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.

3.1.1. Definition. 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.

3.1.2. Examples.
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.

3.1.3. Definition. A subnormal [normal] series $\{K_j\}$ is a **refinement** of a subnormal [normal] series $\{H_i\}$ of a group G if $\{H_i\} \subseteq \{K_j\}$.

3.1.4. Example. The series $\{0\} < 72\mathbb{Z} < 9\mathbb{Z} < 3\mathbb{Z} < \mathbb{Z}$ is a refinement of the series $\{0\} < 9\mathbb{Z} < \mathbb{Z}$.

3.1.5. Definition. 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.

3.1.6. Example. 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.

3.1.7. Theorem. [Schreier] Two subnormal [normal] series of a group G have isomorphic refinements.

3.1.8. Example. 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.

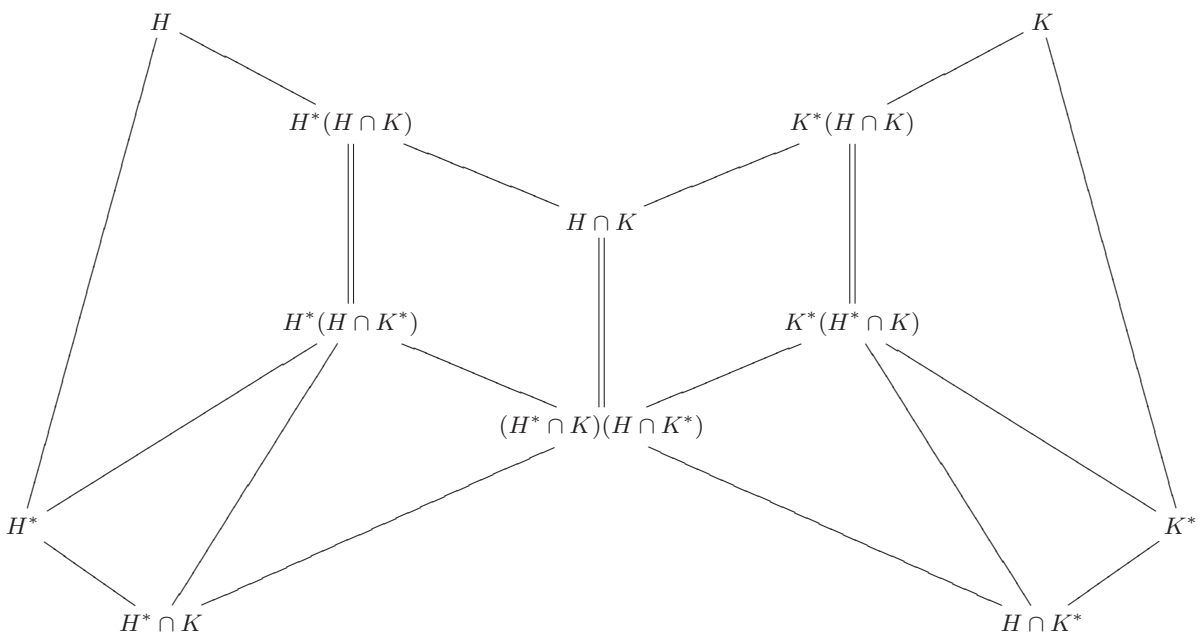
Recall the following fact.

3.1.9. Theorem. 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.

3.1.10. Lemma. [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 3.1.9 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 $0 \le i \le n-1$, we form the chain of (not necessarily distinct) groups

$$H_i = H_i(H_{i+1} \cap K_0) \le H_i(H_{i+1} \cap K_1) \le \cdots \le H_i(H_{i+1} \cap K_m) = H_{i+1}.$$

We refine the first subnormal series by inserting the above chain between H_i and H_{i+1}. In a symmetric fashion, for $0 \le j \le m-1$, we insert the chain

$$K_j = K_j(K_{j+1} \cap H_0) \le K_j(K_{j+1} \cap H_1) \le \cdots \le K_j(K_{j+1} \cap H_n) = K_{j+1}$$

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 $0 \le i \le n-1$ and $0 \le j \le m-1$. 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. □

3.1.11. Definition. 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.

3.1.12. Theorem. 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 1.5.8).

3.1.13. Theorem. M is a maximal normal subgroup of a group G if and only if G/M is simple.

3.1.14. Definition. 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 3.1.13 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:

3.1.15. Theorem. If G is a finite group, then G has a composition series.

Note that by Theorem 3.1.13 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.

3.1.16. Theorem. [Jordan-Hölder] Any two composition [principal] series of a group G are isomorphic.

Proof. Let $\{H_i\}$ and $\{K_j\}$ be two composition [principal] series of G. By Schreier's theorem, they have isomorphic refinements. But since all factor groups are already simple, Theorem 3.1.13 shows that neither series has any further refinement. Hence, $\{H_i\}$ and $\{K_j\}$ must already be isomorphic. □

3.1.17. Examples. (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 primes 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.

3.1. Exercises.

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 $s_1 s_2 \ldots s_n$.
7. Show that an infinite abelian group can have no composition series.

