and only if} \quad \sum_{i=1}^{\nu}(a_ib_{\nu+i}-a_{\nu+i}b_i) \in R_j^\times. \tag{4.7.3}$$

for all $j \in \{1, 2, \ldots, t\}$. Therefore, it follows from Eq. (4.7.3) that

$$\mathcal{G}_{\mathrm{Sp}_R(V)} \cong \mathcal{G}_{\mathrm{Sp}_{R_1}(V^{(1)})} \otimes \mathcal{G}_{\mathrm{Span}_{R_2}(V^{(2)})} \otimes \cdots \otimes \mathcal{G}_{\mathrm{Sp}_{R_t}(V^{(t)})}, \tag{4.7.4}$$

as a graph isomorphism. Here, for two graphs G and H, we define their *tensor product* $G \otimes H$ to be the graph with vertex set $\mathcal{V}(G) \times \mathcal{V}(H)$, where (u, v) is adjacent to (u', v') if and only if u is adjacent to u' and v is adjacent to v'.

From Theorem 4.7.5 (1) and the above discussion, we have the number of vertices of $\mathcal{G}_{\mathrm{Sp}_R(V)}$ is equal to

$$|\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_R(V)})| = \prod_{j=1}^{t}|\mathcal{V}(\mathcal{G}_{\mathrm{Sp}_{R_j}(V^{(j)})})| = \prod_{j=1}^{t}\frac{|R_j|^{2\nu}-|M_j|^{2\nu}}{|R_j^\times|}$$

and $\mathcal{G}_{\mathrm{Sp}_R(V)}$ is regular of degree $|R_1|^{2\nu-1}|R_2|^{2\nu-1}\ldots|R_t|^{2\nu-1} = |R|^{2\nu-1}$. Moreover, every two adjacent vertices of $\mathcal{G}_{\mathrm{Sp}_R(V)}$ has $|R|^{2\nu-2}|R^\times|$ common neighbors by Theorem 4.7.5 (2). We record these results in the next theorem.

Theorem 4.7.6. *Let R be a finite commutative ring and (V, β) be the induced symplectic space of rank 2ν, $\nu \geq 1$, discussed above.*

1. *The symplectic graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$ is a $|R|^{2\nu-1}$-regular and isomorphic to the graph*

$$\mathcal{G}_{\mathrm{Sp}_{R_1}(V^{(1)})} \otimes \mathcal{G}_{\mathrm{Sp}_{R_2}(V^{(2)})} \otimes \cdots \otimes \mathcal{G}_{\mathrm{Sp}_{R_t}(V^{(t)})}.$$

2. *Every two adjacent vertices of $\mathcal{G}_{\mathrm{Sp}_R(V)}$ has $|R|^{2\nu-2}|R^\times|$ common neighbors.*

Remark. Other topics for symplectic graphs over a finite commutative ring such as vertex and arc transitivity, automorphism groups and the chromatic number can be found in the following exercises and [32, 33].

Exercises 4.7. 1. Let R be a local ring with unique maximal ideal M and let (V, β) be a symplectic space of R-dimension 2ν, where $\nu \geq 1$. Let $\vec{a} = a_1\vec{e}_1 + \cdots + a_{2\nu}\vec{e}_{2\nu}$ and $\vec{b} = b_1\vec{e}_1 + \cdots + b_{2\nu}\vec{e}_{2\nu}$ be unimodular vectors in V and assume that $a_i \in R^\times$ for some $i \in \{1, \ldots, 2\nu\}$. If $R\vec{a}$ is adjacent to $R\vec{b}$, prove that $a_ib_l - a_lb_i$ is a unit for some $l \in \{1, \ldots, 2\nu\}$ and $l \neq i$.

2. The **chromatic number** of a graph G is the smallest number of colors needed to color the vertices of G so that no two adjacent vertices share the same color. Tang and Wan [42] showed that if k is the field of q elements and V' is the symplectic space of dimension 2ν, $\nu \geq 1$, then the chromatic number of the symplectic graph $\mathcal{G}_{\mathrm{Sp}_k(V')}$ is $q^\nu + 1$. Let R be a local ring with unique maximal ideal M and residue field k and let (V, β) be a symplectic space of R-dimension 2ν, where $\nu \geq 1$. Determine the chromatic number of the symplectic graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$. (*Hint*. Use the lifting theorem.)
3. Let G and H be graphs. A function σ from $\mathcal{V}(G)$ to $\mathcal{V}(H)$ is a **homomorphism** from G to H if $\sigma(g_1)$ and $\sigma(g_2)$ are adjacent in H whenever g_1 and g_2 are adjacent in G. It is called an **isomorphism** if it is a bijection and σ^{-1} is a homomorphism from H onto G. Moreover, an isomorphism on G is called an **automorphism**. The set of all automorphisms of a graph G is denoted by $\mathrm{Aut}\,(G)$. It is a group under composition, called the **automorphism group of** G. Prove that for graphs G and H, $\mathrm{Aut}\,(G) \times \mathrm{Aut}\,(H) \subseteq \mathrm{Aut}\,(G \otimes H)$.
4. A graph G is **vertex transitive** if its automorphism group acts transitively on the vertex set. That is, for any two vertices of G, there is an automorphism carrying one to the other. An **arc** in G is an ordered pair of adjacent vertices, and G is **arc transitive** if its automorphism group acts transitively on its arcs. Klingenberg [28] showed that for a local ring R, if $\{\vec{x}, \vec{a}\}$ and $\{\vec{x}, \vec{b}\}$ are hyperbolic pairs of unimodular vectors in V, then there exists an isometry σ in $\mathrm{Sp}_R(V)$ which leaves $\vec{x}$ invariant and carries $\vec{a}$ to $\vec{b}$. Let R be a finite local ring and let (V, β) be a symplectic space of dimension 2ν. Show that
   (a) $\mathrm{Sp}_R(V)$ acts transitively on unimodular vectors and on hyperbolic planes.
   (b) The symplectic graph $\mathcal{G}_{\mathrm{Sp}_R(V)}$ is vertex transitive and arc transitive.

   Show further that (b) holds for any finite commutative ring R. (*Hint*. Use 3.)

Project 23 (Orthogonal graphs)**.** Similar to symplectic graphs, we may study orthogonal graphs over a finite commutative ring defined as follows.

Let R be a commutative ring and let V be a free R-module of rank n, where $n \geq 2$. Assume that we have a function $\beta : V \times V \to R$ which is R-bilinear, symmetric and the $R-$ module morphism from V to $V^* = \hom_R(V, R)$ given by $\vec{x} \mapsto \beta(\cdot, \vec{x})$ is an isomorphism. For $\vec{x} \in V$, we call $\beta(\vec{x}, \vec{x})$ the **norm of** $\vec{x}$. The pair (V, β) is called an **orthogonal space**.

Let R be a commutative ring and let (V, β) be an orthogonal space, where V is a free R-module of rank $n \geq 2$. A vector $\vec{x}$ in V is said to be **unimodular** if there is an f in $\hom_R(V, R)$ with $f(\vec{x}) = 1$; equivalently, if $\vec{x} = \alpha_1\vec{b}_1 + \ldots + \alpha_n\vec{b}_n$, where $\{\vec{b}_1, \ldots, \vec{b}_n\}$ is a basis for V, then the ideal $(\alpha_1, \ldots, \alpha_n) = R$. If $\vec{x}$ is unimodular, then the **line** $R\vec{x}$ is free R-direct summand of rank one. Moreover, it is easy to see that if $\vec{x}$ and $\vec{y}$ are unimodular vectors in V, then $R\vec{x} = R\vec{y}$ if and only if $\vec{x} = \lambda\vec{y}$ for some $\lambda \in R^\times$.

Define the graph $\mathscr{G}_{\mathrm{O}_R(V)}$ whose vertex set $\mathcal{V}(\mathscr{G}_{\mathrm{O}_R(V)})$ is the set of lines

$$\{R\vec{x} : \vec{x} \text{ is a unimodular vector in } V \text{ and } \beta(\vec{x}, \vec{x}) = 0\}$$

and its adjacency condition is given by

$$R\vec{x} \text{ is adjacent to } R\vec{y} \Longleftrightarrow \beta(\vec{x}, \vec{y}) \in R^\times (\text{or equivalently}, \beta(\vec{x}, \vec{y}) = 1).$$

We call $\mathscr{G}_{\mathrm{O}_R(V)}$ the **orthogonal graph of** (V, β) **over** R.

(a) Show that the above adjacency condition is well defined.

(b) If k is a finite field of odd characteristic and V'_δ is an orthogonal space over k of dimension $2\nu + \delta$, where $\nu \geq 1$ and $\delta \in \{0, 1, 2\}$, then Gu and Wan [23] showed that $\mathcal{G}_{\mathrm{O}_k(V'_\delta)}$ is a $|k|^{\nu+\delta-1} + 1$-partite graph with partite sets $X_1, X_2, \ldots, X_{|k|^{\nu+\delta-1}+1}$ such that $|X_i| = \frac{|k|^\nu - 1}{|k| - 1}$ for all $i \in \{1, 2, \ldots, |k|^{\nu+\delta-1} + 1\}$. They also had that (Theorem 2.1 of [23]) the orthogonal graph $\mathscr{G}_{\mathrm{O}_k(V'_\delta)}$ is $|k|^{2\nu+\delta-2}$-regular on $\dfrac{(|k|^\nu - 1)(|k|^{\nu+\delta-1} + 1)}{|k| - 1}$ many vertices. Moreover, if $\nu = 1$, then it is a complete graph, and if $\nu \geq 2$, then the graph is a strongly regular graph with parameters

$$\lambda = |k|^{2\nu+\delta-2} - |k|^{2\nu+\delta-3} - |k|^{\nu-1} + |k|^{\nu+\delta-2} \text{ and } \mu = |k|^{2\nu+\delta-2} - |k|^{2\nu+\delta-3},$$

respectively. Construct the lifting theorem for orthogonal graphs and determine the results similar to Theorem 4.7.5 for orthogonal graphs over a finite local ring.

(c) Let R be a finite commutative ring of odd characteristic and (V_δ, β) be an orthogonal space of rank $2\nu+\delta$, $\nu \geq 1$ and $\delta \in \{0,1,2\}$. Write $R = R_1 \times R_2 \times \cdots \times R_t$ as a direct product of finite local rings R_i, $i = 1, 2, \ldots, t$. Prove that the orthogonal graph $\mathscr{G}_{\mathrm{O}_R(V_\delta)}$ is a $|R|^{2\nu-2+\delta}$-regular and isomorphic to the graph

$$\mathscr{G}_{\mathrm{O}_{R_1}(V_\delta^{(1)})} \otimes \mathscr{G}_{\mathrm{O}_{R_2}(V_\delta^{(2)})} \otimes \cdots \otimes \mathscr{G}_{\mathrm{O}_{R_t}(V_\delta^{(t)})}.$$

5 | Field Theory

In Section 2.6, we learn about extensions of a field. Here, we give more details on a construction of extension fields. We prepare the readers to Galois theory which yields a connection between field theory and group theory. Applications of Galois theory are provided in proving fundamental theorem of algebra, finite fields, and cyclotomic fields. We discuss some results on a transcendental extension in the final section.

5.1 Splitting Fields

Let F be a field. Given a polynomial $f(x) \in F[x]$ we would like to have at hand an extension field E of F which in some sense contains all the roots of the equation $f(x) = 0$. We recall that $f(r) = 0$ if and only if $f(x)$ is divisible by $x - r$.

We say that $f(x)$ **splits** in an extension field E if $f(x) = \prod_{i=1}^{n} c(x - r_i)$, that is, it is a product of linear factors in $E[x]$ and $c \in F$. We shall first study some facts about the roots of $f(x) \in F[x]$ as follows.

Theorem 5.1.1. *If $f(x) \in F[x]$ and $\deg f(x) = n \geq 1$, then $f(x)$ can have at most n roots counting multiplicities in any extension field of F.*

Proof. We shall prove the theorem by induction on the degree of $f(x)$. If $\deg f(x) = 1$, then $f(x) = ax + b$ for some $a, b \in F$ and $a \neq 0$. Then $-b/a$ is the unique root of $f(x)$ and $-b/a \in F$, so we are done.

Let $\deg f(x) = n > 1$ and assume that the result is true for all polynomials of degree $< n$. Let E be any extension field of F. If $f(x)$ has no roots in E, then we are done. Let $r \in E$ be a root of $f(x)$ of multiplicity $m \geq 1$. Then there exists $q(x) \in E[x]$ such that $f(x) = (x - r)^m q(x)$ and $q(r) \neq 0$. Thus, $\deg q(x) = n - m$. By the inductive hypothesis $q(x)$ has at most $n - m$ roots in E counting multiplicities. Hence, $f(x)$ has at most $m+(n-m)$ roots in E counting multiplicities. □

Theorem 5.1.2. [Kronocker] *If $p(t) \in F[t]$ is irreducible over F, then there exists an extension field E of F such that $[E : F] = \deg p(t)$ and $p(t)$ has a root in E.*

Proof. Let $E = F[x]/(p(x))$ where x is an indeterminate. Then E is a field containing $\{a + (p(x)) : a \in F\}$ as a subfield. But $F \cong \{a + (p(x)) : a \in F\}$ by $\varphi : a \mapsto a + (p(x))$, so E can be considered as an extension field of F by considering a as $a + (p(x))$ for all $a \in F$. Then $E = F[x]/(p(x)) = F(\bar{t})$ where $\bar{t} = x + (p(x))$ is a root of $p(t)$. Since $E = F(\bar{t})$ and $p(t)$ is irreducible over F, $[E : F] = [F(\bar{t}) : F] = \deg p(t)$ by Corollary 2.6.5. □

Corollary 5.1.3. *If $p(t) \in F[t]$ is a nonconstant polynomial, then there exists a finite extension field E of F containing a root of $p(t)$ and $[E : F] \leq \deg p(t)$.*

Proof. Since $F[t]$ is a UFD, $p(t)$ has an irreducible factor in $F[t]$ say $p_1(t)$. By Theorem 5.1.2, there exists an extension field E of F such that E contains a root of $p_1(t)$ and $[E : F] = \deg p_1(t)$. Hence, $[E : F] \leq \deg p(t)$ and E contains a root of $p(t)$. □

Let F be a field and $f(x)$ a monic polynomial in $F[x]$. An extension field E of F is a **splitting field** of $f(x)$ over F if

$$f(x) = (x - r_1) \dots (x - r_n)$$

in $E[x]$ and

$$E = F(r_1, \dots, r_n),$$

that is, E is generated by the roots of $f(x)$. The next results demonstrate the existence of a splitting field for a monic polynomial.

Theorem 5.1.4. [Existence of Splitting Fields] *Let $f(x)$ be a monic polynomial of degree $n \geq 1$. Then there exists an extension field E of F such that $[E : F] \leq n!$ and E contains n roots of $f(x)$ counting multiplicities. Hence, in $E[t]$, $f(x) = c(x - r_1) \dots (x - r_n)$ for some $c \in F$ and $r_1, \dots, r_n \in E$, so that $r_1, \dots, r_n$ are n roots of $f(x)$ in E.*

Proof. We shall prove the theorem by induction on the degree of $f(x)$. If $\deg f(x) = 1$, then $f(x)$ has exactly one root in F and $[F : F] = 1 = 1!$.

Let $\deg f(x) = n > 1$ and assume that the theorem is true for the case of polynomials of degree $< n$. By Corollary 5.1.3, there exists an extension field E_0 of F such that $f(x)$ has a root, say $r \in E_0$ and $[E_0 : F] \leq n$. Since r is a root of $f(x)$, $f(x) = (x - r)q(x)$ for some $q(x) \in E_0[x]$, so $\deg q(x) = n - 1$. By the inductive hypothesis, there exists an extension field E of E_0 such that $[E : E_0] \leq (n - 1)!$ and E contains $n - 1$ roots of $q(x)$. Then E is an extension field of F, $[E : F] = [E : E_0][E_0 : F] \leq n!$ and E contains n roots of $f(x)$ counting multiplicities. □

Corollary 5.1.5. *Let F be a field and $f(x)$ a nonconstant polynomial over F of degree n. Then there exists a splitting field E of $f(x)$ over F. Moreover, $[E : F] \leq n!$.*

Proof. We have seen from Theorem 5.1.4 that there exists an extension field E of F such that $f(x) = c(x - r_1) \dots (x - r_n)$, for some $c \in F$ and $r_1, \dots, r_n \in E$, is a product of linear factors in $E[x]$ and $[E : F] \leq n!$. Hence, $E = F(r_1, \dots, r_n)$ is a desired field. □

Examples 5.1.1 (Examples of splitting fields). 1. Let $f(x) = x^2 + ax + b$. If $f(x)$ is reducible in $F[x]$ (F arbitrary) then F is a splitting field. Otherwise, put $E = F[x]/(f(x)) = F(r_1)$ where $r_1 = x + (f(x))$. Then E is a splitting field since $f(r_1) = 0$, so $f(x) = (x - r_1)(x - r_2)$ in $E[x]$. Thus, $E = F(r_1) = F(r_1, r_2)$. Since $f(x)$ is the minimal polynomial of r_1 over F, $[E : F] = 2$.

2. Let the base field F be $\mathbb{Z}/(2)$, the field of two elements, and let $f(x) = x^3 + x + 1$. Since $1+1+1 \neq 0$ and $0+0+1 \neq 0$, $f(x)$ has no roots in F; hence $f(x)$ is irreducible in $F[x]$. Put $r_1 = x + (f(x))$ in $F[x]/(f(x))$ so $F(r_1)$ is a field and $x^3 + x + 1 = (x + r_1)(x^2 + ax + b)$ in $F(r_1)[x]$. (Note that we can write $+$ for $-$ since characteristic is two.) Comparison of coefficients shows that $a = r_1$, $b = 1 + r_1^2$. The elements of $F(r_1)$ can be listed as $c+dr_1+er_1^2, c, d, e \in F$. There are eight of these: $0, 1, r_1, 1+r_1, r_1^2, 1+r_1^2, r_1+r_1^2$ and $1+r_1+r_1^2$. Substituting these in $x^2+r_1x+1+r_1^2$, we reach $(r_1^2)^2+r_1(r_1^2)+1+r_1^2 = r_1^4+r_1^3+1+r_1^2 = 0$ since $r_1^3 = r_1 + 1$ and $r_1^4 = r_1^2 + r_1$. Hence, $x^2 + ax + b$ factors into linear factors in $F(r_1)[x]$ and $E = F(r_1)$ is a splitting field of $x^3 + x + 1$ over F.

3. Let $F = \mathbb{Q}$, $f(x) = (x^2 - 2)(x^2 - 3)$. Since the rational roots of $x^2 - 2$ and $x^2 - 3$ must be integral, it follows that $x^2 - 2$ and $x^2 - 3$ are irreducible in $\mathbb{Q}[x]$. Form $K = \mathbb{Q}(r_1), r_1 = x + (x^2 - 2)$ in $\mathbb{Q}[x]/(x^2 - 2)$. The elements of K have the form $a + br_1, a, b \in \mathbb{Q}$. We claim that $x^2 - 3$ is irreducible in $K[x]$. Otherwise, we have rational numbers a, b such that

$(a+br_1)^2=3$. Then $(a^2+2b^2)+2abr_1=3$ so that $ab=0$ and $a^2+2b^2=3$. If $b=0$ we obtain $a^2=3$ which is impossible since $\sqrt{3}$ is not rational, and if $a=0, b^2=3/2$. Then $(2b^2)=6$ and since $\sqrt{6}$ is not rational, we again obtain an impossibility. Thus, x^2-3 is irreducible in $K[x]$. Now form $E=K[x]/(x^2-3)$. Then this is a splitting field over $\mathbb{Q}$ of $(x^2-2)(x^2-3)$ and $[E:\mathbb{Q}]=[E:K][K:\mathbb{Q}]=2\cdot 2=4$.

4. Let $F=\mathbb{Q}, f(x)=x^p-1$, p a prime. We have $x^p-1=(x-1)(x^{p-1}+x^{p-2}+\cdots+x+1)$ and we know that $x^{p-1}+x^{p-2}+\cdots+x+1$ is irreducible in $\mathbb{Q}[x]$. Let $E=\mathbb{Q}(z)$ where $z=x+(x^{p-1}+x^{p-2}+\cdots+x+1)$ in $\mathbb{Q}[x]/(x^{p-1}+x^{p-2}+\cdots+x+1)$. We have $1, z, \ldots, z^{p-1}$ are distinct. Also $(z^k)^p=(z^p)^k=1$ so every z^k is a root of x^p-1. It follows that $x^p-1=\prod_{k=1}^{p}(x-z^k)$ in $E[x]$. Thus, E is a splitting field over $\mathbb{Q}$ of x^p-1 and $[E:\mathbb{Q}]=p-1$.
5. Since $x^3-2=(x-\sqrt[3]{2})(x-\sqrt[3]{2}\omega)(x-\sqrt[3]{2}\omega^2)$ where $\omega\neq 1$ and $\omega^3=1$, $\mathbb{Q}(\sqrt[3]{2})$ is not a splitting field of $f(x)=x^3-2$ over $\mathbb{Q}$. A splitting field of $f(x)$ is $E=\mathbb{Q}(\sqrt[3]{2},\omega)$. Since $g(x)=x^2+x+1$ is irreducible over $\mathbb{Q}(\sqrt[3]{2})$ and $g(\omega)=0$, $[E:\mathbb{Q}(\sqrt[3]{2})]=2$, so $[E:F]=[E:\mathbb{Q}(\sqrt[3]{2})][\mathbb{Q}(\sqrt[3]{2}):\mathbb{Q}]=2\cdot 3=6$.
6. A splitting field over $\mathbb{Z}/(p)$ of $x^{p^e}-1$, $e\in\mathbb{N}$, is $\mathbb{Z}/(p)$.

Theorem 5.1.6. [Uniqueness of Splitting Fields] *Let $\eta: F\to F_1$ be an isomorphism of fields and let $\eta: F[x]\to F_1[x]$ be the isomorphism which extends η and satisfies $\eta(x)=x$. Suppose $f(x)$ is a monic polynomial in $F[x]$, let $f_1(x)=\eta(f(x))$ and suppose that E/F and E_1/F_1 are splitting fields of $f(x)$ and $f_1(x)$, respectively. Then there exists an isomorphism $\eta^*: E\to E_1$ which extends η.*

Proof. Let $\hat{f}(x)$ be an irreducible factor of $f(x)$ and let $\hat{f}_1(x)=\eta(\hat{f}(x))$. Let $r\in E$ be a root of $\hat{f}(x)$ and let $r_1\in E_1$ be a root of $\hat{f}_1(x)$. Then we have a commutative diagram in which the vertical arrows are isomorphisms and the horizontal arrows are inclusion maps

$$\begin{array}{ccccc}
F & \longrightarrow & F[r] & \longrightarrow & E \\
 & & \uparrow{\scriptstyle i} & & \\
 & & F[x]/\hat{f}(x)F[x] & & \\
\eta\downarrow & & \downarrow{\scriptstyle \hat{\eta}} & & \\
 & & F_1[x]/\hat{f}_1(x)F_1[x] & & \\
 & & \downarrow{\scriptstyle j} & & \\
F_1 & \longrightarrow & F_1[r_1] & \longrightarrow & E_1.
\end{array}$$

The map $j\hat{\eta}i^{-1}=\bar{\eta}$ is an isomorphism of fields extending η. Also, $\bar{\eta}(f(x)/(x-r))=f_1(x)/(x-r)$ and $E/F[r]$, $E_1/F_1[r_1]$ are splitting fields of $f(x)/(x-r)$ and $f_1(x)/(x-r_1)$, respectively.

Now, by induction on $\deg f(x)$, $\bar{\eta}: F[r]\to F_1[r_1]$ has an extension to $\eta^*: E\to E_1$ and this is the required extension of η. □

Theorem 5.1.7. *Assume $f(x)$ has no multiple factors as an element of $F[x]$. Under the hypothesis of Theorem 5.1.6, the number of distinct extensions of $\eta: F\to F_1$ to $\eta: E\to E_1$ is at most $[E:F]$. Moreover, the number of distinct extensions is equal to $[E:F]$ if and only if $f(x)$ has distinct roots in E.*

Proof. Proceeding as in the proof of Theorem 5.1.6, let $\hat{f}(x)$ be an irreducible factor of $f(x)$, let d be the degree of $\hat{f}(x)$, let $\hat{f}_1(x)=\eta(\hat{f}(x))$, let $r_1,\ldots,r_e$ be the distinct roots of $\hat{f}(x)$ in E and let $r_1',\ldots,r_e'$ be the roots of $\hat{f}_1(x)$ in E_1. (Note that $e\leq d$ and $e=d$ if $\hat{f}_1(x)$ has no multiple roots, but this is not always the case.)

Next fix a root $r = r_1$ of $\hat{f}(x)$. The argument of Theorem 5.1.6 shows that for each root $r'_1, \ldots, r'_e$ of $\hat{f}_1(x)$ there is an isomorphism $\bar{\eta}_j : F[r] \to F_1[r'_j]$ extending η, where $\hat{\eta}_j(r) = r'_j$.

$$\begin{array}{ccccc} F & \longrightarrow & F[r] & & \\ \eta\downarrow & & & & \\ F_1 & \longrightarrow & F_1[r'_j] & \hookrightarrow & E_1 \end{array}$$

On the other hand, any isomorphism of $F[r]$ *into* E_1 must carry r into a root of $\hat{f}_1(x)$, and so must one of the $\bar{\eta}_j$. Furthermore, as noted above

$$\text{the number of roots of } \hat{f}(x) = e \leq d = [F[r] : F].$$

By induction, the number of ways each $\hat{\eta}_j$ can be extended to an isomorphism $E \to E_1$ is at most $[E : F[r]]$. Thus,

$$\begin{aligned} &\text{the number of extensions of } \eta : F \to F_1 \text{ to } \eta^* : E \to E_1 \\ &\qquad\qquad \leq e[E : F[r]] \leq [F[r] : F][E : F[r]] = [E : F]. \end{aligned}$$

Now we want to answer the question: When is there equality – that is, the number of extensions = $[E : F]$?

Looking at the first step above we see that the number of roots of $\hat{f}(x) = e = d = [F[r] : F]$ if and only if $\hat{f}(x)$ has $d = \deg \hat{f}(x)$ roots – that is if and only if $\hat{f}(x)$ has distinct roots.

To continue inductively, we now have the set up

$$\begin{array}{ccc} F[r] & \longrightarrow & E \\ \hat{\eta}_j\downarrow & & \\ F_1[r'_j] & \longrightarrow & E_1 \end{array}$$

The key point is that E is the splitting field over $F[r]$ of the polynomial $f(x)/(x - r)$. This polynomial has no multiple factor so inductively the number of extensions of $\hat{\eta}_j$ to an isomorphism $\eta^* : E \to E_1$ is equal to $[E : F[r]]$ if and only if $f(x)/(x-r)$ has distinct roots. Combining this with the result for $\hat{f}(x)$ we get the number of extensions of $\eta : F \to F_1$ to an isomorphism $\eta : E \to E_1$ is equal to $[E : F]$ if and only if $f(x)$ has distinct roots in E. □

Remarks. (1) If $f(x)$ is an irreducible polynomial over a field F and r is a root of $f(x)$ in some extension field of F, then

$$F[x]/f(x)F[x] \cong F[r].$$

However, if $f(x) = g(x)h(x)$ where $g(x)$ and $h(x)$ are irreducible polynomials, then by Chinese remainder theorem

$$F[x]/f(x)F[x] \cong F[x]/g(x)F[x] \times F[x]/h(x)F[x]$$

a direct product of fields. If $f(x) = g(x)^2$, then $F[x]/f(x)F[x]$ even has nilpotent elements.

In general, E/F arises from a succession of simple extensions

$$\begin{aligned} F \subseteq F_1 &\cong F[x]/f_1(x)F[x], \\ F_1 \subseteq F_2 &\cong F_1[x]/f_2(x)F_1[x], \\ &\vdots \\ F_{r-1} \subseteq F_r &\cong F_{r-1}[x]/f_r(x)F_{r-1}[x] = E. \end{aligned}$$

We shall see that in some important cases (but not all), the splitting field E/F of the polynomial $f(x)$ can be achieved as a simple extension $F \subseteq F[x]/g(x)F[x] = E$, but usually $g(x) \neq f(x)$.
(2) If $f(x)$ and $g(x)$ have the same roots in some extension field E of F ($f(x), g(x) \in F[x]$), then they have the same splitting field. However, one cannot guarantee that the roots of $f(x)$ are distinct (or simple, or one fold). The basic example is the polynomial

$$f(x) = x^p - a \in F[a]$$

where F is a field of characteristic $p > 0$. If r is a root of $f(x)$ in some extension field E of $F[a]$, then $r^p = a$ and the factorization of $f(x)$ in $E[x]$ is

$$f(x) = x^p - a = x^p - r^p = (x - r)^p.$$

Exercises 5.1.
1. Construct a splitting field over $\mathbb{Q}$ of $x^5 - 2$. Find its dimension over $\mathbb{Q}$.
2. Let $f(x) = x^4 + x^2 + 1$. Find the splitting field of $f(x)$ over $\mathbb{Q}$ and determine its dimension.
3. Let E/F be a splitting field of over F of $f(x)$ and let K be a subfield of E/F. Show that any monomorphism of K/F into E/F can be extended to an automorphism of E.
4. If $f(x) \in F[x]$ has degree n and K is a splitting field of $f(x)$ over F, prove that $[K : F]$ divides $n!$.
5. Let F be a field of characteristic $p > 0$ and let $b \in F$. Show that either $x^p - b$ is irreducible in $F[x]$ or $b = a^p$ and $x^p - b = (x - a)^p$ for some $a \in F$.

5.2 Algebraic Closure of a Field

We know about the *prime field* which is the smallest field such that every other field is an extension of it. However, we does not know if we can algebraically extend our field F forever to obtain a field that every polynomial in $F[x]$ has a root in it. We shall assure it in this section.

A field F is called **algebraically closed** if every monic polynomial $f(x)$ of positive degree with coefficients in F has a root in F.

Theorem 5.2.1. *Let F be a field. The following statements are equivalent.*
(i) F is algebraically closed.
(ii) An irreducible polynomial in $F[x]$ is linear, and hence every polynomial of $F[x]$ of positive degree is a product of linear factors.
(iii) F has no proper algebraic extension field.

Proof. Since r is a root, that is $f(r) = 0$, if and only if $x - r$ is a factor of $f(x)$ in $F[x]$, we have (i) $\Leftrightarrow$ (ii). Next, we show (i) $\Leftrightarrow$ (iii). If E is an extension field of F and $a \in E$ is algebraic over F, then $[F(a) : F]$ is the degree of the minimal polynomial $f(x)$ of a over F, and $f(x)$ is monic and irreducible. Then $a \in F$ if and only if $\deg f(x) = 1$. Hence, E is algebraic over F and $E \supset F$ implies there exist irreducible monic polynomials in $F[x]$ of degree ≥ 2; hence F is not algebraically closed. Conversely, if F is not algebraically closed, then there exists a monic irreducible $f(x) \in F[x]$ with $\deg f(x) \geq 2$, Thus, the field $F[x]/(f(x))$ is a proper algebraic extension of F. □

We recall that (Corollary 2.6.7) if E is an extension field of the field F, then the set of elements of E that are algebraic over F constitutes a subfield A of E/F (that is, a subfield of E containing F). Evidently $E = A$ if and only if E is algebraic over F. At the other extreme, if $A = F$, then F is said to be **algebraically closed in** E. In any case A is algebraically closed in E, since any element of E that is algebraic over A is algebraic over F and so is contained in A. This result shows that if a field F has an algebraically closed extension field, then it has one that is algebraic

over F. We call an extension field E/F an **algebraic closure of** F if E is algebraic over F and E is algebraically closed.

For example, assuming the truth of the fundamental theorem of algebra (Theorem 5.5.6), that $\mathbb{C}$ is algebraically closed, it follows that the field of A of algebraic numbers is an algebraic closure of $\mathbb{Q}$, and thus A is algebraically closed.

We proceed to prove the existence and uniqueness up to isomorphism of an algebraic closure of any field F. For a countable F a straightforward argument is available to establish these results. We begin by enumerating the monic polynomials of positive degree as $f_1(x), f_2(x), \ldots$. Evidently this can be done. We now define inductively a sequence of extension fields beginning with $F_0 = F$ and letting F_i be a splitting field over F_{i-1} of $f_i(x)$. The construction of such splitting fields was given at the end of the previous section. It is clear that every F_i is countable, so we can realize all of these constructions in some large *set* S. Then we can take $E = \bigcup F_i$ in the set. Alternatively we can define E to be a direct limit of the fields F_i. It is easily seen that E is an algebraic closure of F. We showed that (Theorem 5.1.6) there exists an isomorphism of K_1/F onto K_2/F. This can be used to prove the isomorphism theorem for algebraic closures of a countable field by a simple inductive argument.

The pattern of the proof sketched above can be carried over to the general case by using "transfinite induction". This is what was done by E. Steinitz, who first proved these results. There are several alternative proofs available that are based on Zorn's lemma. We shall give one that makes use of the following lemma.

Lemma 5.2.2. *If E is an algebraic extension of a field F, then the cardinality of E cannot exceed the cardinality of $F[x]$.*

Proof. Let S be the set of all ordered pairs (f, α) where $f(x) \in F[x]$ is nonzero and $\alpha \in E$ with $f(\alpha) = 0$. Since for each polynomial $f(x)$, the number of α such that (f, α) lies in S is finite, we have $|S| \leq |F[x]|\aleph_0 = |F[x]|$. On the other hand, E maps injectively into S via $\alpha \mapsto (f_\alpha, \alpha)$ where f_α is the minimal polynomial of α, and thus $|E| \leq |S|$. □

Recall that $|F[x]| = |F|\aleph_0$. If F is infinite, then $|F[x]| = |F|$ and it follows that $|E| = |F|$. When F is finite, $F[x]$ is countable, and hence E is either finite or countably infinite.

Corollary 5.2.3. *There exist real numbers transcendental over $\mathbb{Q}$.*

Proof. There are only countably many polynomials in $\mathbb{Q}[x]$. Since $\mathbb{R}$ is uncountable, the above lemma guarantees that $\mathbb{R}$ is not algebraic over $\mathbb{Q}$. □

We can now prove the existence of algebraic closures.

Theorem 5.2.4. *Any field F has an algebraic closure.*

Proof. We first embed F in a set S in which we have a lot of elbow room. Precisely, we assume that $|S| > |F|$ if F is infinite and that S is uncountable if F is finite. We now define a set Λ whose elements are $(E, +, \cdot)$ where E is a subset of S containing F and $+, \cdot$ are binary compositions in E such that $(E, +, \cdot)$ is an algebraic extension field of F. We partially order Λ by declaring that $(E, +, \cdot) > (E', +', \cdot')$ if E is an extension field of E'. By Zorn's lemma there exists a maximal element $(E, +, \cdot)$. Then E is an algebraic extension of F. We claim that E is algebraically closed. Otherwise we have a proper algebraic extension $E' = E(a)$ of E. Then $|E'| < |S|$, so we can define an injective map of E' into S that is the identity on E and then we can transfer the addition and multiplication on E' to its image. This gives an element of Λ bigger than $(E, +, \cdot)$. This contradiction shows that E is an algebraic closure of F. □

Next we take up the question of uniqueness of algebraic closures. It is useful to generalize the concept of a splitting field of a polynomial to apply to sets of polynomials.

If $\Gamma = \{f_\alpha(x)\}$ is a set of monic polynomials with coefficients in F, then an extension field E/F is called a **splitting field over F of the set Γ** if

1. every $f_\alpha(x) \in \Gamma$ is a product of linear factors in $E[x]$ and
2. E is generated over F by the roots of the $f_\alpha(x) \in \Gamma$.

It is clear that if E is a splitting field over F of Γ, then no proper subfield of E/F is a splitting field of Γ and if K is any intermediate field, then E is a splitting field of Γ. Since an algebraic closure E of F is algebraic, it is clear that E is a splitting field over F of the complete set of monic polynomials of positive degree in $F[x]$. The isomorphism theorem for algebraic closures will therefore be a consequence of a general result on isomorphisms of splitting fields that we shall now prove. Our starting point is the following result, which is Theorem 5.1.6.

Let $\eta : a \mapsto \tilde{a}$ be an isomorphism of a field F onto a field $\tilde{F}$, $f(x) \in F[x]$ be monic of positive degree, $\tilde{f}(x)$ the corresponding polynomial in $\tilde{F}[x]$ (under the isomorphism, which is η on F and sends $x \mapsto x$), and let E and $\tilde{E}$ be splitting fields over F and $\tilde{F}$ of $f(x)$ and $\tilde{f}(x)$, respectively. Then η can be extended to an isomorphism of E onto $\tilde{E}$.

We shall now extend this to a set of polynomials.

Theorem 5.2.5. *Let $\eta : a \mapsto \tilde{a}$ be an isomorphism of a field F onto a field $\tilde{F}$, Γ a set of monic polynomials $f_\alpha(x) \in F[x]$, $\tilde{\Gamma}$ the corresponding set of polynomials $\tilde{f}(x) \in \tilde{F}[x]$, E and $\tilde{E}$ splitting fields over F and $\tilde{F}$ of Γ and $\tilde{\Gamma}$, respectively. Then η can be extended to an isomorphism of E onto $\tilde{E}$.*

Proof. The proof is a straightforward application of Zorn's lemma. We consider the set of extensions of η to monomorphisms of subfields of E/F into $\tilde{E}/\tilde{F}$ and use Zorn's lemma to obtain a maximal one. This must be defined on the whole E, since otherwise we could get a larger one by applying the result quoted to one of the polynomials $f_\alpha(x) \in \Gamma$. Now if ζ is a monomorphism of E into $\tilde{E}$ such that $\zeta|_F = \eta$, then it is clear that $\zeta(E)$ is a splitting field over $\tilde{F}$ of $\tilde{\Gamma}$. Hence, $\zeta(E) = \tilde{E}$ and ζ is an isomorphism of E onto $\tilde{E}$. □

As we have observed, this result applies in particular to algebraic closures. If we take $\tilde{F} = F$ and $\eta = \mathrm{id}$, we obtain

Theorem 5.2.6. *Any two algebraic closures of a field F are isomorphic over F.*

From now on we shall appropriate the notation $\bar{F}$ for any determination of an algebraic closure of F. If A is any algebraic extension of F, its algebraic closure $\bar{A}$ is an algebraic extension of A, hence of F, and so $\bar{A}$ is an algebraic closure of F. Consequently, we have an isomorphism of $\bar{A}/F$ *into* $\bar{F}/F$. This maps A/F into a subfield of $\bar{F}/F$. Thus, we see that every algebraic extension A/F can be realized as a subfield of the algebraic closure $\bar{F}/F$.

Exercises 5.2.

1. No finite field F is algebraically closed. [*Hint*. If $F = \{0, 1, a_2, \ldots, a_n\}$, consider the polynomial $1 + x(x-1)(x-a_2)\ldots(x-a_n) \in F[x]$.]
2. Let E be an algebraic extension of a field F and A an algebraic closure of F. Show that E/F is isomorphic to a subfield of A/F. [*Hint*. Consider the algebraic closure $\bar{A}$ of A and note that this is an algebraic closure of F.]

5.3 Multiple Roots and Separability

Recall the following facts from Subsection 2.6.2 about the multiple roots.

Let R be an integral domain and $f(x) \in R[x]$. If α is a root of $f(x)$, then there exist $m \in \mathbb{N}$ and $g(x) \in R[x]$ such that $f(x) = (x-\alpha)^m g(x)$ and $g(\alpha) \neq 0$. m is called the **multiplicity** of the root α of $f(x)$ and if $m > 1$, α is called a **multiple root** of $f(x)$.

If $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in R[x]$, we define $f'(x) \in R[x]$, the **derivative of** $f(x)$, to be

$$f'(x) = a_1 + a_2 x + \cdots + n a_n x^{n-1}.$$

We record the straightforward properties of the derivative of polynomials in the next lemma.

Lemma 5.3.1. *If $f(x)$ and $g(x)$ are polynomials over an integral domain R and $c \in R$, then*

1. $(cf(x))' = cf'(x)$,
2. $(f(x) + g(x))' = f'(x) + g'(x)$,
3. $(f(x)g(x))' = f(x)g'(x) + f'(x)g(x)$,
4. $((f(x))^n)' = n(f(x))^{n-1} f'(x)$ *where* $n \in \mathbb{N}$.

Theorem 5.3.2. *Let E be an extension of a field F and $f(x) \in F[x]$.*

1. *For $\alpha \in E$, α is a multiple root of $f(x)$ if and only if α is a root of both $f(x)$ and $f'(x)$.*
2. *If $f(x)$ and $f'(x)$ are relatively prime, then $f(x)$ has no multiple root.*
3. *If $f(x)$ is irreducible over F having a root in E, then $f(x)$ has no multiple root in E if and only if $f'(x) \neq 0$.*

Proof. (1) is clear.
(2) Since $f(x)$ and $f'(x)$ are relatively prime, there exist $h(x)$ and $k(x)$ in $F[x]$ such that $1 = f(x)h(x) + f'(x)k(x)$. If $\alpha \in E$ is a multiple root of $f(x)$, by (1), $f(\alpha) = 0 = f'(\alpha)$, so $1 = 0$, a contradiction.
(3) Since $f(x)$ is irreducible, $f'(x) \neq 0$ and $\deg f'(x) < \deg f(x)$, we have $f(x)$ and $f'(x)$ are relatively prime, so $f(x)$ has no multiple roots. Conversely, if $f'(x) = 0$, then $f(\alpha) = 0 = f'(\alpha)$ for some $\alpha \in E$ since $f(x)$ has a root in E. Hence, by (1), α is a multiple root of $f(x)$. □

Let F be a field. A polynomial $f(x) \in F[x]$ is **separable** if every root (in some splitting field over F) of its irreducible factor is not a multiple root. If E is an extension of F and $\alpha \in E$ is algebraic over F, then α is **separable over** F if its minimal polynomial over F is separable.

Let $F \subset K \subset E$ be field extensions. Note that if α is separable over F, then α is separable over K since $m_{\alpha,K}(x) \mid m_{\alpha,F}(x)$. Here $m_{\alpha,-}(x)$ stands for the minimal polynomial of α over the indicated field.

Examples 5.3.1.

1. Consider $f(x) = x^2+1$. Over $\mathbb{Q}$, we have $f(x)$ is irreducible and separable but over $\mathbb{Z}/(2)$, we have $f(x) = x^2 + 1 = (x+1)^2$ is not irreducible but is separable since the only irreducible factor is $x + 1$ which is separable over $\mathbb{Z}/(2)$.
2. Let K be a field of characteristic p and $F = K(y)$ be the field of rational functions over K with indeterminate y. Since $K[y]$ is UFD, y is irreducible element in $K[y]$, so the polynomial $f(x) = x^p - y$ in $F[x]$ is irreducible over F by Eisenstien criterion. Since $f'(x) = 0$ and $f(x)$ has a root, say α in some splitting field E of F, α is a multiple root of $f(x)$, so $f(x)$ is not separable over F. However, if we consider $f(x) = x^p - y \in E[x]$, we have $f(x) = (x-\alpha)^p$ and its irreducible factor in $E[x]$ is only $x - \alpha$ which is separable over E, so $f(x)$ is separable over E.

Suppose that F is a field of characteristic zero and $f(x)$ is a monic irreducible polynomial over F, say $f(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + x^n$. Then $f'(x) = a_1 + 2a_2 x + \cdots + n x^{n-1}$. The *key* point is that $n \neq 0$, so $f'(x) \neq 0$. Since $\deg f'(x) < \deg f(x)$ and $f(x)$ is irreducible, $f(x)$ and $f'(x)$ are relatively prime, so all roots of $f(x)$ are simple. Thus, we have shown:

Theorem 5.3.3. *Let F be a field of characteristic zero. Then every polynomial $f(x) \in F[x]$ is separable.*

We call an *algebraic* extension field E of a field F a **separable extension** if the minimal polynomial of every element of E is separable. Hence, if F is of characteristic zero, then every algebraic extension is a separable extension. A field F is **perfect** if every polynomial $f(x)$ over F is separable. Thus, all fields of characteristic zero are perfect.

Remark. Suppose F is a field (or even a commutative ring) of characteristic $p > 0$. Then the identities

$$(ab)^p = a^p b^p \quad \text{and} \quad (a+b)^p = a^p + b^p$$

show that the map $\varphi : F \to F$ defined by $\varphi(a) = a^p$ is a ring homomorphism. Since F is a field, φ has to be one-to-one. But φ does *not* have to be onto - for example

$$\varphi : (\mathbb{Z}/p\mathbb{Z})(x) \to (\mathbb{Z}/p\mathbb{Z})(x)$$

is not onto; the image is $(\mathbb{Z}/p\mathbb{Z})(x^p)$. However, if F is *finite* of order p^n, then $a^{p^n} = a$ for all $a \in F$, so φ is onto and φ^n is the identity map, called the **Frobenius' automorphism**.

Theorem 5.3.4. *Let F be a field of characteristic $p > 0$, and let $a \in F$.*
(1) If $a \in F^p$ and $a = r^p$, then $x^p - a = (x - r)^p$.
(2) If $a \notin F^p$, then $x^p - a$ is irreducible.

Proof. (1) is trivial.
(2) In a splitting field for F, $x^p - a = (x - r)^p$ (r may not be in F). Any proper factor of $x^p - a$ (after being made monic) has the form $(x - r)^i$ where $1 \leq i \leq p - 1$. Thus, if $x^p - a$ has a proper factor over F, then $r^i \in F$ for some $1 \leq i \leq p - 1$. But then r^i and $r^p = a \in F$, so $r \in F$ since $(i, p) = 1$. Hence, $a = r^p \in F^p$. □

Theorem 5.3.5. *Let F be a field of characteristic $p > 0$. Then F is perfect if and only if $F = F^p$.*

Proof. Suppose $F \neq F^p$ and choose $a \in F \setminus F^p$. By Theorem 5.3.4, $x^p - a$ is irreducible. But $x^p - a$ does not have distinct roots in a splitting field of F. Hence, F is not perfect.

Conversely, assume that F is not perfect. Then there is an irreducible polynomial $f(x)$ over F which does not have simple roots. By Theorem 5.3.2, this means that $f(x)$ and $f'(x)$ are not relatively prime. Since $f(x)$ is irreducible and $\deg f'(x) < \deg f(x)$, $f'(x) = 0$. Thus, $f(x)$ is a polynomial in x^p, i.e.,

$$f(x) = a_0 + a_p x^p + a_{2p} x^{2p} + \cdots + a_{(m-1)p} x^{(m-1)p} + x^{mp}.$$

We shall claim that some $a_{jp} \notin F^p$. For if each $a_{jp} \in F^p$, say $a_{jp} = (b_j)^p$, then $f(x) = g(x)^p$ where

$$g(x) = b_0 + b_1 x + \cdots + b_{m-1} x^{m-1} + x^m$$

which contradicts the irreducibility of $f(x)$ over F. This establishes the claim. Hence, $a_{jp} \notin F^p$ and $F \neq F^p$. □

Corollary 5.3.6. *Every finite field is perfect.*

Proof. The characteristic of a finite field F is a prime p. The monomorphism $a \mapsto a^p$ of F is an isomorphism since F is finite. Hence, $F = F^p$ is perfect by Theorem 5.3.5. □

We shall end this section by proving the "primitive element theorem" which is a classic of field theory. We first recall that an extension field E of a field F is said to be a **simple extension** of F if $E = F(\alpha)$ for some $\alpha \in E$. Such an element α is called a **primitive element**.

Theorem 5.3.7. *If F is a field and G is a finite subgroup of the multiplicative group of nonzero elements of F, then G is a cyclic group. In particular, the multiplicative group of all nonzero elements of a finite field is cyclic.*

Proof. If $G = \{1\}$, then G is cyclic. Assume that $G \neq \{1\}$. Since G is a finite abelian group,

$$G \cong \mathbb{Z}/(m_1) \oplus \cdots \oplus \mathbb{Z}/(m_k)$$

where $k \geq 1, m_1 > 1$ and $m_1 \mid \cdots \mid m_k$. Since $m_k(\sum_{i=1}^{k} \mathbb{Z}/(m_i)) = 0$, u is a root of the polynomial $x^{m_k} - 1 \in F[x]$ for all $u \in G$. By Theorem 5.1.1, this polynomial has at most m_k distinct roots in F, we must have $k = 1$ and $G \cong \mathbb{Z}/(m_1)$ which is a cyclic group. □

Theorem 5.3.8. [Primitive Element Theorem] *Let E be a finite separable extension of a field F. Then there exists $\alpha \in E$ such that $E = F(\alpha)$. That is, a finite separable extension of a field is a simple extension.*

Proof. If F is a finite field, then E is also finite. Let α be a generator for the cyclic group of all nonzero elements of E under multiplication. Clearly, $E = F[\alpha]$, so α is a primitive element in this case.

We now assume that F is infinite and prove our theorem in the case that $E = F(\beta, \gamma)$. The induction argument from this to the general case is obvious. Let $m_{\beta,F}(x)$ and $m_{\gamma,F}(x)$ be the minimal polynomials over F of β and γ, respectively. Assume that $m_{\beta,F}(x)$ has distinct roots $\beta = \beta_1, \ldots, \beta_n$ and $m_{\gamma,F}(x)$ has distinct roots $\gamma = \gamma_1, \ldots, \gamma_m$ in $\bar{F}$ where all roots are of multiplicity 1, since E is a separable extension of F. Since F is infinite, we can find $a \in F$ such that

$$a \neq \frac{\beta_i - \beta}{\gamma - \gamma_j}$$

for all i and j, with $j \neq 1$. That is, $a(\gamma - \gamma_j) \neq \beta_i - \beta$. Letting $\alpha = \beta + a\gamma$, we have $\alpha = \beta + a\gamma \neq \beta_i + a\gamma_j$, so

$$\alpha - a\gamma_j \neq \beta_i$$

for all i and all $j \neq 1$. Consider $h(x) = m_{\beta,F}(\alpha - ax) \in F(\alpha)[x]$. Now, $h(\gamma) = m_{\beta,F}(\beta) = 0$. However, $h(\gamma_j) \neq 0$ for $j \neq 1$ by construction, since the β_i were the only roots of $m_{\beta,F}(x)$. Hence, $h(x)$ and $m_{\gamma,F}(x)$ have a common factor in $F(\alpha)[x]$, namely the minimal polynomial of γ over $F(\alpha)$, which must be linear, since γ is the only common root of $m_{\gamma,F}(x)$ and $h(x)$. Thus, $\gamma \in F(\alpha)$, and therefore $\beta = \alpha - a\gamma$ is in $F(\alpha)$. Hence, $F(\beta, \gamma) = F(\alpha)$. □

Exercises 5.3.

1. Suppose that $F \subseteq K \subseteq E$ and that E is separable extension of F. Prove that E is separable over K and K is separable over F.
2. Let F be of characteristic p and let $a \in F$. Show that $f(x) = x^p - x - a$ has no multiple roots and $f(x)$ is irreducible in $F[x]$ if and only if $a \neq c^p - c$ for any $c \in F$.
3. Find a primitive element of $\mathbb{Q}(i, \sqrt[3]{2})$ over $\mathbb{Q}$.
4. Let $K = \mathbb{F}_{25}$ be the field with 5 elements and let $F = \mathbb{Z}/(5)$ be the prime subfield of K. Determine the cardinalities of the following two sets.
   (a) The set of elements of K which generate K as a field over F.
   (b) The set of elements of K which generate the group of nonzero elements of K as an abelian group under multiplication.

5. Let F be a field and let $\overline{F}$ be its algebraic closure. If a monic polynomial $p(x) \in F[x]$ is irreducible over F and has distinct roots $\alpha_1, \alpha_2, \ldots, \alpha_k \in \overline{F}$, prove that the multiplicities of α_j are equal, that is,

$$p(x) = (x - \alpha_1)^m (x - \alpha_2)^m \ldots (x - \alpha_k)^m$$

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

5.4 Automorphisms of Fields and Galois Theory

If F is a field, the set of automorphisms of F, $\operatorname{Aut} F$, forms a group under composition of functions.

Examples 5.4.1 (Examples of automorphism groups). 1. Any automorphism satisfies $\varphi(1) = 1$, so $\varphi(n) = n$ for all $n \in \mathbb{Z}$ and $\varphi(n/m) = n/m$ if $n, m \in \mathbb{Z}$ and $m \neq 0$ in F. This implies that the fields $\mathbb{Q}$ and $\mathbb{F}_p = \mathbb{Z}/(p)$ have only the identity map as an automorphism. That is, $\operatorname{Aut}(F) = \{\mathrm{id}_F\}$ if $F = \mathbb{Q}$ or $\mathbb{F}_p$. Moreover, any field E is an extension of $\mathbb{Q}$ or $\mathbb{F}_p$ (so called the prime subfield) and any automorphism $\varphi : E \to E$ leaves the prime subfield pointwise fixed.

2. The only automorphism $\varphi : \mathbb{R} \to \mathbb{R}$ is the identity map. For, we have known that $\varphi(q) = q$ for all $q \in \mathbb{Q}$. Note that $\varphi(a) = \varphi((\sqrt{a})^2) = (\varphi(\sqrt{a}))^2 > 0$ for all $a > 0$. Thus, if $a < b$, then $\varphi(a) < \varphi(b)$. Let $x \in \mathbb{R}$. Suppose $\varphi(x) \neq x$. Then $\varphi(x) < x$ or $\varphi(x) > x$. If $\varphi(x) < x$, then there exists a $q \in \mathbb{Q}$ such that $\varphi(x) < q < x$. Thus, $q = \varphi(q) < \varphi(x)$, a contradiction. If $x < \varphi(x)$, then there exists a $q \in \mathbb{Q}$ such that $x < q < \varphi(x)$, so $\varphi(x) < \varphi(q) = q$, a contradiction. Hence, $\varphi = \mathrm{id}_{\mathbb{R}}$.

3. Complex conjugation: $\varphi(z) = \bar{z}$ is an automorphism of $\mathbb{C}$ of order two. In fact, $\operatorname{Aut}\mathbb{C}$ is uncountable, but the other automorphisms are "indescribable" and exist only via Zorn's lemma. However, the group of automorphisms of $\mathbb{C}$ which fix all elements of $\mathbb{R}$ is a group of order two.

4. Let F be a field and let $E = F(t)$ where t is transcendental over F. As shall be indicated in the Exercise 5.4 below, $u \in E$ is a generator of E/F if and only if it has the form

$$u = \frac{\alpha t + \beta}{\gamma t + \delta}, \quad \alpha\delta - \beta\gamma \neq 0.$$

Since an automorphism of E/F sends generators into generators, it follows that every automorphism $\varphi : E \to E$ is given by

$$\varphi(a) = a \text{ for all } a \in F \quad \text{and} \quad \varphi(t) = \frac{\alpha t + \beta}{\gamma t + \delta},$$

where $\alpha, \beta, \gamma, \delta \in F$ and $\alpha\delta - \beta\gamma \neq 0$. Note that if $c \in F$ and $c \neq 0$, then

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} c\alpha & c\beta \\ c\gamma & c\delta \end{bmatrix}$$

give rise to the same automorphism of $F(t)$. A computation shows composition of functions corresponds to matrix multiplication. The net result is that

$$\operatorname{Aut} F(t) \cong \mathrm{GL}_2(F)/F^\times = \mathrm{PGL}_2(F),$$

where $F^\times$ is the set of matrices aI, $a \neq 0$.

5. If F is a subfield of K, let

$$\operatorname{Aut}_F K = \{\varphi \in \operatorname{Aut} K : \varphi(a) = a \text{ for all } a \in F\}.$$

The group structure of $\operatorname{Aut}_F F(x, y)$ is known, but very complicated. For $n \geq 3$, almost nothing is known about $\operatorname{Aut}_F F(x_1, \ldots, x_n)$.

The above examples show that $\mathrm{Aut}\, F$ is in general very complicated and probably impossible to describe. Galois theory proceeds in a different direction. One takes a subgroup H of $\mathrm{Aut}\, F$—we shall be almost concerned with finite H—and looks the set

$$F^H = \{a \in F : \varphi(a) = a \text{ for all } \varphi \in H\}.$$

It is easy to see that F^H is a subfield of F. Moreover, if K is a subgroup of H, then

$$1 \subseteq K \subseteq H$$
$$F \supseteq F^K \supseteq F^H.$$

The fundamental result of Galois theory is that of F is separable over F^H, then there is a one-to-one correspondence between subgroups of H and subfields of F which contain F^H. Such correspondences are inclusion reversing and are called "Galois correspondences".

Let E be an extension field of a field F. The **Galois group of E over F** denoted by $\mathrm{Gal}(E/F)$ is the group

$$\{\varphi \in \mathrm{Aut}\, E : \varphi(a) = a \text{ for all } a \in F\}.$$

Let G be a subgroup of $\mathrm{Aut}\, E$ where E is a field. Then the **field of G-invariant of E** or the **fixed field of G acting on E** is the field

$$\{a \in E : \varphi(a) = a \text{ for all } \varphi \in G\}.$$

It is denoted by E^G or $\mathrm{Inv}\, G$.

Theorem 5.4.1. *(1) If $\mathrm{Aut}\, E \supseteq G_1 \supseteq G_2$, then $E^{G_1} \subseteq E^{G_2}$.*
(2) If $E \supseteq F_1 \supseteq F_2$, then $\mathrm{Gal}(E/F_1) \subseteq \mathrm{Gal}(E/F_2)$.
(3) If $G = \mathrm{Gal}(E/F)$, then $E^G \supseteq F$.
(4) If $F = E^G$, then $\mathrm{Gal}(E/F) \supseteq G$.

Proof. These are immediate consequences of the definitions. □

We shall now apply these ideas to splitting fields. Using the present terminology, Theorem 5.1.7 can be restated as follows. If E is a splitting field over F of a polynomial $f(x)$, then $\mathrm{Gal}(E/F)$ is finite and we have the inequality $|\mathrm{Gal}(E/F)| \leq [E : F]$. Moreover, $|\mathrm{Gal}(E : F)| = [E : F]$ if $f(x)$ has distinct roots. We therefore have the following important preliminary result.

Lemma 5.4.2. *Let E/F be a splitting field of a separable polynomial contained in $F[x]$. Then*

$$|\mathrm{Gal}(E/F)| = [E : F].$$

Our next attack will be from the group side. We begin with an arbitrary field E and any finite group of automorphisms G acting in E. Then we have the following

Lemma 5.4.3. [Artin] *Let G be a finite subgroup of $\mathrm{Aut}\, E$ and let $F = E^G$. Then*

$$[E : F] \leq |G|.$$

Proof. Let $|G| = n$ and write $G = \{g_1 = 1, g_2, \ldots, g_n\}$. We have to show that $[E : F] \leq n$, or equivalently:
$(*)$ If $x_1, \ldots, x_{n+1} \in E$, then there exist $u_1, \ldots, u_{n+1} \in F$ not all zero, such that

$$u_1x_1 + \cdots + u_{n+1}x_{n+1} = 0,$$

that is, $x_1, \ldots, x_{n+1}$ are linearly dependent over F.

Consider the following $n \times (n+1)$ matrix with entries in E

$$M = \begin{bmatrix} x_1 & x_2 & \ldots & x_{n+1} \\ g_2(x_1) & g_2(x_2) & \ldots & g_2(x_{n+1}) \\ \vdots & & & \\ g_n(x_1) & g_n(x_2) & \ldots & g_n(x_{n+1}) \end{bmatrix}.$$

This matrix has rank $\leq n$, so there is a nonzero $(n+1) \times 1$ vector $\vec{v} = (v_1, \ldots, v_{n+1})^t$ with entries in E such that $M\vec{v} = \vec{0}_{(n+1)\times 1}$. We wish to find such a vector where entries lie in F. Among all such vectors with entries in E, choose one in which the number of nonzero coordinates, r, is minimal. By renaming the elements $x_1, \ldots, x_{n+1}$, we may suppose that the nonzero coordinates are the first r of them; by multiplying the vector by v_r^{-1} we may suppose that the last nonzero coordinate is equal to 1. Thus,

$$M\vec{v} = \vec{0}_{(n+1)\times 1} \quad \text{where} \quad \vec{v} = (v_1, \ldots, v_{r-1}, 1, 0, \ldots, 0)^t.$$

Claim. *If $h \in G$ and $h(\vec{v}) = (h(v_1), \ldots, h(v_{r-1}), 1, 0, \ldots, 0)^t$, then $Mh(\vec{v}) = \vec{0}$.*
Proof of Claim. The inner product of the j-th row of M with $h(\vec{v})$ is:

$$z = g_j(x_1)h(v_1) + \cdots + g_j(x_{r-1})h(v_{r-1}) + g_j(x_r) \cdot 1.$$

Apply the automorphism h^{-1} to z,

$$\begin{aligned} h^{-1}z &= h^{-1}g_j(x_1)h(v_1) + \cdots + h^{-1}g_j(x_{r-1})h(v_{r-1}) + h^{-1}g_j(x_r) \cdot 1 \\ &= g_i(x_1)v_1 + \cdots + g_i(x_{r-1})v_{r-1} + g_i(x_r) \cdot 1 = 0, \end{aligned}$$

since $h^{-1}g_j = g_i$ for some i. This proves the claim.

Now we consider, for any $h \in G$

$$\begin{aligned} \vec{v} - h(\vec{v}) &= (v_1, \ldots, v_{r-1}, 1, 0, \ldots, 0)^t - (h(v_1), \ldots, h(v_{r-1}), 1, 0, \ldots, 0)^t \\ &= (\overbrace{*, \ldots, *}^{r-1}, 0, \ldots, 0)^t. \end{aligned}$$

Since $M(\vec{v} - h(\vec{v})) = \vec{0}$ and $\vec{v} - h(\vec{v})$ has at most $r-1$ nonzero entries, $\vec{v} - h(\vec{v}) = \vec{0}$ by the minimal choice of r. This means that for all $h \in G$ and $i = 1, \ldots, r-1$, we have $h(v_i) = v_i$. Thus, all the v_i lie in $E^G = F$ and $(u_1, \ldots, u_{n+1}) = (v_1, \ldots, v_{r-1}, 0, \ldots, 0)$ is a set of elements of F which satisfies $(*)$. □

Recall that an algebraic extension field E of a field F is a separable extension if the minimal polynomial of every element of E is separable. We call an *algebraic* extension field E of a field F a **normal extension** if every irreducible polynomial in $F[x]$ which has a root in E splits into linear factors in E. This is equivalent to saying that E contains a splitting field for the minimal polynomial of every element of E. Normality plus separability, called a **Galois extension**, mean that every irreducible polynomial of $F[x]$ which has a root in E is a product of distinct linear factors in $E[x]$. Also, by the results of the last section, if E is algebraic over F, then E is necessarily separable over F if the characteristic is zero or if the characteristic is $p > 0$ and $F^p = F$.

We are now ready to derive our main results, the first of which gives two abstract characterizations of splitting fields of separable polynomials and some important additional information. We state this as

Theorem 5.4.4. *Let E be an extension field of a field F. Then the following conditions on E/F are equivalent.*

(i) E is a splitting field over F of a separable polynomial $f(x)$.

(ii) $F = E^G$ for some finite group G of automorphisms of E.

(iii) E is finite dimensional Galois (normal and separable) over F.

Moreover, if E and F are as in (i) and $G = \mathrm{Gal}(E/F)$, then $F = E^G$ and if G and F are as in (ii), then $G = \mathrm{Gal}(E/F)$.

Proof. (i) $\Rightarrow$ (ii). Let $G = \mathrm{Gal}(E/F)$. Then E^G is a subfield of E containing F. Also it is clear that E is a splitting field over E^G of $f(x)$ as well as over F and $G = \mathrm{Gal}(E/E^G)$. Hence, by Lemma 5.4.2, $|G| = [E : F]$ and $|G| = [E : E^G]$. Since $E \supseteq E^G \supseteq F$, we have $[E : F] = [E : E^G][E^G : F]$. Hence, $[E^G : F] = 1$, and so $E^G = F$. We have prove also that $F = E^G$ for $G = \mathrm{Gal}(E/F)$, which is the first of the two supplementary statements.

(ii) $\Rightarrow$ (iii). By Artin's lemma, $[E : F] \leq |G|$, and so E is finite dimensional over F. Let $f(x)$ be an irreducible polynomial in $F[x]$ having a root r in E. Let $\{r = r_1, r_2, \ldots, r_m\}$ be the orbit of r under the action of G. Thus, this is the set of distinct elements of the form $\sigma(r), \sigma \in G$. Hence, if $\sigma \in G$, then the set $\{\sigma(r_1), \sigma(r_2), \ldots, \sigma(r_m)\}$ is a permutation of $\{r_1, r_2, \ldots, r_m\}$. We have $f(r) = 0$ which implies that $f(r_i) = 0$. Then $f(x)$ is divisible by $x - r_i$, and since the r_i, $1 \leq i \leq m$, are distinct, $f(x)$ is divisible by $g(x) = \prod_{i=1}^{m}(x - r_i)$. We now apply to $g(x)$ the automorphism of $E[x]$, which sends $x \to x$ and $a \to \sigma(a)$ for $a \in E$. This gives $\sigma g(x) = \prod_{i=1}^{m}(x - \sigma(r_i)) = \prod_{i=1}^{m}(x - r_i) = g(x)$. Since this holds for every $\sigma \in G$ we see that the coefficients of $g(x)$ are G-invariant. Hence, $g(x) \in F[x]$. Since we assumed $f(x)$ irreducible in $F[x]$ we see that $f(x) = g(x) = \prod(x - r_i)$, a product of distinct linear factors in $E[x]$. Thus, E is separable and normal over F and (iii) holds.

(iii) $\Rightarrow$ (i). Since we are given that $[E : F] < \infty$ we can write $E = F(r_1, r_2, \ldots, r_k)$ and each r_i is algebraic over F. Let $f_i(x)$ be the minimal polynomial of r_i over F. Then the hypothesis implies that $f_i(x)$ is a product of distinct linear factors in $E[x]$. It follows that $f(x) = \prod_{i=1}^{k} f_i(x)$ is separable and $E = F(r_1, r_2, \ldots, r_k)$ is a splitting field over F of $f(x)$. Hence, we have (i).

It remains to prove the second supplementary statement. We have seen that under the hypothesis of (ii) we have $[E : F] \leq |G|$, and that since (i) holds, we have $|\mathrm{Gal}(E/F)| = [E : F]$. Since $G \subseteq \mathrm{Gal}(E/F)$ and $|G| \geq [E : F] = |\mathrm{Gal}(E/F)|$, equivalently $G = \mathrm{Gal}(E/F)$. □

The above proof also yields

Corollary 5.4.5. *If E/F is the splitting field of $f(x) \in F[x]$ and $r_1, \ldots, r_n$ are distinct roots of $f(x)$ in E, then $G = \mathrm{Gal}(E/F)$ may be identified with a subgroup of S_n, the group of permutations of $\{r_1, \ldots, r_n\}$. However, it is not always the case that $\mathrm{Gal}(E/F)$ is the full group of permutations of the roots of $f(x)$.*

There are two observations underlying the above corollary.

1. Each $\sigma \in G$ permutes $r_1, \ldots, r_n$.
2. $\sigma \in G$ is determined by its action on $r_1, \ldots, r_n$ because $r_1, \ldots, r_n$ generate E as a field over F, i.e., $E = F[r_1, \ldots, r_n] = F(r_1, \ldots, r_n)$.

Example 5.4.2 (Elementary symmetric functions)**.** If K is a field, then the polynomial ring $K[x_1, \ldots, x_n]$ is an integral domain. The quotient field of $K[x_1, \ldots, x_n]$ is denoted by $K(x_1, \ldots, x_n)$ and is called the **field of rational functions** in $x_1, \ldots, x_n$ over K. In the field extension

$$K \subset K(x_1, \ldots, x_n)$$

each x_i is easily seen to be transcendental over K. In fact, every element of $K(x_1, \ldots, x_n)$ not in K itself is transcendental over K (Prove!).

Let S_n be the symmetric group on n letters. A rational function $\varphi \in K(x_1, \ldots, x_n)$ is said to be **symmetric** in $x_1, \ldots, x_n$ over K if for every $\sigma \in S_n$,

$$\varphi(x_1, x_2, \ldots, x_n) = \varphi(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}).$$

Trivially, every constant polynomial is a symmetric function. More generally, the **elementary symmetric functions** in $x_1, \ldots, x_n$ over K are defined to be the polynomials:

$$\begin{aligned}
e_1 &= x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i; \\
e_2 &= \sum_{1 \le i < j \le n} x_i x_j; \\
&\vdots \\
e_k &= \sum_{1 \le i_1 < \cdots < i_k \le n} x_{i_1} x_{i_2} \ldots x_{i_k}; \\
&\vdots \\
e_n &= x_1 x_2 \ldots x_n.
\end{aligned}$$

The verification that the e_i are indeed symmetric follows from the fact that they are simply the coefficients of t in the polynomial $p(t) \in K[x_1, \ldots, x_n][t]$, where

$$p(t) = (t - x_1)(t - x_2) \ldots (t - x_n) = t^n - e_1 t^{n-1} + e_2 t^{n-2} - \cdots + (-1)^{n-1} e_{n-1} t + (-1)^n e_n.$$

If $\sigma \in S_n$, then the assignments $x_i \mapsto x_{\sigma(i)}$, $i = 1, 2, \ldots, n$ and

$$f(x_1, \ldots, x_n)/g(x_1, \ldots, x_n) \mapsto f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})/g(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

define a K-automorphism of the field $E = K(x_1, \ldots, x_n)$ which will also be denoted σ. The map $S_n \to \mathrm{Gal}(E/K)$ given by $\sigma \mapsto \sigma$ is clearly a monomorphism of groups, whence S_n may be consider as a subgroup of the Galois group $\mathrm{Gal}(E/K)$. Clearly, the fixed field $F = E^{S_n}$ consists precisely of symmetric functions; that is, the set of all symmetric functions is a subfield of E containing K. Therefore, by Theorem 5.4.4, E is a Galois extension of F with Galois group $\mathrm{Gal}(E/F) = S_n$ and dimension $|S_n| = n!$.

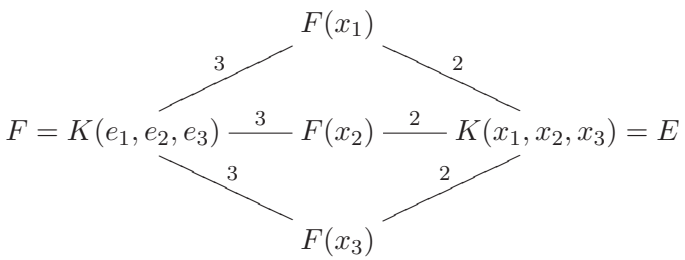
Example 5.4.3. Let K be a field and x_1, x_2, x_3 be indeterminates over K, set

$$e_1 = x_1 + x_2 + x_3, e_2 = x_1 x_2 + x_2 x_3 + x_3 x_1, e_3 = x_1 x_2 x_3$$

and consider the fields

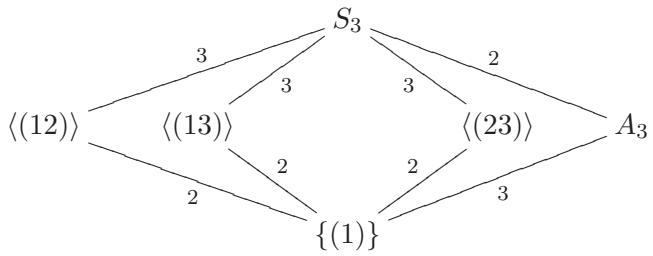
$$F = K(e_1, e_2, e_3) \subseteq K(x_1, x_2, x_3) = E.$$

The relevant subfields of E are indicated in the diagram



The fields $F(x_1), F(x_2)$ and $F(x_3)$ are all isomorphic (over F), but they are distinct subfields of E. Moreover, E is a splitting field for $f(t) = t^3 - e_1 t^2 + e_2 t - e_3$ but $F(x_1), F(x_2)$ and $F(x_3)$ are not.

We know that $G = \mathrm{Gal}(E/F) = S_3$ where S_3 is identified with the group of permutations on 3 letters. We next calculate E^H when H is a subgroup of $G = \mathrm{Gal}(E/F) = S_3$. The following is a diagram of the lattice of subgroups of S_3 and there indices.



We have already calculated that $E^{S_3} = E^G = F$ and of course $E^{\{(1)\}} = E$. It is not hard to verify that

$$E^{\langle(12)\rangle} = F[x_3], E^{\langle(13)\rangle} = F[x_2], E^{\langle(23)\rangle} = F[x_1].$$

It is somewhat more difficult to verify that $E^{A_3} = F[\Delta]$ where

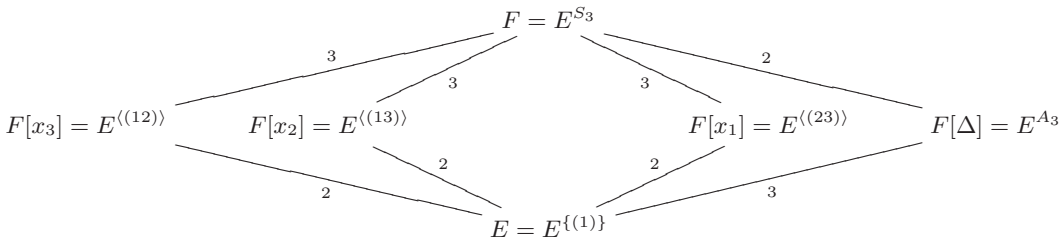
$$\Delta = (x_1 - x_2)(x_2 - x_3)(x_3 - x_1).$$

Note that $\sigma(\Delta) = \Delta$ if $\sigma \in A_3$, but $\sigma(\Delta) = -\Delta$ if $\sigma \in S_3 \setminus A_3$.

We already know that

$$[F[x_1] : F] = [F[x_2] : F] = [F[x_3] : F] = 3$$

and one can verify that $[F[\Delta] : F] = 2$. Thus, we get the following diagrams of *all* (by Galois Theory) subfields of E containing F



The indices are the same as in the lattice diagram for S_3, but inclusions are reversed. Recall that E is the splitting field of a separable polynomial

$$f(t) = (t - x_1)(t - x_2)(t - x_3)$$

for any field in the above diagram. More generally, it is clear that if M/L is a splitting field for $f(t) \in L[t]$ and $M \supseteq N \supseteq L$, then M/N is a splitting field for $f(t)$, regarded as a polynomial in $N[t]$.

Furthermore, for each field L in the above diagram, we have $L = E^H$ for some subgroup H of $G = S_3$ and $\mathrm{Gal}(E/L) = H$. On the other hand, things are not so nice for the extensions L/F. For example, $\mathrm{Gal}(F[x_i]/F) = 1$ for all $i = 1, 2, 3$ and $\mathrm{Gal}(F[\Delta]/F) \cong \mathbb{Z}/(2) = \langle\varphi\rangle$ where the action of φ is $\varphi(\Delta) = -\Delta$. Here $\Delta^2 \in F$ and $F[\Delta]$ is the splitting field of the polynomial $t^2 - \Delta^2$ over F, so it is Galois. However, we may conclude that the fields $F[x_1], F[x_2]$ and $F[x_3]$ are not the splitting fields of any polynomials over F.

The previous example illustrates the fundamental theorem of Galois theory: if E/F is the splitting field of a *separable* polynomial $f(t) \in F[t]$, then the map

$$H \longleftrightarrow E^H = \{a \in E : \varphi(a) = a \text{ for all } \varphi \in H\}$$

is a 1-1 correspondence between

$$\text{subgroups of } \mathrm{Gal}(E/F) \longleftrightarrow \text{subfields of } E$$

which reverses inclusions. In addition, H is a normal subgroup of $\mathrm{Gal}(E/F)$ if and only if E^H is the splitting field of some separable polynomial over F (i.e., E^H is normal over F), and if H is *normal* in $\mathrm{Gal}(E/F)$, then

$$\mathrm{Gal}(E^H/F) \cong \mathrm{Gal}(E/F)/H.$$

In our example, the only proper normal subgroup of S_3 is A_3, and

$$\mathrm{Gal}(E^{A_3}/F) = \mathrm{Gal}(F[\Delta]/F) \cong \mathbb{Z}_2 \cong S_3/A_3 = \mathrm{Gal}(E/F)/A_3.$$

We now formally establish Galois' fundamental group-field pairing as follows.

Theorem 5.4.6. [Fundamental Theorem of Galois Theory] *Let E be a finite dimensional Galois extension of a field F (i.e., the conditions of Theorem 5.4.4 holds) and let $G = \mathrm{Gal}(E/F)$. Let $\Gamma = \{H\}$, the set of subgroups of G, and Σ, the set of intermediate fields between E and F (the subfields of E/F). Then the map $H \mapsto E^H$ and $K \mapsto \mathrm{Gal}(E/K)$, $H \in \Gamma$, $K \in \Sigma$, are inverses to each other. In particular, they are one-to-one correspondences between Γ and Σ. Moreover, the pairing $\Gamma \leftrightarrow \Sigma$ has the following properties:*

1. *$H_1 \supseteq H_2$ if and only if $E^{H_1} \subseteq E^{H_2}$.*
2. *$|H| = [E : E^H]$ and $[G : H] = [E^H : F] = [E^H : E^G]$.*
3. *H is normal in G if and only if E^H is normal over F. In this case,*

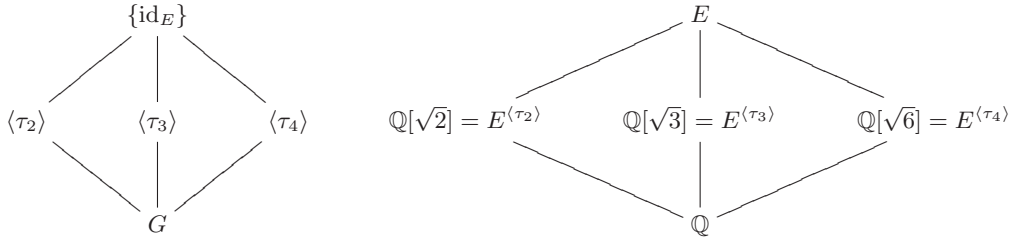
$$\mathrm{Gal}(E^H/F) \cong G/H.$$

This is the main theorem. Most of our remaining field theory will be consequences of it.

Proof. Let H be a subgroup of $G = \mathrm{Gal}(E/F)$. Since $F = E^G$, $F \subseteq E^H$ and E^H is thus a subfield of E containing F. Also, E/E^H is Galois. Applying the second supplementary result of Theorem 5.4.4 to H in place of G we see that $\mathrm{Gal}(E/E^H) = H$. By Lemma 5.4.2, $|H| = |\mathrm{Gal}(E/E^H)| = [E : E^H]$. Now let K be any subfield of E/F. Then $\mathrm{Gal}(E/K) \subseteq G = \mathrm{Gal}(E/F)$, so $\mathrm{Gal}(E/K)$ is a subgroup of G. It is clear also that E is a splitting field over K of a separable polynomial. Hence, the first supplementary result of Theorem 5.4.4 applied to the pair E and K shows that $K = E^{\mathrm{Gal}(E/K)}$. We have now shown that the specified maps between Γ and Σ are inverses. Also, we know that if $H_1 \supseteq H_2$, then $E^{H_1} \subseteq E^{H_2}$. Moreover, if $E^{H_1} \subseteq E^{H_2}$, then we have also that $H_1 = \mathrm{Gal}(E/E^{H_1}) \supseteq \mathrm{Gal}(E/E^{H_2}) = H_2$. Hence, (1) holds. The first part of (2) was noted before. Since $|G| = [E : F] = [E : E^H][E^H : F] = |H|[E^H : F]$ and $|G| = |H|[G : H]$, evidently $[E^H : F] = [G : H]$. This proves (2).

If $H \in \Gamma$, then $E^{\eta H \eta^{-1}} = \eta(E^H)$ for all $\eta \in G$. This is clear since the condition $\sigma(x) = x$ is equivalent to $(\eta\sigma\eta^{-1})(\eta(x)) = \eta(x)$. It now follows that H is normal in G if and only if $\eta(E^H) = E^H$ for every $\eta \in G$. Suppose H is normal in G. Then every $\eta \in G$ maps E^H onto itself and so its restriction $\bar{\eta} = \eta|_{E^H}$ is an automorphism of E^H/F. Thus, we have the restriction homomorphism $\eta \to \bar{\eta}$ of $G = \mathrm{Gal}(E/F)$ into $\mathrm{Gal}(E^H/F)$. The image $\bar{G}$ is a group of automorphisms in E^H and clearly $(E^H)^{\bar{G}} = F$. Hence, $\bar{G} = \mathrm{Gal}(E^H/F)$. The kernel of the homomorphism $\eta \to \bar{\eta}$ is the set of $\eta \in G$ such that $\eta|_{E^H} = \mathrm{id}_{E^H}$. By the pairing, this is $\mathrm{Gal}(E/E^H) = H$. Hence, the kernel is H and $\bar{G} = \mathrm{Gal}(E^H/F) \cong G/H$. Since $F = (E^H)^{\bar{G}}$, E^H is normal over F by Theorem 5.4.4. Conversely, suppose E^H is normal over F. Let $a \in E^H$ and let $f(x)$ be the minimal polynomial of a over F. Then $f(x) = (x-a_1)\dots(x-a_m)$ in $E^H[x]$ where $a = a_1$. If $\eta \in G$, then $f(\eta(a)) = 0$ which implies that $\eta(a) = a_i$ for some i. Thus, $\eta(a) \in E^H$. We have therefore shown that $\eta(E^H) = E^H$. Hence, H is a normal subgroup of G. This completes the proof of (3). □

Example 5.4.4. Let $E = \mathbb{Q}[\sqrt{2}, \sqrt{3}]$ be a splitting field of $f(x) = (x^2 - 2)(x^2 - 3)$. Then E is Galois over $\mathbb{Q}$. Let $G = \mathrm{Gal}(E/\mathbb{Q})$. Then $|G| = [E : \mathbb{Q}] = 4$. Since $\mathbb{Q}(\sqrt{2})$ is a splitting field of $x^2 - 2$, it is Galois over $\mathbb{Q}$ and its Galois group consists of 2 elements, namely $\sigma_1 = \mathrm{id}$ and $\sigma_2 : \sqrt{2} \mapsto -\sqrt{2}$. Each automorphism extends to an automorphism of E in two different ways; $\sqrt{3} \mapsto \sqrt{3}$ or $\sqrt{3} \mapsto -\sqrt{3}$. Then the four elements of G are $\tau_1 = \mathrm{id}_E$, $\tau_2 : \sqrt{2} \mapsto \sqrt{2}, \sqrt{3} \mapsto -\sqrt{3}$, $\tau_3 : \sqrt{2} \mapsto -\sqrt{2}, \sqrt{3} \mapsto \sqrt{3}$ and $\tau_4 : \sqrt{2} \mapsto -\sqrt{2}, \sqrt{3} \mapsto -\sqrt{3}$. Each of these elements except τ_1 has order 2. Thus, $G \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. Hence, the subgroup-intermediate subfield correspondence for the fundamental theorem of Galois theory is shown in the lattice diagrams



Exercises 5.4.

1. Let $E = F(t)$ where t is transcendental over F and write any non-zero element of E as $u = f(t)/g(t)$ where $(f(t), g(t)) = 1$. Call the maximum of degrees of f and g the *degree* of u. Show that if x and y are indeterminates then $f(x) - yg(x)$ is irreducible in $F[x, y]$ and hence is irreducible in $F(y)[x]$. Show that t is algebraic over $F(u)$ with minimal polynomial the monic polynomial which is a multiple in $F(u)$ of $f(x) - ug(x)$. Hence, conclude that $[F(t) : F(u)] = 1$, and $F(u) = F(t)$ if and only if $\deg u = 1$. Note that this implies

$$u = \frac{at + b}{ct + d}$$

where $ad - bc \neq 0$. Therefore, deduce that $\mathrm{Gal}(E/F)$ is the set of maps $h(t) \mapsto h(u)$ where u is of the form indicated.
2. Let $F \subseteq K \subseteq E$ and E Galois over F. Prove that E is Galois over K.
3. Show that every element of $K(x_1, \ldots, x_n)$ which is not in K is transcendental over K.
4. Show that in the subgroup-intermediate subfield correspondence given in the fundamental theorem of Galois theory, the subfield corresponding to the intersection of two subgroups H_1 and H_2 is the subfield generated by the composite field $E^{H_1}E^{H_2}$, the smallest subfield of E generated by E^{H_1} and E^{H_2}, and the intersection of two intermediate fields K_1 and K_2 corresponds to the subgroup generated by $\mathrm{Gal}(E/K_1) \cup \mathrm{Gal}(E/K_2)$.
5. Use the fact that any finite group G is isomorphic to a subgroup of S_n (Cayley's theorem) to prove that given any finite group G, there exist fields E and E/F such that $\mathrm{Gal}(E/F) = G$.
6. Let $E = \mathbb{Q}(r)$ where $r^3 + r^2 - 2r - 1 = 0$. Verify that $r' = r^2 - 2$ is also a root of $x^3 + x^2 - 2x - 1 = 0$. Determine $\mathrm{Gal}(E/\mathbb{Q})$. Show that E is normal over $\mathbb{Q}$.
7. Let $\alpha = \sqrt{2 + \sqrt{2}}$ in $\mathbb{R}$, $f(x)$ the minimal polynomial of α over $\mathbb{Q}$ and E is a splitting field of $f(x)$ over $\mathbb{Q}$.
   (a) Compute $f(x)$ and $[E : \mathbb{Q}]$.
   (b) Find $G = \mathrm{Gal}(E/\mathbb{Q})$ and draw a lattice diagram for the subgroup-intermediate subfield correspondence for the fundamental theorem of Galois theory.
8. Let $(\mathbb{Z}/(p))(t)$ where t is transcendental over $\mathbb{Z}/(p)$. Let G be the group of automorphisms generated by the automorphism of E such that $t \mapsto t + 1$. Determine $F = E^G$ and $[E : F]$.

5.5 Some Consequences of Galois Theory

In this section, we shall derive some consequences of Galois theory including another proof of the fundamental theorem of algebra.

Theorem 5.5.1. *Let K be a finite dimensional separable extension of a field F. Then there are only finitely many fields L such that $K \supseteq L \supseteq F$.*

Proof. Since K/F is finite separable, by primitive element theorem, $K = F[\alpha]$ for some $\alpha \in K$. Let E be the splitting field of $m_{\alpha,F}(x)$. Then E is Galois over F and $E \subseteq K \subseteq F$. By fundamental theorem of Galois theory, the number of intermediate fields between E and F is the number of subgroups of $\mathrm{Gal}(E/F)$. Hence, the number of intermediate fields between K and F is at most the number of subgroups of $\mathrm{Gal}(E/F)$. □

Remark. If $G = \mathrm{Gal}(E/F)$, then $K = E^H$ for some subgroup H of G and the fields L such that $K \supseteq L \supseteq F$ are in 1-1 correspondence with the subgroups J of G such that $G \supseteq J \supseteq H$.

The primitive element theorem and the previous theorem both *fail* for inseparable extensions as shown in the following example.

Example 5.5.1. Let F be an infinite field of prime characteristic p and let u and v be indeterminates over F. Consider

$$F(u, v) \supseteq F(u^p, v^p)$$

It is easy to see that $[F(u, v) : F(u^p, v^p)] = p^2$. On the other hand, if $z \in F(u, v)$, then $z^p \in F(u^p, v^p)$, so

$$[F(u^p, v^p)(z) : F(u^p, v^p)] \leq p.$$

Hence, there is no z such that $F(u, v) = F(u^p, v^p)(z)$, that is, no primitive element.

On the other hand, the nonexistence of a primitive element shows that the fields

$$F(u^p, v^p)(u + \alpha v),$$

for $\alpha \in F$, are all distinct. To see this, assume that $F(u^p, v^p)(u + \alpha v) = F(u^p, v^p)(u + \beta v) = E$ for some $\alpha \neq \beta$ in E. Then $u + \alpha v$ and $u + \beta v$ in E, so

$$\alpha(u + \beta v) - \beta(u + \alpha v) = (\alpha - \beta)u \in E.$$

Since $\alpha - \beta \neq 0$, u is in E which implies that v is also in E. Thus, $E = F(u, v)$, a contradiction. Hence, there are infinitely many fields L such that $F(u, v) \supset L \supset F(u^p, v^p)$.

Let us now recall some concepts from group theory. Suppose a group G acts on a set S. The action is *transitive* if for any $s, t \in S$ there is a $g \in G$ such that $gs = t$.

Remark. The action of G being transitive simply means that the action of G on S has only one orbit. Assuming G acts transitively on S. let $s \in S$ and let

$$H = \{g \in G : gs = s\}$$

be the stabilizer of s. Then S can be identified with the set of left cosets

$$\{gH : g \in G\},$$

with G acting by left multiplication. Note that the subgroup H depends on the choice of s and choosing a different s will give a conjugate of H. More precisely, if $s \in S$ and $x \in G$, and

$$H = \text{stabilizer of } s = \{g \in G : gs = s\}$$

then

$$xHx^{-1} = \text{stabilizer of } xs = \{g \in G : g(xs) = xs\}.$$

(If $gs = s$, then $(xgx^{-1})(xs) = xs$.)

A basic example of this phenomenon is the action of S_n on $\{1, 2, \ldots, n\}$. The stabilizer of $i \in \{1, 2, \ldots, n\}$ is $\mathrm{Sym}\{1, \ldots, i-1, i+1, \ldots, n\}$ which may be identified with S_{n-1}, but S_{n-1} has n conjugates in S_n.

Theorem 5.5.2. *Let E be the splitting field over F of a separable polynomial $f(x) \in F[x]$ which is irreducible over F. Then $\mathrm{Gal}(E/F)$ acts transitively on the roots of $f(x)$. Hence, $\mathrm{Gal}(E/F)$ may be identified with a subgroup of $\mathrm{Sym}\{r_1, \ldots, r_n\}$ which acts transitively on $\{r_1, \ldots, r_n\}$, the roots of $f(x)$ in E.*

Proof. This is implicit in the proof of Theorem 5.1.6. For, if r and s are roots of $f(x)$ in E, then

$$F(r) \cong F[x]/(f(x)) \cong F(s) \quad \text{with} \quad r \mapsto x + (f(x)) \mapsto s$$

by an isomorphism which fixes F pointwise. Let $\eta : F(r) \to F(s)$ be this isomorphism. By Theorem 5.1.6, η extends to an isomorphism $\hat{\eta} : E \to E$. Then $\hat{\eta} \in \mathrm{Gal}(E/F)$ and $\hat{\eta}(r) = s$, which is what we need to prove. □

Remarks.

1. The hypothesis that $f(x)$ be irreducible over F is essential. For, example, if $f(x) = f_1(x) \ldots f_k(x)$ where $f_1(x), \ldots, f_k(x)$ are distinct irreducible polynomials, then all one can say is that $\mathrm{Gal}(E/F)$ permutes the roots of each $f_i(x)$ among themselves. It is still true that $\mathrm{Gal}(E/F)$ can be identified with a subgroup of the group of permutations of the roots, but not a transitive one.
2. Assume that $f(x)$ is irreducible and separable over F of degree n, E/F is a splitting field for $f(x)$ over F and r is one root of $f(x)$. Then the fundamental theorem of Galois theory gives the following picture

$$\begin{array}{ccc} E & & \{\mathrm{id}_E\} \\ | & & | \\ F[r] & & \mathrm{Gal}(E/F[r]) = H \\ |\, n & & |\, n \\ F & & \mathrm{Gal}(E/F). \end{array}$$

   Thus, $F[r] = E^H$ where H is a subgroup of index n in G.

The basic Theorems 5.1.4 and 5.1.6 give the existence and uniqueness of splitting fields. That is, if F is a field and $f(x)$ is a monic polynomial in $F[x]$, then

1. A splitting field E for $f(x)$ exists. E is generated over F by the roots of $f(x)$ and $f(x)$ splits into linear factors in $E[x]$.
2. The splitting field E/F is unique up to isomorphism over F. In other words, if E'/F is another splitting field for $f(x)$ over F, then there is an isomorphism

$$\varphi : E \to E'$$

   which is identity on F.

What does this means if we are searching for the splitting field of some $f(x) \in \mathbb{Q}[x]$?

It means that we can realize E as a subfield of $\mathbb{C}$. More precisely, $f(x)$ is a product of linear factors in $\mathbb{C}[x]$, say $f(x) = (x-\alpha_1) \ldots (x-\alpha_k)$ and we can take E to be the field $\mathbb{Q}(\alpha_1, \ldots, \alpha_k) \subseteq \mathbb{C}$. This could be very helpful because it allows us to work in a concrete and explicit field.

The fundamental theorem of algebra (*every $f(x) \in \mathbb{C}[x]$ is a product of linear factors*) is usually proved in complex analysis and there is also a topological proof. Here we present a proof based on Galois theory and the intermediate value theorem from real analysis or calculus. We shall start with some basic results.

Theorem 5.5.3. *Let $f(x) \in \mathbb{R}[x]$ be a polynomial of odd degree. Then $f(x)$ has a root in $\mathbb{R}$.*

Proof. It is enough to prove for a monic polynomial

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$$

with $a_i \in \mathbb{R}$ and n is odd. If $a = |a_0| + \cdots + |a_{n-1}|$, then it is easy to see that $f(a) > 0$ and $f(-a) < 0$. By intermediate value theorem (because $f(x)$ is continuous), there exists $r \in \mathbb{R}$ such that $f(r) = 0$. □

Consider $\alpha + \beta i$ with $\alpha, \beta \in \mathbb{R}$. If $\gamma = \sqrt{\alpha^2 + \beta^2}$, then

$$(\sqrt{(\gamma + \alpha)/2} + i\sqrt{(\gamma - \alpha)/2})^2 = \alpha + \beta i.$$

Hence, we have proved

Theorem 5.5.4. *Every complex number has a square root.*

Theorem 5.5.5. *If K is a field containing $\mathbb{C}$, then $[K : \mathbb{C}] \neq 2$.*

Proof. Suppose conversely that $[K : \mathbb{C}] = 2$ and let $K = \mathbb{C} + \mathbb{C}u$ for some $u \in K$. Then u satisfies a polynomial

$$f(x) = x^2 - bx + c$$

of degree two over $\mathbb{C}$, since $1, u, u^2$ are linearly dependent over $\mathbb{C}$. The roots of $f(x)$ are

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2}$$

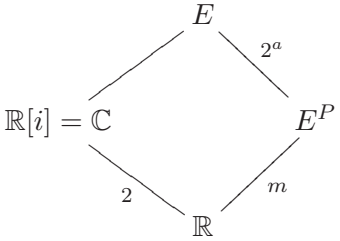
which lie in $\mathbb{C}$, since every element of $\mathbb{C}$ has a square root in $\mathbb{C}$. Thus, $u \in \mathbb{C}$, a contradiction. □

Recall that a finite p-group G is nilpotent, so by Exercise 3.3, a maximal subgroup M of G is normal and $[G : M] = p$, i.e., if G is a nontrivial finite p-group, then G has a normal subgroup of index p.

Theorem 5.5.6. [Fundamental Theorem of Algebra] *Let $f(x) \in \mathbb{C}[x]$. Then $f(x)$ is a product of linear factors in $\mathbb{C}[x]$.*

Proof. Let $^- : \mathbb{C} \to \mathbb{C}$ denote the complex conjugation. Then $g(x) = f(x)\overline{f(x)} \in \mathbb{R}[x]$. Let E be a splitting field for $g(x)(x^2 + 1)$ over $\mathbb{R}$ and identify $\mathbb{C}$ with the subfield of E generated by the roots of $x^2 + 1$. Since the characteristic is zero, all polynomials are separable, so E is the splitting field of a separable polynomial. Hence, E is Galois over $\mathbb{R}$ by Theorem 5.4.4.

Let $G = \mathrm{Gal}(E/\mathbb{R})$, $|G| = 2^a m$, where m is odd, and let P be a Sylow 2-subgroup of G. Consider the diagram of fields



Thus, $E^P = \{\alpha \in E : \varphi(\alpha) = \alpha \text{ for all } \varphi \in P\}$ is an extension of $\mathbb{R}$ of odd degree m, by the fundamental Galois correspondence. If $u \in E^P$, the minimal polynomial $q(x)$ of u over $\mathbb{R}$ is an irreducible polynomial in $\mathbb{R}[x]$ of odd degree, so it has a root in $\mathbb{R}$ by Theorem 5.5.3. Since $q(x)$ is irreducible, it has degree one. Hence, $E^P = \mathbb{R}$ and $G = P$, so $|G| = 2^a$. By the fundamental

theorem of Galois theory, $\mathbb{C} = E^H$ where H is a subgroup of G of index 2. If $H \neq \{1\}$, it has a subgroup K of index 2, so

$$\mathbb{R} \text{ —}^{2}\text{— } \mathbb{C} = E^H \text{ —}^{2}\text{— } E^K \text{ —— } E.$$

Thus, $[E^K : \mathbb{C}] = 2$ which contradicts Theorem 5.5.5. Hence, $|G| = 2$, $H = \{1\}$ and $\mathbb{C} = E^H = E$. Therefore, $\mathbb{C}$ is a splitting field for $g(x)(x^2+1) = f(x)\overline{f(x)}(x^2+1)$ over $\mathbb{R}$, so $g(x)(x^2+1)$ (and hence $f(x)$) splits into linear factors in $\mathbb{C}[x]$. □

The fundamental theorem of algebra was first rigorously proved by Gauss in 1816 (his doctoral dissertation in 1798 provides a proof using geometric considerations requiring some topological justification). There was a proof due to Laplace in 1795. However, Laplace's proof was deemed unacceptable because he assumed the existence of a splitting field for polynomials (i.e., that the roots existed somewhere in some field), which had not been established at that time. The elegant above proof was given by Artin.

5.6 Finite Fields

Let k be a field of q elements. Then $(k, +)$ is an abelian group, so $q \cdot 1 = 0$. Thus, F is of characteristic prime $p > 0$ and $p \mid q$, so it contains $\mathbb{Z}/p\mathbb{Z}$ as a subfield and it is a finite extension of $\mathbb{Z}/p\mathbb{Z}$. Its cardinality $|k| = q = p^d$ is a power of p, with $d = [k : \mathbb{Z}/p\mathbb{Z}]$. This also indicates that the additive group of k is a direct sum of d copies of cyclic group of order p. We shall restate the following fact (Theorem 5.3.7).

Theorem 5.6.1. *$k^\times$ is cyclic of order $q - 1$.*

Some immediate consequences of the above theorem are as follows.

Corollary 5.6.2. *The field k consists of the solutions to $x^q - x = 0$ in an algebraic closure of $\mathbb{Z}/p\mathbb{Z}$ containing k.*

Corollary 5.6.3. *There is an element $\alpha \in k$ such that $k = (\mathbb{Z}/p\mathbb{Z})[\alpha]$, that is, k is a simple extension of the prime field $\mathbb{Z}/p\mathbb{Z}$.*

Corollary 5.6.4. *For each positive divisor r of $q - 1 (= |k^\times|)$ there are exactly $\phi(r)$ elements in $k^\times$ of order r.*

Corollary 5.6.5. *Let p be a prime and d a positive integer. Then, up to isomorphism, there is exactly one field of order $q = p^d$.*

Proof. Let E be a splitting field of $f(t) = t^{p^d} - t$ over $\mathbb{Z}/p\mathbb{Z}$ in an algebraic closure of $\mathbb{Z}/p\mathbb{Z}$. By Theorem 5.1.6, E is unique up to isomorphism. It consists of the roots of $t^{p^d} = t$ in the algebraic closure of $\mathbb{Z}/p\mathbb{Z}$. Thus, $|E|$ is the number of roots of $t^{p^d} - t$. Since $f'(t) = -1$, $f(t)$ is separable, so $|E| = p^d$. Thus, we have constructed a field of order $q = p^d$, namely E, the splitting field of $f(t)$ over $\mathbb{Z}/p\mathbb{Z}$. □

For $q = p^d$, we may write $\mathbb{F}_q$ for the (*unique up to isomorphism*) field of q elements. Also, we may write $\mathbb{F}_p$ for $\mathbb{Z}/p\mathbb{Z}$.

Corollary 5.6.6. *Given any positive integer d, there exists an irreducible polynomial of degree n over $\mathbb{F}_p$.*

Proof. By Corollary 5.6.3, $\mathbb{F}_{p^d} = \mathbb{F}_p[\alpha]$ for some $\alpha \in \mathbb{F}_{p^d}$. Let $f(t)$ be the minimal polynomial of α over $\mathbb{F}_p$. Then $\mathbb{F}_{p^d} = \mathbb{F}_p[\alpha] \cong \mathbb{F}_p[t]/(f(t))$ shows $\deg f(t) = [\mathbb{F}_{p^d} : \mathbb{F}_p] = d$. □

Next, we shall study finite extensions of a finite field. For simplicity, k stands for the finite field $\mathbb{F}_q$. Let k_n be a degree n field extension of k. If k_m is an intermediate field of degree m over k, then k_n is a vector space over k_m, so m divides n. Conversely, any degree m extension of k within an algebraic closure of k with $m \mid n$ is a subfield of k_n by Corollary 5.6.2 since $m \mid n$ implies $(q^m - 1) \mid (q^n - 1)$.

Consider the map σ on k_n which sends x to x^q. From

$$\sigma(x+y) = (x+y)^q = x^q + y^q = \sigma(x) + \sigma(y) \text{ and } \sigma(xy) = (xy)^q = x^q y^q = \sigma(x)\sigma(y),$$

we see that σ is an endomorphism. Furthermore, $\sigma(x) = x^q = 0$ implies $x = 0$. So σ is one-to-one. As k_n is finite, we have shown that σ is an automorphism of k_n. Finally, $\sigma(x) = x^q = x$ for $x \in k$, this shows that $\sigma \in \mathrm{Gal}(k_n/k)$, called the **Frobenius' automorphism**. Let r be the order of σ. Then

$$\sigma^r(x) = x^{q^r} = x \text{ for all } x \in k_n$$

implies $r = n$ since $k_n^\times$ is cyclic of order $q^n - 1$. Hence, $\mathrm{Gal}(k_n/k)$ contains the cyclic group $\langle\sigma\rangle$ of order n. Since $|\mathrm{Gal}(k_n/k)| \leq [k_n : k] = n$, $\mathrm{Gal}(k_n/k) = \langle\sigma\rangle$ and so the field k_n is Galois over k. We record this in

Theorem 5.6.7. *The field k_n is Galois over k with the Galois group $\mathrm{Gal}(k_n/k)$ cyclic of order n, generated by the Frobenius' automorphism σ.*

Note that an element $x \in k_n$ lies in k if and only if it satisfies $x^q = x$, in other words, if and only if it is fixed by the Frobenius' automorphism, or equivalently, by the group $\mathrm{Gal}(k_n/k)$. Using $G = \mathrm{Gal}(k_n/k)$, we define two important maps, called **trace** and **norm**, denoted by $\mathrm{Tr}_{k_n/k}$ and $\mathrm{N}_{k_n/k}$, respectively, from k_n to k as follows:

$$\mathrm{Tr}_{k_n/k} : x \mapsto \sum_{\tau \in G} \tau(x) = \sum_{i=1}^{n} \sigma^i(x),$$

$$\mathrm{N}_{k_n/k} : x \mapsto \prod_{\tau \in G} \tau(x) = \prod_{i=1}^{n} \sigma^i(x).$$

One check easily that the images of trace and norm maps are in k. It is clear that $\mathrm{Tr}_{k_n/k}$ is a homomorphism from the additive group k_n to the additive group k and $\mathrm{N}_{k_n/k}$ is a homomorphism from $k_n^\times$ to $k^\times$. Next we investigate their images. We shall first need

Lemma 5.6.8. *If E is an extension field of a field F, then the automorphisms in $\mathrm{Gal}(E/F)$ are E-linearly independent F-linear transformations.*

Proof. Suppose otherwise. Let $a_1\tau_1 + \cdots + a_r\tau_r = 0$ be a shortest nontrivial linear relation with $a_1, \ldots, a_r \in E^\times$ and $\tau_1, \ldots, \tau_r \in \mathrm{Gal}(E/F)$. Then $r \geq 2$ and τ_i are distinct. Let $y \in E$ be such that $\tau_1(y) \neq \tau_2(y)$. From $\sum_{i=1}^r a_i\tau_i = 0$ we get

$$\sum_{i=1}^{r} a_i\tau_i(yx) = \sum_{i=1}^{r} a_i\tau_i(y)\tau_i(x) = 0$$

for all $x \in E$, so $\sum_{i=1}^{r} a_i\tau_i(y)\tau_i = 0$. This yields another nontrivial relation

$$\sum_{i=1}^{r} a_i\tau_i(y)\tau_i - \tau_1(y)\sum_{i=1}^{r} a_i\tau_i = \sum_{i=2}^{r} a_i(\tau_i(y) - \tau_1(y))\tau_i = 0,$$

which is shorter than the relation we started with, a contradiction. □

Theorem 5.6.9. [Hilbert Theorem 90]
1. *The norm map* $\mathrm{N}_{k_n/k}$ *from* $k_n^\times$ *to* $k^\times$ *is surjective with the kernel consisting of* $x/\sigma(x)$, $x \in k_n^\times$.
2. *The trace map* $\mathrm{Tr}_{k_n/k}$ *from* k_n *to* k *is surjective with the kernel consisting of* $x - \sigma(x)$, $x \in k_n$.

Proof. (1) Since $\mathrm{N}_{k_n/k}(\sigma(x)) = \prod_{i=1}^{n} \sigma^{i+1}(x) = \prod_{i=1}^{n} \sigma^i(x) = \mathrm{N}_{k_n/k}(x)$, so $x/\sigma(x)$ lies in the kernel of the norm map for all $x \in k_n^\times$. Further, $x/\sigma(x) = y/\sigma(y)$ if and only if $xy^{-1} \in k^\times$, hence the elements $x/\sigma(x)$ with $x \in k_n^\times$ form a subgroup of $k_n^\times$ of order $(q^n-1)/(q-1)$. Thus, it is equal to the whole kernel if and only if the norm map is surjective. To see $\mathrm{N}_{k_n/k}$ is onto, observe that

$$\mathrm{N}_{k_n/k}(x) = \prod_{i=1}^{n} \sigma^i(x) = x \cdot x^q \cdot x^{q^2} \cdot \cdots \cdot x^{q^{n-1}} = x^{1+q+q^2+\cdots+q^{n-1}} = x^{(q^n-1)/(q-1)}$$

for all $x \in k_n^\times$. Hence, any generator x of $k_n^\times$ has $\mathrm{N}_{k_n/k}(x)$ of order $q-1$.
(2) Since elements in $\mathrm{Gal}(k_n/k)$ are k-linear maps, the image of $\mathrm{Tr}_{k_n/k}(k_n)$ is a vector space over k, hence $\mathrm{Tr}_{k_n/k}(k_n) = 0$ or k. If $\mathrm{Tr}_{k_n/k} = 0$, then $\sum_{i=1}^{n} \sigma_i = 0$, which is a nontrivial linear relation among elements of $\mathrm{Gal}(k_n/k)$, so impossible by Lemma 5.6.8. Therefore, $\mathrm{Tr}_{k_n/k}$ is surjective. Then its kernel has order q^{n-1}. Clearly, $\mathrm{Tr}_{k_n/k}(\sigma(x)) = \mathrm{Tr}_{k_n/k}(x)$ so that kernel contains $x - \sigma(x)$ for all $x \in k_n$. Further, $x - \sigma(x) = y - \sigma(y)$ if and only if $x - y \in k$, so the group $\{x - \sigma(x) : x \in k_n\}$ has order q^n/q, thus is equal to the kernel. □

Remark. The Hilbert Theorem 90 for norm and trace maps is usually proved using first cohomology group of the Galois group (à la Noether). When the base field is finite, we may use counting argument, as shown above.

Given $z \in k_n$, it defines a k-linear transformation L_z on k_n by $x \mapsto zx$, that is, multiplication by z. The **trace** and **determinant** of L_z are defined as the trace and determinant of any $n \times n$ matrix representing L_z. They are in fact given by $\mathrm{Tr}_{k_n/k}$ and $\mathrm{N}_{k_n/k}$ of z. More precisely, we have

Theorem 5.6.10. *Let* $z \in k_n$ *and define* L_z *as above. Then*
1. $\mathrm{Tr}\, L_z = \mathrm{Tr}_{k_n/k}(z)$ *and* $\det L_z = \mathrm{N}_{k_n/k}(z)$.
2. *Suppose* $k(z) = k_n$. *Let* $f(t) = t^n + a_1t^{n-1} + \cdots + a_{n-1}t + a_n$ *be the minimal polynomial of* z *over* k. *Then*
$$a_1 = -\mathrm{Tr}_{k_n/k}(z) \quad \textit{and} \quad a_n = (-1)^n\mathrm{N}_{k_n/k}(z).$$

Proof. We shall prove (1) and (2) under the assumption (2) and leave (1) for the case $k(z)$ being a proper subfield k_n as an exercise. For each $\tau \in \mathrm{Gal}(k_n/k)$, $0 = \tau(f(z)) = f(\tau(z))$, hence $\tau(z)$ is also a root of $f(x)$. Further, if τ and τ' are two different elements in $\mathrm{Gal}(k_n/k)$, then $\tau(z) \neq \tau'(z)$ (otherwise they would agree on $k(z) = k_n$). This shows that z has n distinct images under $\mathrm{Gal}(k_n/k)$ and they are the roots of $f(t)$. Therefore,

$$-a_1 = \text{the sum of roots of } f(t) = \mathrm{Tr}_{k_n/k}(z)$$

and

$$(-1)^n a_n = \text{the product of roots of } f(t) = \mathrm{N}_{k_n/k}(z).$$

This proves (2). For (1), we know that L_z satisfies $f(t) = 0$. As $f(t)$ is irreducible over k and $[k_n : k] = n$, $f(t)$ is the characteristic polynomial of L_z. The companion matrix attached to L_z is

$$\begin{bmatrix} 0 & & & & & -a_n \\ 1 & 0 & & & & -a_{n-1} \\ & 1 & 0 & & & -a_{n-2} \\ & & & \ddots & & \vdots \\ & & & & 0 & \vdots \\ & & & & 1 & -a_1 \end{bmatrix},$$

which has trace $= -a_1$ and determinant $= (-1)^n a_n$. This proves (1). □

Exercises 5.6.

1. Let $k_6 = \mathbb{F}_{5^6}$ be the field with 15625 elements and let $k = \mathbb{F}_5$ be its prime subfield.
   (a) Determine the cardinality of the set of elements of k_6 which generate k_6 as a field over k.
   (b) Draw a lattice diagram for the subgroup-intermediate subfield correspondence for the fundamental theorem of Galois theory of k_6/k.
2. Let k be a finite field with finite extensions k_m and k_{mn} of degrees m and mn, respectively. Show that
$$\mathrm{Tr}_{k_{mn}/k} = \mathrm{Tr}_{k_m/k} \circ \mathrm{Tr}_{k_{mn}/k_m} \quad \text{and} \quad \mathrm{N}_{k_{mn}/k} = \mathrm{N}_{k_m/k} \circ \mathrm{N}_{k_{mn}/k_m}.$$
3. Let $z \in k_n$. Suppose $k(z) = k_m$ is a proper subfield of k_n. Prove that
$$\mathrm{Tr}\, L_z = \mathrm{Tr}_{k_n/k}(z) = (n/m)\mathrm{Tr}_{k_m/k}(z) \quad \text{and} \quad \det L_z = \mathrm{N}_{k_m/k}(z)^{n/m}.$$
4. (a) (Normal Basis Theorem) There exists an element $z \in k_n$ such that the set $\{\tau(z) : \tau \in \mathrm{Gal}(k_n/k)\}$ is a basis of k_n over k. [Hint: Consider the minimal polynomial of the Frobenius' automorphism σ.]
   (b) For z in (a), we have $\mathrm{Tr}_{k_n/k}(z) \neq 0$. [Hint: Express an element in k_n as a k-linear combination of $\{\tau(z)\}$. Then show $\mathrm{Tr}_{k_n/k}(k_n) = k\mathrm{Tr}_{k_n/k}(z)$.]

Project 24 (Primitive elements in a finite field)**.** The polynomial $p(x) = x^2 - 2$ is irreducible in $\mathbb{Z}_5[x]$. Then $\mathbb{Z}[x]/(x^2 - 2)$ is a field with 25 elements and we denote it by $\mathbb{F}_{25}$. We shall investigate a way to find a primitive element for $\mathbb{F}_{25}$ in this project. Let $\alpha = x + (p(x))$.

(a) Prove that the order of α is 8 and the order of $\alpha + 1$ is 12.
(b) Use (a) to obtain a primitive element (the element of order 24) in $\mathbb{F}_{25}$. (*Hint*. Theorem 1.4.12 is useful.)
(c) Find a primitive element for the fields $\mathbb{Z}_2[x]/(x^4 + x + 1)$, $\mathbb{Z}_3[x]/(x^3 + 2x + 1)$ and $\mathbb{Z}[x]/(x^2 - 2)$.
(d) Write an algorithm to obtain a primitive element for finite fields.

Project 25 (Paley Graph)**.** Let q be an odd prime power and consider the finite field $\mathbb{F}_q$ of q elements. For the additive group $(\mathbb{F}_q, +)$, let $P(q)$ be the Cayley graph $\mathrm{Cay}(\mathbb{F}_q, (\mathbb{F}_q^*)^2)$, called the **Paley graph** (named after Raymond Paley). Here, $(\mathbb{F}_q^*)^2 = \{x^2 : x \in \mathbb{F}_q^*\}$

(a) Prove that -1 is in $(\mathbb{F}_q^*)^2$ if and only if $q \equiv 1 \pmod 4$. (*Hint*. Use Theorem 5.6.1.)
(b) Deduce that the Paley graph is undirected if and only if $q \equiv 1 \pmod 4$.
(c) Assume that $q \equiv 1 \pmod 4$. Then the Paley graph $P(q)$ is a regular graph. What is its degree? Prove also that if x and y are adjacent the number of common neighbors of x and y is $(q - 5)/4$ and if x and y are adjacent the number of common neighbors of x and y is $(q - 1)/4$.
(d) Can we generalize the definition of the Paley graph to a finite local ring and a finite commutative ring? (*Hint*. See Example 3.5.11 and Meemark and Suntornpoch [35].)

5.7 Cyclotomic Extensions

In this section, we shall study other important examples of Galois extension, called cyclotomic fields, and compute their Galois groups. Note that "Cyclotomy" is Greek for the art of dividing a circle into equal parts.

Theorem 5.7.1. *Let K be a field of characteristic 0 and let E be a splitting field of $x^n - 1$ over K. Then $\mathrm{Gal}(E/K)$ is isomorphic to a subgroup of $\mathrm{Aut}\,\mathbb{Z}/(n) \cong (\mathbb{Z}/(n))^\times$. In particular, $\mathrm{Gal}(E/K)$ is abelian.*

Proof. Since $(x^n - 1)' = nx^{n-1} \neq 0$, the roots of $x^n - 1$ (in E) are distinct, say

$$x^n - 1 = (x - 1)(x - \alpha_2) \dots (x - \alpha_n).$$

Then $A = \{z \in E : z^n = 1\} = \{1, \alpha_2, \dots, \alpha_n\}$ is a finite subgroup of $E^\times$, so it is cyclic of order n by Theorem 5.3.7. Any automorphism of E, $\theta : E \to E$ induces an automorphism $\theta : A \to A$, so there is a group homomorphism from $\mathrm{Gal}(E/K)$ to $\mathrm{Aut}\,A$ defined by $\theta \mapsto \theta|_A$. This homomorphism is 1-1 since any automorphism of E/K is completely determined by its action on the roots of $x^n - 1$. Hence, $\mathrm{Gal}(E/K)$ is isomorphic to a subgroup of $\mathrm{Aut}\,A = \mathrm{Aut}\,\mathbb{Z}/(n)$. □

We call a Galois extension field E/F **abelian [cyclic] over** F if $\mathrm{Gal}(E/F)$ is abelian [cyclic]. Hence, the above theorem provides an example of abelian extension.

Our next objective is to show that if E is a splitting field of $x^n - 1$ over $\mathbb{Q}$, then $\mathrm{Gal}(E/\mathbb{Q}) \cong \mathrm{Aut}\,\mathbb{Z}/(n) \cong (\mathbb{Z}/(n))^\times$. We first recall some properties of the cyclic group of order n. Let $\mathbb{Z}/(n) = \langle a \rangle$. Then

1. For each divisor d of n, $\mathbb{Z}/(n)$ has a unique subgroup of order d, generated by $a^{n/d}$.
2. All subgroups of $\mathbb{Z}/(n)$ are as in (1). Thus, the number of subgroups of $\mathbb{Z}/(n)$ is equal to the number of divisors of n.
3. If $x, y \in \mathbb{Z}/(n)$, then

$$\begin{aligned} \langle x \rangle = \langle y \rangle &\iff o(x) = o(y) \\ &\iff \theta(x) = y \text{ for some } \theta \in \mathrm{Aut}\,\mathbb{Z}/(n) \\ &\iff x \text{ and } y \text{ lie in the same orbit under the action of } \mathrm{Aut}\,\mathbb{Z}/(n). \end{aligned}$$

An element ω in a field K is an n**th root of unity** if $\omega^n = 1$, it is a **primitive** n**th root of unity** if $o(\omega) = n$ in $K^\times$, that is, $\omega^n = 1$ and $\omega^m \neq 1$ if $1 \leq m < n$. In the complex numbers $\mathbb{C}$, the nth roots of unity are the powers of

$$\omega = e^{2\pi i/n} = \cos(2\pi/n) + i\sin(2\pi/n) \text{ and } \omega^t = e^{2\pi it/n} = \cos(2\pi t/n) + i\sin(2\pi t/n).$$

Thus, $\mathbb{Q}[\omega]$ is the splitting field of $x^n - 1$ over $\mathbb{Q}$, so $[\mathbb{Q}[\omega] : \mathbb{Q}]$ is the degree of the minimal polynomial of ω over $\mathbb{Q}$. We know that the set U of the nth roots of unity is a cyclic group of order n under multiplication. Hence, the number of primitive nth roots of 1, that is, the number of generators of U, is $\phi(n)$.

For a positive integer d and x an indeterminate, the d**th cyclotomic polynomial**, $\Phi_d(x)$ is the product

$$\Phi_d(x) = \prod \{(x - \varepsilon) : \varepsilon \text{ is a primitive } d\text{th root of unity}\}.$$

If $\eta \in \mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q})$ and z is primitive nth root of unity, then $\eta(z)$ is primitive. Hence, $\eta(\Phi_n(x)) = \Phi_n(x)$ and so $\Phi_n(x) \in \mathbb{Q}[x]$. It is clear that $\Phi_n(x) \mid (x^n - 1)$ and, in fact, since any nth root of unity has an order $d \mid n$ we see that

$$x^n - 1 = \prod_{d|n} \Phi_d(x). \tag{5.7.1}$$

Remark. The formula (5.7.1) provides us with an algorithm for calculating the polynomial $\Phi_n(x)$. To begin with we have

$$\Phi_1(x) = x - 1$$

and assuming we already know the $\Phi_d(x)$ for proper divisors d of n then (5.7.1) gives us $\Phi_n(x)$. For example, for a prime p, $\Phi_1(x)\Phi_p(x) = x^p - 1$, so we get

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1.$$

Then $\Phi_2(x) = x + 1$ and $\Phi_3(x) = x^2 + x + 1$, so

$$\Phi_4(x) = \frac{x^4 - 1}{\Phi_1(x)\Phi_2(x)} = x^2 + 1$$

$$\Phi_6(x) = \frac{x^6 - 1}{\Phi_1(x)\Phi_2(x)\Phi_3(x)} = x^2 - x + 1$$

$$\Phi_{12}(x) = \frac{x^{12} - 1}{\Phi_1(x)\Phi_2(x)\Phi_3(x)\Phi_4(x)\Phi_6(x)} = x^4 - x^2 + 1.$$

Next, we observe that $\Phi_n(x)$ has integer coefficients. This holds for $n = 1$ and assuming it holds for every $\Phi_d(x)$, $d < n$, we have $x^n - 1 = \Phi_n(x)g(x)$ where $g(x) = \prod_{d|n;d<n} \Phi_d(x)$ is a monic polynomial with integer coefficients. The division algorithm gives integral polynomials $q(x)$ and $r(x)$ with $\deg r(x) < \deg g(x)$ such that $x^n - 1 = q(x)g(x) + r(x)$. Since $q(x)$ and $r(x)$ are unique in $\mathbb{Z}[x]$ and $x^n - 1 = \Phi_n(x)g(x)$ in $\mathbb{Q}[x]$, we see that $\Phi_n(x) = q(x) \in \mathbb{Z}[x]$.

We shall now prove

Theorem 5.7.2. *The nth cyclotomic polynomial $\Phi_n(x)$ has integer coefficients and is an irreducible polynomial in $\mathbb{Q}[x]$.*

Proof. Suppose that $\Phi_n(x) = h(x)k(x)$, where $h(x), k(x) \in \mathbb{Z}[x]$ and $h(x)$ is irreducible in $\mathbb{Z}[x]$, hence, in $\mathbb{Q}[x]$ (Gauss' lemma). We may also assume that $h(x)$ and $k(x)$ are monic and so $\deg h(x) \geq 1$. Let p be a prime integer not dividing n and let δ be a root of $h(x)$. Since $(p, n) = 1$, δ^p is a primitive nth root of unity. Assume that δ^p is not a root of $h(x)$. Then δ^p is a root of $k(x)$; consequently δ is a root of $k(x^p)$. Since $h(x)$ is irreducible and has δ as a root also, $(h(x), k(x^p)) \neq 1$ and thus $h(x) \mid k(x^p)$. It follows (as mentioned earlier) that $k(x^p) = h(x)l(x)$, where $l(x)$ is monic with integral coefficients. Since $x^n - 1 = \Phi_n(x)g(x)$, we have $x^n - 1 = h(x)k(x)g(x)$. We now pass to congruences modulo p or, which is the same thing, to equations in $(\mathbb{Z}/(p))[x]$. This gives

$$x^n - \bar{1} = \bar{h}(x)\bar{k}(x)\bar{g}(x) \tag{5.7.2}$$

where, in general, if $f(x) = a_0x^m + a_1x^{m-1} + \cdots + a_m \in \mathbb{Z}[x]$, then $\bar{f}(x) = \bar{a}_0x^m + \bar{a}_1x^{m-1} + \cdots + \bar{a}_m, \bar{a}_i = a_i + (p)$ in $\mathbb{Z}/(p)$. Similarly, we have $\bar{k}(x^p) = \bar{h}(x)\bar{l}(x)$. Now, using $\bar{a}^p = \bar{a}$ for any $a \in \mathbb{Z}$, we see that

$$\begin{aligned}\bar{f}(x)^p &= (\bar{a}_0x^m + \bar{a}_1x^{m-1} + \cdots + \bar{a}_m)^p \\ &= \bar{a}_0^p x^{pm} + \bar{a}_1^p x^{p(m-1)} + \cdots + \bar{a}_m^p \\ &= \bar{a}_0 x^{pm} + \bar{a}_1 x^{p(m-1)} + \cdots + \bar{a}_m \\ &= \bar{f}(x^p)\end{aligned}$$

for any $f(x) \in \mathbb{Z}[x]$. Thus, $\bar{k}(x)^p = \bar{k}(x^p) = \bar{h}(x)\bar{l}(x)$ which implies that $(\bar{h}(x), \bar{k}(x)) \neq 1$. Then (5.7.2) shows that $x^n - \bar{1}$ has multiple roots in its splitting field over $\mathbb{Z}/(p)$. Since the derivative $(x^n - \bar{1})' = \bar{n}x^{n-1}$ and $\bar{n} \neq 0$, we have $(x^n - \bar{1}, (x^n - \bar{1})') = \bar{1}$, contrary to the derivative criterion for multiple roots. This contradiction shows that δ^p is a root of $h(x)$ for every prime $p \not| n$. A repetition of this shows that δ^r is a root of $h(x)$ for every integer r prime to n. Since every primitive nth root of 1 has the form δ^r, $(r, n) = 1$, we see that $h(x)$ is divisible by every $x - \delta'$, δ' primitive. Hence, $h(x) = \Phi_n(x)$ and $\Phi_n(x)$ is irreducible in $\mathbb{Q}[x]$. □

As an immediate consequence of Theorem 5.7.2, we get

Theorem 5.7.3. *Let ω be a primitive nth root of unity. Then*
1. *$\Phi_n(x)$ is the minimal polynomial of ω over $\mathbb{Q}$.*
2. *$[\mathbb{Q}[\omega] : \mathbb{Q}] = \deg \Phi_n(x) = \phi(n)$, the Euler's ϕ-function.*
3. *$\mathbb{Q}[\omega]$ is the splitting field of $\Phi_n(x)$ over $\mathbb{Q}$.*
4. *$\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q}) \cong (\mathbb{Z}/(n))^\times$.*

Proof. (1), (2) and (3) are obvious. To prove (4), recall that by Theorem 5.7.1, $\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q})$ is isomorphic to a subgroup of $(\mathbb{Z}/(n))^\times$. Since $[\mathbb{Q}[\omega] : \mathbb{Q}] = \phi(n) = |(\mathbb{Z}/(n))^\times|$, it must be isomorphic to all of $(\mathbb{Z}/(n))^\times$. □

Theorem 5.7.3 implies that $\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q})$ is isomorphic to the multiplicative group U_n of units of the ring $\mathbb{Z}/(n)$. If n is a prime then we know that this is a cyclic group of order $p-1$. Moreover, if $n = p_1^{e_1} p_2^{e_2} \dots p_s^{e_s}$, p_i distinct primes, then U_n is isomorphic to the direct product of the groups $U_{p^{e_i}}$. In addition, we know the structures of U_{p^e} from the knowledge of primitive roots in number theory as follows.

Theorem 5.7.4. [Structure of U_{p^e}]
1. *U_2 and U_4 are cyclic and if $e > 3$, then U_{2^e} is a direct product of a cyclic group of order 2 and one of order 2^{e-2}.*
2. *If p is an odd prime, the multiplicative group U_{p^e} of units of $\mathbb{Z}/(p^e)$ is cyclic.*

Example 5.7.1. If $\omega = e^{2\pi i/72}$ is a primitive 72^{nd} root of unity, then

$$\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q}) \cong U_{72} \cong \mathbb{Z}/(2) \times \mathbb{Z}/(2) \times \mathbb{Z}/(6).$$

A finite-dimensional field extension of $\mathbb{Q}$ is called a **cyclotomic field** if it is a subfield of $\mathbb{Q}[\omega]$ for some root of unity ω.

Theorem 5.7.5. *Let K be a cyclotomic field. Then K is Galois over $\mathbb{Q}$ and $\mathrm{Gal}(K/\mathbb{Q})$ is abelian.*

Proof. Consider $\mathbb{Q} \subset K \subset \mathbb{Q}[\omega]$ for some nth root of unity ω. By the fundamental theorem of Galois theory $K = \mathbb{Q}[\omega]^H$ for some subgroup H of $G = \mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q}) \cong (\mathbb{Z}/(n))^\times$. Since G is abelian, H is normal in G, so the fundamental theorem says that K is Galois over $\mathbb{Q}$ with Galois group G/H, an abelian group. □

Remark. A deep theorem of Kronecker and Weber says that the converse of Theorem 5.7.5 is true, namely, *"if K is Galois over $\mathbb{Q}$ and $\mathrm{Gal}(K/\mathbb{Q})$ is abelian, then K is a cyclotomic field, that is, $K \subset \mathbb{Q}[\omega]$ for some root of unity ω."*

Example 5.7.2. Let $\omega = e^{2\pi i/71}$ be a primitive 71^{st} root of unity. Then

$$G = \mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q}) \cong U_{71} \cong \mathbb{Z}/(70) \cong \mathbb{Z}/(2) \times \mathbb{Z}/(5) \times \mathbb{Z}/(7).$$

Let $H = \mathbb{Z}/(2) \times \mathbb{Z}/(5)$ be the subgroup of G of order 10. Then H is normal in G and consequently we have $\mathbb{Q}[\omega]^H$ is a Galois extension over $\mathbb{Q}$ of degree $[\mathbb{Q}[\omega]^H : \mathbb{Q}] = [G : H] = 7$ and $\mathrm{Gal}(\mathbb{Q}[\omega]^H/\mathbb{Q}) \cong G/H \cong \mathbb{Z}/(7)$.

We now have enough tools to find the Galois groups of splitting fields of irreducible separable polynomials $x^n - a$. Note that $(x^n - a)' = nx^{n-1}$, so $x^n - a$ is separable over a field F if and only if $\mathrm{char}\, F \not| n$. In particular, if F contains a primitive nth root of unity, then $\mathrm{char}\, F \not| n$.

Theorem 5.7.6. *Let F be a field which contains a primitive nth root of unity ω, i.e., $\operatorname{char} F$ not divide n. Let $a \in F$, $f(x) = x^n - a$, E the splitting field for E over F and r a root of $f(x)$ in E. Then*
(1) The factorization of $f(x)$ in $E[x]$ is

$$x^n - a = (x - r)(x - \omega r) \dots (x - \omega^{n-1} r)$$

and $E = F[r]$.
(2) Let d be the least positive integer such that $r^d = b \in F$. Then d divides n and

$$x^d - b = (x - r)(x - \varepsilon r) \dots (x - \varepsilon^{d-1} r)$$

is the minimal polynomial of r over F where $\varepsilon = \omega^{n/d}$, a primitive dth root of unity. In addition, $[E : F] = d$ and $\operatorname{Gal}(E/F) \cong \mathbb{Z}/(d)$. The automorphism $\alpha : E \to E$ defined by $\alpha(r) = \varepsilon r$ generates $\operatorname{Gal}(E/F)$.

Proof. (1) Since $r, \omega r, \dots, \omega^{n-1} r$ are all roots of $x^n - a$, $(x - r)(x - \omega r) \dots (x - \omega^{n-1} r)$ must divide $x^n - a$. Since both polynomials are monic of degree n, they must be equal. Also, $\omega \in F$ by hypothesis, so $F[r]$ contains all the roots of $x^n - a$ and is generated over F by them. Hence, $E = F[r]$ by the definition of splitting field.
(2) Since d is the generator of the group $\{m \in \mathbb{Z} : r^m \in F\}$ and n is in this group, d divides n. Certainly, r is a root of $x^d - b \in F[x]$. If $x^d - b$ had a proper factor of degree c, $0 < c < d$, looking at its constant term would show that $r^c \in F$, contradicting the minimality of d. Thus, $x^d - b$ is irreducible. Hence, $[E : F] = [F[r] : F] = d$, so $|\operatorname{Gal}(E/F)| = d$. On the other hand, one sees that $\alpha^i(r) = \varepsilon^i r$, so α is an element of $\operatorname{Gal}(E/F)$ of order d. Therefore, $\operatorname{Gal}(E/F) = \langle \alpha \rangle \cong \mathbb{Z}/(d)$. □

For the sake of clarity, we reformulate Theorem 5.7.6 slightly to emphasize the case where $f(x)$ is irreducible, which is the important one.

Theorem 5.7.7. *Let F be a field which contains a primitive nth root of unity ω and let $a \in F$. Then $x^n - a$ is irreducible if and only if no divisor d of n, $d \neq 1$, such that $a = b^d$ for some $b \in F$. If $x^n - a$ is irreducible and E/F is its splitting field, then $[E : F] = n$ and $\operatorname{Gal}(E/F) \cong \mathbb{Z}/(n)$.*

Example 5.7.3. Let $f(x) = x^n - p \in \mathbb{Q}[x]$ where p is prime. (The essential point is not that p is prime, but that it is not a proper power.) By Eisenstein's criterion $f(x)$ is irreducible over $\mathbb{Q}$. If we let $r = \sqrt[n]{p}$ denote the positive real nth root of p and $\omega = e^{2\pi i/n}$, a primitive nth root of unity, then the factorization of $f(x)$ in $\mathbb{C}[x]$ is

$$x^n - p = (x - r)(x - \omega r) \dots (x - \omega^{n-1} r).$$

Now let $E = \mathbb{Q}[r, \omega r, \dots, \omega^{n-1} r]$ be a splitting field for $f(x)$, and let $\varphi \in \operatorname{Gal}(E/\mathbb{Q})$. Then φ permutes $\{r, \omega r, \dots, w^{n-1} r\}$ and φ is completely defined by its action on the set $\{r, \omega r, \dots, \omega^{n-1} r\}$. This gives rise to an embedding

$$\operatorname{Gal}(E/\mathbb{Q}) \hookrightarrow S_n = \operatorname{Sym}\{r, \omega r, \dots, \omega^{n-1} r\}.$$

Note that $\omega = (\omega r) r^{-1}$, so $\omega \in E$. This makes it clear that

$$E = \mathbb{Q}[r, \omega r, \dots, \omega^{n-1} r] = \mathbb{Q}[\omega, r] = \mathbb{Q}[\omega][r].$$

Thus, E is generated over $\mathbb{Q}$ by two elements ω and r. We also know that E can be generated over $\mathbb{Q}$ by a primitive element. However, using such an element would not simplify the description of $\operatorname{Gal}(E/\mathbb{Q})$.

Now consider $\varphi \in \mathrm{Gal}(E/\mathbb{Q})$. Then

$$\varphi(\omega) = \omega^i \quad \text{and} \quad \varphi(r) = \omega^j r$$

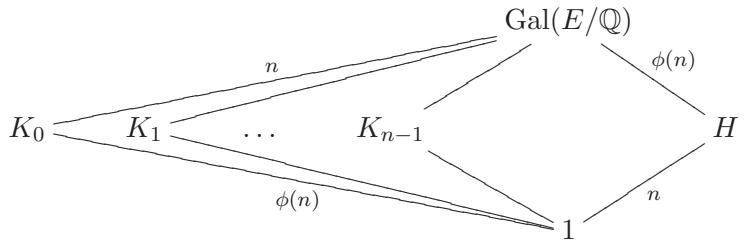
for some $1 \leq i \leq n-1$ such that $\gcd(i,n) = 1$ and $0 \leq j \leq n-1$. The choice of i and j completely determines φ and it turns out that all of the above choices do determine automorphisms of E. Thus,

$$|\mathrm{Gal}(E/\mathbb{Q})| = n \cdot \phi(n).$$

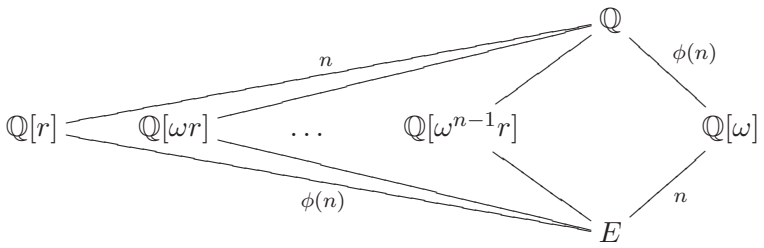
To describe $\mathrm{Gal}(E/\mathbb{Q})$ more precisely, let $\mathbb{Q}[\omega] = E^H$, and for $0 \leq j \leq n-1$, let $\mathbb{Q}[\omega^j r] = E^{K_j}$. Since $\mathbb{Q}[\omega]$ is Galois over $\mathbb{Q}$, H is normal in $\mathrm{Gal}(E/\mathbb{Q})$. Moreover, by Theorem 5.7.7, $H = \mathrm{Gal}(E/\mathbb{Q}[\omega]) = \langle \tau \rangle \cong \mathbb{Z}/(n)$ is cyclic of order n with generator τ defined by

$$\tau(\omega) = \omega \quad \text{and} \quad \tau(r) = \omega r.$$

The group K_j are more difficult to describe explicitly, but they are all conjugate in $\mathrm{Gal}(E/\mathbb{Q})$ and isomorphic as abstract groups to $\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q}) \cong (\mathbb{Z}/(n))^\times$. We have the following diagram of subgroups of $\mathrm{Gal}(E/\mathbb{Q})$ which does *not* include all subgroups.



The corresponding invariant fields are



As a group, $\mathrm{Gal}(E/\mathbb{Q})$ is a semi-direct product $H \rtimes K_i$ for any i.

We conclude this section with the statement of the following theorem on the Galois group of splitting fields of irreducible separable polynomials $x^n - a$ without proof.

Theorem 5.7.8. *Let $F[\omega]$ be a splitting field for $x^n - 1$ over F where ω is a primitive nth root of unity. Suppose that $a \in F$ and $f(x) = x^n - a$ is irreducible over F and let E be a splitting field for $f(x)$ over F. Let d be the largest divisor of n such that $b^d = a$ for some $b \in F[\omega]$ (possibly $d = 1$). Let $G = \mathrm{Gal}(E/F)$ and $H = \mathrm{Gal}(E/F[\omega])$. Then H is cyclic of order d and normal in G, $\mathrm{Gal}(F[\omega]/F) \cong G/H$ is isomorphic to a subgroup of $(\mathbb{Z}/(n))^\times$ and G is isomorphic to a semi-direct product of H by G/H.*

Using the cyclotomic polynomials, we now present the proof of Wedderburn's theorem as follows.

Theorem 5.7.9. [Wedderburn, 1909] *A finite division ring is a field.*

Proof. Let D be a finite division ring. Then the center of D, denoted by F, is a finite field (see Exercises 2.1). Assume that $|F| = q$. Since D is a vector space over F, $|D| = q^n$ for some $n \in \mathbb{N}$. Also, for an element $d \in D$, the set $C(d) = \{r \in D : rd = dr\}$ is a division ring containing F and $|C(d)| = q^m$ for some $m \leq n$, which is strictly less than if $d \notin F$. Thus, the class equation (Corollary 1.3.8) for the multiplicative group $D \setminus \{0\}$ is

$$q^n - 1 = |F \setminus \{0\}| + \sum_{i=1}^{s} [D \setminus \{0\} : C(d_i) \setminus \{0\}] = q - 1 + \sum_{i=1}^{s} \frac{q^n - 1}{q^{m_i} - 1},$$

where $d_1, d_2, \ldots, d_s$ represent the conjugacy classes of $D \setminus \{0\}$ which contains more than one element and $|C(d_i)| = q^{m_i}$ for some $m_i < n$ for all i. Because each $(q^n - 1)/(q^{m_i} - 1) = [D \setminus \{0\} : C(d_i) \setminus \{0\}]$ is an integer, m_i is a proper divisor of n. Thus, the quotient

$$\frac{x^n - 1}{\Phi_n(x)(x^{m_i} - 1)}$$

is a polynomial in $\mathbb{Z}[x]$. Substitute q for x, we see that $\Phi_n(q)$ divides $(q^n - 1)/(q^{m_i} - 1)$. It follows from the class equation that $\Phi_n(q)$ divides $q - 1$ because it divides all the other terms. Then $|\Phi_n(q)| \leq q - 1$. On the other hand, since 1 is the closest point, on the unit circle $\{z \in \mathbb{C} : |z| = 1\}$, to the positive integer q, we have that for every primitive nth root of unity ω^j,

$$|q - \omega^j| \geq q - 1 \geq 1,$$

and the first inequality is strict unless $\omega^j = 1$, that is, unless 1 is a primitive nth root of unity which means $n = 1$. So the product $|\Phi_n(q)|$ of the $|q - \omega^j|$'s is greater than or equal to $q - 1$, with equality only if $n = 1$. Because $|\Phi_n(q)|$ is both at most $q - 1$ and at least $q - 1$, we get $|\Phi_n(q)| = q - 1$ and hence $n = 1$. Therefore, $|D| = q = |C(D)|$, so $D = C(D)$ which implies D is commutative as desired. □

Given a field F and a polynomial $p(x) \in F[x]$, we say that $p(x)$ is **solvable by radicals over** F if we can find a finite sequence of fields $F_1 = F(\omega_1)$, $F_2 = F_1(\omega_2)$, ..., $F_k = F_{k-1}(\omega_k)$ such that $\omega_1^{r_1} \in F$, $\omega_2^{r_2} \in F_1$, ..., $\omega_k^{r_k} \in F_{k-1}$ and all roots of $p(x)$ lie in F_k.

If K is the splitting field of $p(x)$ over F, then $p(x)$ is solvable by radicals over F if we can find a finite sequence of fields as above such that $K \subseteq F_k$. An important remark, and one we shall use later, in the proof of Theorem 5.7.10, is that if such an F_k can be found, we can, without loss of generality, assume it to be a normal extension of F. We leave its proof as an exercise.

Theorem 5.7.10. [Galois] *Let F be a field which contains a primitive nth root of unity for every positive integer n. If a polynomial $p(x) \in F[x]$ is solvable by radical over F, then the Galois group over F of $p(x)$ is solvable.*

Proof. Let K be the splitting field of $p(x)$ over F. Since $p(x)$ is solvable by radicals, there exists a finite sequence of fields

$$F = F_0 \subset F_1 = F(\omega_1) \subset F_2 = F_1(\omega_2) \subset \ldots \subset F_k = F_{k-1}(\omega_k),$$

where $\omega_1^{r_1} \in F$, $\omega_2^{r_2} \in F_1$, ..., $\omega_k^{r_k} \in F_{k-1}$ and $K \subseteq F_k$ such that F_k is normal over F. As a normal extension of F, F_k is also a normal of any intermediate fields, hence F_k is a normal extension of each F_i. Theorem 5.7.6 implies that F_i is a normal extension of F_{i-1} and $\mathrm{Gal}(F_i/F_{i-1})$ is abelian for all i. Thus, by the Galois correspondence, $\mathrm{Gal}(F_k/F_i)$ is a normal subgroup in $\mathrm{Gal}(F_k/F_{i-1})$. Consider the normal series

$$\mathrm{Gal}(F_k/F_0) \supset \mathrm{Gal}(F_k/F_1) \supset \mathrm{Gal}(F_k/F_2) \supset \ldots \supset \mathrm{Gal}(F_k/F_{k-1}) \supset \{1\}.$$

Since $\mathrm{Gal}(F_k/F_{i-1})/\mathrm{Gal}(F_k/F_i) \cong \mathrm{Gal}(F_i/F_{i-1})$ is abelian for all i, $\mathrm{Gal}(F_k/F)$ is solvable. It follows that $\mathrm{Gal}(K/F) \cong \mathrm{Gal}(F_k/F)/\mathrm{Gal}(F_k/K)$ is solvable by Theorem 3.2.4 (2). □

We make two remarks without proof.

1. The converse of Theorem 5.7.10 is also true; that is, if the Galois group of $p(x)$ over F is solvable, then $p(x)$ is solvable by radicals over F.
2. Theorem 5.7.10 and its converse are true even if F does not contain roots of unity.

Recall that for $n \geq 5$, S_n is not solvable. Thus we have

Corollary 5.7.11. *The general polynomial of degree $n \geq 5$ over $\mathbb{Q}$ is not solvable by radical.*

Exercises 5.7. 1. Prove the following statements. (a) If p is a prime number, then $\Phi_{p^k}(x) = \Phi_p(x^{p^{k-1}})$.
(b) If $n > 1$ is odd, then $\Phi_{2n}(x) = \Phi_n(-x)$.
(c) If p is a prime number, then $\Phi_{pn}(x) = \begin{cases} \dfrac{\Phi_n(x^p)}{\Phi_n(x)}, & \text{if } p \not| n, \\ \Phi_n(x^p), & \text{if } p|n. \end{cases}$

2. Let $\omega = e^{2\pi i/18}$ be a primitive 18th root of unity.
(a) Find the minimal polynomial of ω over $\mathbb{Q}$.
(b) Draw a lattice diagram for the subgroup-intermediate subfield correspondence for the fundamental theorem of Galois theory of $\mathbb{Q}[\omega]/\mathbb{Q}$.
3. Give an example of field E containing the field of rational numbers $\mathbb{Q}$ such that E is Galois over $\mathbb{Q}$ and $\mathrm{Gal}(E/\mathbb{Q})$ is a cyclic group of order five.
4. Let K be a finite separable extension over F and E its normal closure (smallest normal extension over F containing K).
(a) Prove that $[E : F]$ is finite. (b) If $\mathrm{Gal}(E/F)$ is abelian, show that K is normal over F.
5. If $p(x)$ is solvable by radicals over F, prove that we can find a finite sequence of fields

$$F \subset F_1 = F(\omega_1) \subset F_2 = F_1(\omega_2) \subset \ldots \subset F_k = F_{k-1}(\omega_k),$$

where $\omega_1^{r_1} \in F$, $\omega_2^{r_2} \in F_1$, $\ldots$, $\omega_k^{r_k} \in F_{k-1}$ containing all the roots of $p(x)$ such that F_k is *normal* over F.
6. Assume that $x^p - a$, $a \in \mathbb{Q}$, is irreducible in $\mathbb{Q}[x]$. Show that the Galois group of $x^p - a$ over $\mathbb{Q}$ is isomorphic to the group of transformations of $\mathbb{Z}/(p)$ of the form $y \mapsto ky + l$ where $k, l \in \mathbb{Z}/(p)$ and $k \neq 0$.

Project 26 (Insolvability of a quintic)**.** Consider $g(x) = 3x^5 - 15x + 5$. By Eisenstein's criterion, $g(x)$ is irreducible over $\mathbb{Q}$.

(a) Use the intermediate value theorem to show that $g(x)$ has a real root between -2 and -1 and also has a real root between 0 and 1 and between 1 and 2.
(b) Use Rolle's theorem to assure that there is no other real roots. Hence, the other two roots of $g(x)$ are non real complex numbers, say $a + bi$ and $a - bi$.
(c) Let K be the splitting field of $g(x)$ in $\mathbb{C}$. Show that $\mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to S_5.
(d) Since S_5 is not solvable, deduce that $g(x)$ is not solvable by radical by Galois (Theorem 5.7.10).
(e) Give another example of irreducible polynomial in $\mathbb{Z}[x]$ of degree five that is solvable by radical, compute the Galois group of its splitting field over $\mathbb{Q}$ and show that this group is solvable.

5.8 Normal Bases

Let E be an extension field of a field F. We have known from Lemma 5.6.8 that the automorphism in $\mathrm{Gal}(E/F)$ are E-linearly independent F-linear transformations.

Theorem 5.8.1. *If E/F is a finite Galois extension with Galois group $G = \{1, \sigma_2, \ldots, \sigma_n\}$. Then $\{u_1, u_2, \ldots, u_n\}$ is a basis for E/F if and only if*

$$\det \begin{bmatrix} u_1 & u_2 & \ldots & u_n \\ \sigma_2(u_1) & \sigma_2(u_2) & \ldots & \sigma_2(u_n) \\ \vdots & & & \\ \sigma_n(u_1) & \sigma_n(u_2) & \ldots & \sigma_n(u_n) \end{bmatrix} \neq 0.$$

Proof. Call the above matrix M and suppose that $\det M = 0$. Since $M \in M_n(E)$, there are $\alpha_1, \alpha_2, \ldots, \alpha_n \in E$, not all zero, such that

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_n \end{bmatrix} M = \vec{0}.$$

This translates to $\theta(u_1) = \theta(u_2) = \cdots = \theta(u_n) = 0$ where

$$\theta = \alpha_1 1 + \alpha_2 \sigma_2 + \cdots + \alpha_n \sigma_n : E \to E.$$

But $\theta : E \to E$ is a F-linear map, so θ is the zero map, since it vanishes on $u_1, u_2, \ldots, u_n$. Since $1, \sigma_2, \ldots, \sigma_n$ are linearly independent over K, Lemma 5.6.8 says that $\theta \neq 0$, so we have a contradiction.

Conversely, if $u_1, u_2, \ldots, u_n$ are not a basis for E/F, then there are $\beta_1, \beta_2, \ldots, \beta_n \in F$, not all zero, such that

$$u_1\beta_1 + u_2\beta_2 + \cdots + u_n\beta_n = 0.$$

Then for any $\sigma_i \in G$,

$$\sigma_i(u_1)\beta_1 + \sigma_i(u_2)\beta_2 + \cdots + \sigma_i(u_n)\beta_n = \sigma_i(u_1\beta_1 + u_2\beta_2 + \cdots + u_n\beta_n) = 0,$$

so $M \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_n \end{bmatrix}^T = \vec{0}$. Hence, $\det M = 0$. □

Note that if $|K| = q$, then $\alpha^q - \alpha = 0$ for all $\alpha \in K$, so $f(x) = x^q - x$ is a nonzero polynomial but it is a zero function. The next theorem says that such a polynomial cannot exist if K is infinite.

Theorem 5.8.2. *Let F be an infinite field and $F \subseteq E$. If $f(x_1, \ldots, x_n)$ is a nonzero polynomial in $E[x_1, \ldots, x_n]$, then there exist $\alpha_1, \ldots, \alpha_n \in F$ such that $f(\alpha_1, \ldots, \alpha_n) \neq 0$.*

Proof. We shall use induction on n. For $n = 1$, since $f(x_1)$ has only finitely many roots and F is infinite, there is $\alpha_i \in F$ such that $f(\alpha_1) \neq 0$. Assume that the statement holds for n, and let

$$f(x_1, \ldots, x_{n+1}) = f_0(x_1, \ldots, x_n) + f_1(x_1, \ldots, x_n)x_{n+1} + \cdots + f_t(x_1, \ldots, x_n)x_{n+1}^t.$$

Since $f \neq 0$, at least one of $f_0(x_1, \ldots, x_n), \ldots, f_t(x_1, \ldots, x_n)$ is nonzero, so there are $\alpha_1, \ldots, \alpha_n \in F$ such that $f(\alpha_1, \ldots, \alpha_n, x_{n+1}) \neq 0$ in $E[x_{n+1}]$. By the one variable case, there is $\alpha_{n+1} \in F$ such that $f(\alpha_1, \ldots, \alpha_n, \alpha_{n+1}) \neq 0$. □

Theorem 5.8.3. *Let F be an infinite field and E/F Galois with Galois group $G = \mathrm{Gal}(E/F) = \{1, \sigma_2, \ldots, \sigma_n\}$. Suppose that $0 \neq f(x_1, \ldots, x_n) \in F[x_1, \ldots, x_n]$ where $x_1, \ldots, x_n$ are indeterminates over F. Then there exists $u \in E$ such that $f(u, \sigma_2(u), \ldots, \sigma_n(u)) \neq 0$.*

Proof. Let $\{u_1, \ldots, u_n\}$ be a basis for E/F. By Theorem 5.8.1, the matrix

$$M = \begin{bmatrix} u_1 & u_2 & \ldots & u_n \\ \sigma_2(u_1) & \sigma_2(u_2) & \ldots & \sigma_2(u_n) \\ \vdots & & & \\ \sigma_n(u_1) & \sigma_n(u_2) & \ldots & \sigma_n(u_n) \end{bmatrix} \in M_n(E)$$

is invertible. This means that the map on $E[x_1, \ldots, x_n]$ defined by

$$g(x_1, \ldots, x_n) \mapsto g(u_1x_1 + \cdots + u_nx_n, \ldots, \sigma_n(u_1)x_1 + \cdots + \sigma_n(u_n)x_n)$$

is an isomorphism. Thus,

$$h(x_1, \ldots, x_n) = f(u_1x_1 + \cdots + u_nx_n, \ldots, \sigma_n(u_1)x_1 + \cdots + \sigma_n(u_n)x_n)$$

is a nonzero polynomial in $E[x_1, \ldots, x_n]$. By Theorem 5.8.2, there are $a_1, \ldots, a_n$ in F such that $h(a_1, \ldots, a_n) \neq 0$. Let $u = u_1a_1 + \cdots + u_na_n$, this translates to

$$\begin{aligned} 0 \neq h(a_1, \ldots, a_n) &= f(u_1a_1 + \cdots + u_na_n, \ldots, \sigma_n(u_1)a_1 + \cdots + \sigma_n(u_n)a_n) \\ &= f(u, \sigma_2(u), \ldots, \sigma_n(u)), \end{aligned}$$

since $\sigma_i(u_1)a_1 + \cdots + \sigma_i(u_n)a_n = \sigma_i(u_1a_1 + \cdots + u_na_n) = \sigma_i(u)$. □

Consider $E = \mathbb{Q}[i]$ is a Galois extension over $\mathbb{Q}$. Its Galois group is of order two and consists of the identity map and the complex conjugation. A basis over $\mathbb{Q}$ for it is $\{1, i\}$. This basis is not invariant under the Galois action, namely after acting by the complex conjugation, we obtain $\{1, -i\}$. We are showing the existence of a basis for a finite Galois extension which forms a single orbit under the action of the Galois group. For example, for $\mathbb{Q}[i]$, we may use $\{1 + i, 1 - i\}$. In the case of finite fields, this means that each of the basis elements is related to any one of them by applying the Frobenius' automorphism repeatedly.

Let E/F be Galois with Galois group $G = \mathrm{Gal}(E/F) = \{\sigma_1, \ldots, \sigma_n\}$. A **normal basis** for E/F is a basis of the form $\{\sigma_1(u), \ldots, \sigma_n(u)\}$ for some $u \in E$. Eisenstein conjectured the existence of a normal basis in 1850 for finite extensions of finite fields and Hensel gave a proof for finite fields in 1888. Dedekind used such bases in number fields in his work on the discriminant in 1880, but he had no general proof. (See the quote by Dedekind on the bottom of page 51 of Curtis's "Pioneers of Representation Theory: Frobenius, Burnside, Schur, and Brauer".) In 1932 Noether gave a proof for some infinite fields while Deuring gave a uniform proof for all fields (also in 1932). This basis is frequently used in cryptographic applications that are based on the discrete logarithm problem such as elliptic curve cryptography. Hardware implementations of normal basis arithmetic typically have far less power consumption than other bases.

Theorem 5.8.4. [Normal Basis Theorem] *Let E/F be a Galois extension with Galois group $G = \mathrm{Gal}(E/F) = \{\sigma_1, \ldots, \sigma_n\}$. Then E/F has a normal basis.*

Proof. We shall assume that F is infinite and leave the finite case as an exercise (see Exercise 5.6). Let $u \in E$. By Theorem 5.8.1, $\{\sigma_1(u), \sigma_2(u), \ldots, \sigma_n(u)\}$ is a basis for E/F if and only if

$$\det \begin{bmatrix} \sigma_1^2(u) & \sigma_1\sigma_2(u) & \ldots & \sigma_1\sigma_n(u) \\ \sigma_2\sigma_1(u) & \sigma_2^2(u) & \ldots & \sigma_2\sigma_n(u) \\ \vdots & & & \\ \sigma_n\sigma_1(u) & \sigma_n\sigma_2(u) & \ldots & \sigma_n^2(u) \end{bmatrix} \neq 0.$$

Note that the entries in each row or column of the above matrix, call M, are a permutation of the elements $\sigma_1(u), \ldots, \sigma_n(u)$. In other words, each $\sigma_i(u)$ occurs exactly once in each row and column of M. Thus,

$$M = \sigma_1(u)A_1 + \cdots + \sigma_n(u)A_n$$

where each A_i is a permutation matrix (a matrix with a single entry 1 in each row and column and the remaining entries zero). Since $\det A_i = \pm 1$, we see by inspection that if $x_1, \ldots, x_n$ are indeterminates over E

$$f(x_1, \ldots, x_n) = \det(x_1A_1 + \cdots + x_nA_n) = \pm x_1^n \pm \cdots \pm x_n^n + \text{ other terms}$$

In particular, $f(x_1, \ldots, x_n)$ is a nonzero polynomial in $E[x]$. By Theorem 5.8.3, there is a $\bar{u} \in E$ such that $f(\sigma_1(\bar{u}), \ldots, \sigma_n(\bar{u})) \neq 0$. This translates to

$$0 \neq f(\sigma_1(\bar{u}), \ldots, \sigma_n(\bar{u})) = \det(\sigma_1(\bar{u})A_1 + \cdots + \sigma_n(\bar{u})A_n) = \det M.$$

Hence, $\sigma_1(\bar{u}), \ldots, \sigma_n(\bar{u})$ is a desired normal basis for E/F. □

Exercises 5.8.
1. Determine a normal basis for the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$ by using the Galois group in Example 5.4.4.
2. Determine a normal basis for the cyclotomic field $\mathbb{Q}(e^{2\pi i/p})$ over $\mathbb{Q}$ where p is a prime number.

5.9 Transcendental Extensions

Most of extension fields seen in the previous section are algebraic. In this section, we shall present some results on transcendental extension. The final theorem, namely Lüroth's theorem, has many applications in algebraic geometry and function field theory.

Let F be a subfield of a field E and let $x_1, x_2, \ldots$ be independent indeterminates over E. An element $z \in E$ is **transcendental over** F if the homomorphism $F[x_1] \to E$ defined by $f(x_1) \mapsto f(z)$ is one-to-one. We call $z \in E$ **algebraic over** F if it is not transcendental over F. A finite set $\{z_1, \ldots, z_n\} \subset E$ is **algebraically independent over** F if the homomorphism $F[x_1, \ldots, x_n] \to E$ defined by $f(x_1, \ldots, x_n) \mapsto f(z_1, \ldots, z_n)$ is one-to-one. (Note that the empty set is algebraically independent since $F \hookrightarrow E$ is one-to-one.) An arbitrary subset Z of E is **algebraically independent over** F if all of its finite subsets are algebraically independent. A subset Z of E is **algebraically dependent** if it is not algebraically independent.

Remarks.
1. If z is transcendental over F, then $F[z] \cong F[x_1]$, so $F[z]$ is not a field and $F[z]$ is infinite dimensional over F.
2. If z is algebraic over F, then $F[z] \cong F[x_1]/(f(x_1))$ where $f(x_1)$ is the minimal polynomial of z over F. Thus, $F[z] = F(z)$ is a field and $F[z]$ is finite dimensional over F.

Example 5.9.1. Let $F \subset F(y, z) \subset E$ where y and z are independent indeterminates over F. Then $\{y^2, z^2\}$ is an algebraically independent set but $\{y^2, yz, z^2\}$ is not (for, if $f(x_1, x_2, x_3) = x_1x_3 - x_2^2$, then $f(y^2, yz, z^2) = 0$).

Let F be a subfield of a field E. E is **algebraic over** F if each element of E is algebraic over F. E is **purely transcendental over** F if it is isomorphic (by an isomorphism which is the identity on F) to $F(\{x_\alpha\})$ where $\{x_\alpha\}$ is a (possibly infinite) set of independent indeterminates.

Theorem 5.9.1. *Let F be a subfield of a field E.*
1. *There exists a subset X of E (possibly X is empty) such that*
   (a) *X is algebraically independent over F.*
   (b) *X is maximal among algebraically independent sets, in the sense: If $X \subseteq Y \subseteq E$ and $X \neq Y$, then Y is not algebraically independent.*
2. *$F(X)$ is purely transcendental over F and E is algebraic over $F(X)$.*

$$\begin{array}{c} E \\ \Big| \ \textit{algebraic} \\ F(X) \\ \Big| \ \textit{purely transcendental} \\ F \end{array}$$

Proof. (1) Let $\mathscr{S} = \{X \subseteq E : X \text{ is algebraically independent}\}$. Since the empty set is algebraically independent, $\mathscr{S}$ is nonempty. Let $\{X_\alpha\}_{\alpha\in\Lambda}$ be a chain in $\mathscr{S}$. Let $\{z_1, \ldots, z_n\} \subseteq \bigcup_{\alpha\in\Lambda} X_\alpha$. Then $\forall i, \exists \alpha_i \in \Lambda, z_i \in X_{\alpha_i}$. Since $\{X_\alpha\}_{\alpha\in\Lambda}$ is a chain, we may rearrange α_i so that there exists $j \in \Lambda$ such that $z_i \in X_{\alpha_j}$ for all i. Since X_{α_i} is algebraically independent, so is $\{z_1, \ldots, z_n\}$. Thus, $\bigcup_{\alpha\in\Lambda} X_\alpha$ is an upper bound of this chain in $\mathscr{S}$. By Zorn's Lemma, $\mathscr{S}$ has a maximal element, say X. Hence, $F(X)$ is purely transcendental over F. The maximality of X implies that E must be algebraic over F.
(2) The definition of algebraically independent means that $F(X)$ is purely transcendental over F. Consider $z \in E$. If $z \in X \subset F(X)$, then z is algebraic over $F(X)$. If $z \notin X$, the set $X \cup \{z\}$ is algebraically dependent, so for some n there is a nonzero polynomial $f(x_1, \ldots, x_n, x_{n+1})$ ($x_1, \ldots, x_{n+1}$ are indeterminates over F) and $a_1, \ldots, a_n \in X$ such that $f(a_1, \ldots, a_n, z) = 0$. The polynomial $f(x_1, \ldots, x_n, x_{n+1})$ cannot be a polynomial in only $x_1, \ldots, x_n$, since $\{a_1, \ldots, a_n\}$ is an algebraically independent set. Write

$$f(x_1, \ldots, x_n, x_{n+1}) = f_0(x_1, \ldots, x_n) + f(x_1, \ldots, x_n)x_{n+1} + \cdots + f_r(x_1, \ldots, x_n)x_{n+1}^r.$$

Thus, $f(a_1, \ldots, a_n, x_{n+1}) \in F(X)[x_{n+1}]$ is a nonzero polynomial having z as a root, so z is algebraic over $F(X)$. Hence, E is algebraic over $F(X)$. □

Remark. There is no uniqueness for the field $F(X)$. For example, if $E = F(t)$ where t is an indeterminate, then we can take $X = \{p(t)/q(t)\}$ where $p(t)/q(t)$ is any element of E which is not in F. In this case $[E : F(p(t)/q(t))] = n$ where $n = \max\{\deg p(t), \deg q(t)\}$ (Theorem 5.9.3). However, we shall see shortly that the number of elements in the set X is independent of particular set X.

Let F be a subfield of E. A maximal algebraically independent (over F) subset of E is called a **transcendence basis** for E/F.

Remark. By Theorem 5.9.1, a transcendence basis for E/F exists. It may be empty, which happens precisely when E is algebraic over F. Also, E is purely transcendental over F if it has a transcendence base B such that $E = F(B)$.

Theorem 5.9.2. *Let F be a subfield of E. Then any two transcendence bases for E/F have the same cardinality.*

We call the number of elements of transcendence bases of E the **transcendence degree** of E/F. For example, an algebraic extension has transcendence degree zero; $F(x)$ has transcendence degree one over F; in general, $F((x_\alpha)_{\alpha\in\Lambda})$ has transcendence degree $|\Lambda|$ over K.

The purely transcendental extension fields E/F, especially those having a finite transcendence degree, appear to be the simplest type of extension fields. It is clear that such a field is isomorphic to the field of fractions $F(x_1, \ldots, x_n)$ of the polynomial ring $F[x_1, \ldots, x_n]$ in indeterminates $x_1, \ldots, x_n$. Even though these fields look quite innocent, there are difficult and unsolved problems particularly on the nature of the subfields of $F(x_1, \ldots, x_n)/F$. The one case where the situation is quite simple is that in which E has transcendence degree one. We shall consider this case and close this chapter.

Let $E = F(t)$, t transcendental, and let $u \in E, \notin F$. We can write $u = f(t)/g(t)$ where $f(t), g(t) \in F[t]$ and $(f(t), g(t)) = 1$. If n is the larger of the degrees of $f(t)$ and $g(t)$, then we can write

$$f(t) = a_0 + a_1t + \cdots + a_nt^n \quad \text{and} \quad g(t) = b_0 + b_1t + \cdots + b_nt^n,$$

$a_i, b_i \in F$, and either a_n or $b_n \neq 0$. We have $f(t) - ug(t) = 0$, so

$$(a_n - ub_n)t^n + (a_{n-1} - ub_{n-1})t^{n-1} + \cdots + (a_0 - ub_0) = 0 \tag{5.9.1}$$

and $a_n - ub_n \neq 0$ since either $a_n \neq 0$ or $b_n \neq 0$ and $u \notin F$. Thus, (5.9.1) shows that t is algebraic over $F(u)$ and $[F(t) : F(u)] \leq n$. We prove the following more precise result.

Theorem 5.9.3. *Let $E = F(t)$, t transcendental over F, and let $u \in F(t), \notin F$. Write $u = f(t)/g(t)$ where $(f(t), g(t)) = 1$, and let $n = \max\{\deg f(t), \deg g(t)\}$. Then u is transcendental over F, t is algebraic over $F(u)$, and $[F(t) : F(u)] = n$. Moreover, the minimal polynomial of t over $F(u)$ is a multiple in $F(u)$ of $f(x, u) = f(x) - ug(x)$.*

Proof. Put $f(x, y) = f(x) - yg(x) \in F[x, y]$, x, y indeterminates. This polynomial in x and y is of first degree in y and it has no factor $h(x)$ of positive degree since $(f(x), g(x)) = 1$. Thus, it is irreducible in $F[x, y]$. Now t is algebraic over $F(u)$ so if u were algebraic over F, then t would be algebraic over F, contrary to the hypothesis. Hence, u is transcendental over F. Then $F[x, u] \cong F[x, y]$ under the isomorphism over F fixing x and mapping u into y and hence $f(x, u)$ is irreducible in $F[x, u]$. It turns out that $f(x, u)$ is irreducible in $F(u)[x]$. Since $f(t, u) = f(t) - ug(t) = 0$, it follows that $f(x, u)$ is a multiple in $F(u)[x]$ of the minimal polynomial of t over $F(u)$. Therefore, $[F(t) : F(u)]$ is the degree in x of $f(x, u)$. This degree is n, so the proof is complete. □

We can determine all of the subfields E/F for $E = F(t)$, t transcendental: These have the form $F(u)$ for some u. This important result is called the Lüroth's Theorem. Lüroth proved it in case $K = \mathbb{C}$ in 1876. It was first proved for general fields K by Steinitz in 1910, by the following argument.

Theorem 5.9.4. [Lüroth] *If $E = F(t)$, t transcendental over F, then any subfield K of E/F, $K \neq F$, has the form $F(u)$, u transcendental over F.*

Proof. Let $v \in K, \notin F$. Then we have seen that t is algebraic over $F(v)$. Thus, t is algebraic over K. Let $f(x) = x^n + k_1x^{n-1} + \cdots + k_n$ be the minimal polynomial of t over K, so the $k_i \in K$ and $n = [F(t) : K]$. Since t is not algebraic over F, some $k_j \notin F$. We shall show that $K = F(u), u = k_j$. We can write $u = g(t)/h(t)$ where $g(t), h(t) \in F[t], (g(t), h(t)) = 1$ and $m = \max\{\deg g(t), \deg h(t)\} > 0$. Then, by Theorem 5.9.3, $[E : F(u)] = m$. Since $K \supset F(u)$ and $[E : K] = n$, we evidently have $m \geq n$ and equality holds if and only if $K = F(u)$. Now t is a root of the polynomial $g(x) - uh(x) \in K[x]$. Hence, we have a $q(x) \in K[x]$ such that

$$g(x) - uh(x) = q(x)f(x). \tag{5.9.2}$$

The coefficient k_i of $f(x)$ is in $F(t)$, so there exists a nonzero polynomial $c_0(t)$ of least degree such that $c_0(t)k_i = c_i(t) \in F[t]$ for $1 \leq i \leq n$. Then $c_0(t)f(x) = f(x, t) = c_0(t)x^n + c_1(t)x^{n-1} + \cdots + c_n(t) \in F[x, t]$, and $f(x, t)$ is primitive as a polynomial in x, that is, the $c_i(t)$ are relatively prime. The x-degree of $f(x, t)$ is n. Since $k_j = g(t)/h(t)$ with $(g(t), h(t)) = 1$, the t-degree of $f(x, t)$ is $\geq m$. Now replace u in (5.9.2) by $g(t)/h(t)$ and the coefficients of $q(x)$ by their expressions in t. There exist, therefore, $\varphi(t)$ and $q(x, t) \in F[x, t]$ such that

$$\varphi(t)[g(x)h(t) - g(t)h(x)] = f(x, t)q(x, t).$$

Since the coefficients $c_0(t), c_1(t), \ldots, c_n(t)$ of $f(x, t)$ have no common factor, we know that $\varphi(t)$ divides $q(x, t)$. Hence, we may assume $\varphi(t) = 1$. It turns out that there exists a polynomial $q'(x, t) \in F[x, t]$ such that

$$g(x)h(t) - g(t)h(x) = f(x, t)q'(x, t).$$

Since the t-degree of the left-hand side is $\leq m$ and that of $f(x, t)$ is $\geq m$, it follows that this degree is m and $q'(x, t) = q'(x) \in F[x]$. Then the right-hand side is primitive as a polynomial in x and so is the left-hand side. By symmetry the left-hand side is primitive as a polynomial in t also. Hence, $q'(x) = q' \in F$. Thus, $f(x, t)$ has the same x-degree and t-degree so $m = n$, which implies that $K = F(u)$. □

Exercises 5.9.
1. Prove that there is no intermediate field K with $\mathbb{Q} \subseteq K \subsetneq \mathbb{C}$ with $\mathbb{C}$ purely transcendental over K.
2. Prove that a purely transcendental proper extension of a field is never algebraically closed.
3. Let $E = F(t, v)$, where t is transcendental over F and $v^2 + t^2 = 1$. Show that E is purely transcendental over F.

Project 27 (More on Lüroth's theorem)**.** Prove more general fact that if $F \subseteq L \subseteq E$ and E is finitely generated over F (finite transcendence degree), then L is also finitely generated over F. We can ask more generally about minimal numbers of generators of finitely-generated extensions. For instance, suppose $K \subsetneq L \subseteq K(x_1, \ldots, x_n)$ where the x_i are algebraically independent over K. If L/K has transcendence degree 1, then $L = K(\alpha)$, This was proved for $K = \mathbb{C}$ by Gordan in 1887, and for arbitrary K by Igusa in 1951. If $\mathbb{C} \subsetneq L \subseteq \mathbb{C}(x_1, \ldots, x_n)$ where $L/\mathbb{C}$ has transcendence degree 2, then $L = \mathbb{C}(\alpha, \beta)$. This was proved by Castelnuovo in 1894. All known proofs are difficult. The result is not true in general for other types of fields K, such as $\mathbb{Q}$ or $\mathbb{R}$. Finally, there are fields L with $\mathbb{C} \subsetneq L \subsetneq \mathbb{C}(x_1, x_2, x_3)$ such that $L/\mathbb{C}$ has transcendence degree 3 but cannot be generated by three elements.

Bibliography

General References

[1] D. S. Dummit and R. M. Foote, *Abstract Algebra*, 2nd edn, Prentice-Hall Inc., London, 1999.

[2] J. B. Fraleigh, *A First Course in Abstract Algebra*, 7th edn, Addison Wesley, New York, 2002.

[3] P. Grillet, *Algebra*, 2nd edn, Springer, New York, 2007.

[4] T. W. Hungerford, *Algebra*, Springer, New York, 1974.

[5] I. M. Isaacs, *Algebra, a graduate course*, Brooks/Cole Publishing Company, Pacific Grove, 1993.

[6] N. Jacobson, *Basic Algebra I and II*, W. H. Freeman & Co, 1996.

[7] A. W. Knapp, *Basic Algebra*, Birkhäuser, Boston, 2006.

[8] A. W. Knapp, *Advanced Algebra*, Birkhäuser, Boston, 2007.

[9] S. Lang, *Algebra*, 3rd edn, Springer, New York, 2002.

[10] W. K. Nicholson, *Introduction to Abstract Algebra*, John Wiley & Sons, Inc., New Jersey, 2007.

[11] J. J. Rotman, *Advanced Modern Algebra*, Prentice Hall, 2002.

Technical References

[12] M. F. Atiyah and I. G. MacDonald, *Introduction to Commutative Algebra*, Addison-Wesley Publishing Co, Reading, Mass.-London-Don Mills, Ontario, 1969.

[13] D. M. Burton, *Elementary Number Theory*, 7th edn, McGraw-Hill Higher Education, Dubuque, 2010.

[14] C. Godsil and G. Royle, *Algebraic Graph Theory*, Spinger, New York, 2001.

[15] T. Head, *Modules: A Primer of Structure Theorems*, Brooks/Cole Publishing Company, Monterey, 1974.

[16] W.-C. Winnie Li, *Number Theory with Applications*, World Scientific, Singapore, 1996.

[17] B. R. McDonald, *Finite Rings with Identity*, Marcel Dekker, New York, 1974

[18] M. Reid, *Undergraduate Commutative Algebra*, Cambridge University Press, New York, 1995.

[19] I. Stewart, *Galois Theory*, 3rd edn, Chapman & Hall/CRC mathematics, Boca Raton, 2004.

[20] D. J. Winter, *The Structure of Fields*, Springer, New York, 1974.

Research Articles

[21] R. A. Brualdi, *Energy of a Graph*, `http://www.public.iastate.edu/lhogben/energyB.pdf`.

[22] R. D. Carmichael, Note on a new number theory function, *Bull. Amer. Math. Soc.* 16 (1910) 232–238.

[23] Z. Gu and Z. Wan, Orthogonal graphs of odd characteristic and their automorphisms, *Finite Fields Appls.* 14 (2008) 291–313.

[24] Z. Gu, Subconstituents of symplectic graphs modulo p^n, *Linear Algebra Appl.* 439 (2013) 1321–1329.

[25] I. Gutman, *The Energy of a Graph: Old and New Results, Algebraic Combinatorics and Applications*, Springer, Berlin, 2001.

[26] I. M. Isaacs and M. R. Pournaki, Generalizations of FermatâĂŹs little theorem using group theory, *Amer. Math. Monthly* 112 (2005), 734–740.

[27] D. Kiani, M.M.H. Aghaei, Y. Meemark and Suntornpoch B., Energy of unitary Cayley graphs and gcd-graphs, *Linear Algebra Appl.* 435 (2011) 1336–1343.

[28] W. Klingenberg, Symplectic groups over local rings, *Amer. J. Math.* 85 (1963) 232–240.

[29] W. Klotz and T. Sander, Some properties of unitary Cayley graphs, *The Electronic J. Comb.*, 14 (2007), #R45.

[30] F. Li, K. Wang and J. Guo, More on symplectic graphs modulo p^n, *Linear Algebra Appl.* 438 (2012) 2651–2660.

[31] F. Li, K. Wang and J. Guo, Symplectic graphs modulo pq, *Discrete Math.* 313 (2013), 650–655.

[32] Y. Meemark and T. Prinyasart, On symplectic graphs modulo p^n, *Discrete Math.* 311 (2011) 1874–1878.

[33] Y. Meemark and T. Puirod, Symplectic graphs over finite local rings, *Europ. J. Combin.* 34 (2013) 1114–1124.

[34] Y. Meemark and T. Puirod, Symplectic graphs over finite commutative rings, *Europ. J. Combin.* 41 (2014) 298–307.

[35] Y. Meemark and B. Suntornpoch, Eigenvalues and energy of restricted unitary Cayley graphs induced from the square mapping, *ScienceAsia* 39 No.6 (2013) 649–652.

[36] Y. Meemark and N. Wiroonsri, The quadratic digraph on polynomial rings over finite fields, *Finite Fields Appl.* 16 (2010) 334–346.

[37] Y. Meemark and N. Wiroonsri, The digraph of the kth power mapping of the quotient ring of polynomials over finite fields, *Finite Fields Appl.* 18 (2012) 179–191.

[38] J. W. Sander and T. Sander, The energy of integral circulant graphs with prime power order, *Appl. Anal. Discrete Math.*, 5 (2011), pp. 22–36.

[39] W. So, Integral circulant graphs, *Discrete Math.*, 306 (2006), 153–158.

[40] L. Somer and M. Křížek, On a connection of number theory with graph theory, *Czechoslovak Math J.*, 54 (2004) 465–485.

[41] L. Somer and M. Křížek, Structure of digraphs associated with quadratic congruences with composite moduli, *Discrete Math.*, 306 (2006) 2174–2185.

[42] Z. Tang and Z. Wan, Symplectic graphs and their automorphisms, *Europ. J. Combin.* 27 (2006) 38–50.

Index