3.2 Solvable Groups

3.2.1. Definition. 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 $n \geq 1$.

3.2.2. Theorem. 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, $N \triangleleft G$. 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. □

3.2.3. Remark. The quotient G/G' is the largest abelian homomorphic image of G.

3.2.4. Definition. 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 5.8.

3.2.5. Definition. 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 can deduce that every characteristic subgroup is normal in G.

3.2.6. Lemma. 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 $\varphi \in \operatorname{Aut} G$. □

3.2.7. Theorem. 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. □

3.2.8. Remark. From Lemma 3.2.6, 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 $G^{(n)} = \{e\}$.

3.2.9. Examples. (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 5.8) 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 1.6). 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.

3.2.10. Theorem. 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 3.2.6 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.

3.2.11. Theorem. 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).

3.2. Exercises. 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, then we let $[A, B]$ be 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 $[A, B] = [B, A]$.

3.3.1. Example. $G' = [G, G]$, $G^{(2)} = [G', G']$, $\ldots$, $G^{(n+1)} = [G^{(n)}, G^{(n)}]$.

3.3.2. Definition. 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.

3.3.3. Definition. A group G is said to be **nilpotent** of class $\leq n$ if $\Gamma_{n+1} = \{e\}$.

3.3.4. 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

$$G \text{ is abelian} \Leftrightarrow G \text{ is solvable of length} \leq 1 \Leftrightarrow G \text{ is nilpotent of class} \leq 1.$$

3.3.5. Examples. (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.

3.3.6. Theorem. 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. □

3.3.7. 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.

3.3.8. 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

$$\{e\} = Z_0(G) \leq Z_1(G) \leq Z_2(G) \leq \ldots$$

and $Z_n(G)$ is called the **n-th term of the upper series of** G.

3.3.9. 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 $i \in \mathbb{N}$.

3.3.10. Definition. A subnormal series $G = G_1 \geq G_2 \geq \ldots$ is called a **central series** for G if $[G, G_i] \leq G_{i+1}$ for all i.

3.3.11. 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:

$$G = Z_n(G) \geq Z_{n-1}(G) \geq \cdots \geq Z_1(G) \geq Z_0(G) = \{e\}.$$

Now, we wish to collect equivalence definitions of a nilpotent group in terms of lower central series, upper central series and central series.

3.3.12. Theorem. Let G be a group.
1. If $G = G_1 \geq G_2 \geq G_3 \geq \ldots$ is a central series for G, then $G_n \geq \Gamma_n(G)$ for all n.
2. G has a central series $G = G_1 > G_2 > \cdots > G_{n+1} = \{e\}$ 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 $G = G_1 \leq G_2 \leq \ldots \leq G_{n+1} = \{e\}$ 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 $G = \Gamma_1(G) \geq \ldots \geq \Gamma_{n+1}(G) = \{e\}$ is a central series of the required length. □

3.3.13. Theorem. 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 3.3.12. □

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 3.3.12 and 3.3.13 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 3.3.12 and 3.3.13 as follows.

3.3.14. Theorem. 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 1.7.5.

3.3.15. Theorem. 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.

3.3.16. Theorem. 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. □

3.3.17. Theorem. 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:

3.3.18. Theorem. 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.

3.3.19. Theorem.
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 3.2.10 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 1.8.16. This theorem characterizes all finite nilpotent groups.

3.3.20. Theorem. [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) $\Rightarrow$ (ii). Assume that G is nilpotent. Recall Theorem 1.7.14 that if P is a Sylow p-subgroup, then $N(N(P)) = N(P)$. But Theorem 3.3.18 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) $\Rightarrow$ (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) $\Rightarrow$ (i). A finite p-group is nilpotent (Theorem 3.3.16) and a finite direct product of nilpotent groups is nilpotent (Theorem 3.3.19). 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 1.8.16.

3.3.21. Corollary. A finite abelian group is the direct product of its Sylow subgroups.

3.3. Exercises.

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 $[x, y^{-1}, z]^y [y, z^{-1}, x]^z [z, x^{-1}, y]^x = e$. Here, $[x, y] = x^{-1}y^{-1}xy$ and $a^g = g^{-1}ag$.
   (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.

---

17. Project. 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).

3.4.1. Definition. 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 the quotient group $\mathrm{GL}_n(K)/\mathrm{SL}_n(K)$ isomorphic to $K^\times = K \smallsetminus \{0\}$.

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.

3.4.2. Theorem. Let K be a field. Then

$$Z(\mathrm{GL}_n(K)) = \{\lambda I_n : \lambda \in K^\times\} \quad \text{and} \quad Z(\mathrm{SL}_n(K)) = \{\lambda I_n : \lambda \in K \text{ and } \lambda^n = 1\},$$

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.

3.4.3. Definition. 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.

3.4.4. Theorem. 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. □

3.4.5. Corollary. 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 3.4.4. □

3.4.6. Lemma. 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. □

3.4.7. Theorem. Let K be a field. The elementary matrices $X_{ij}(a)$, defined in the proof of Theorem 3.4.2, generate $\mathrm{SL}_n(K)$.

Proof. If $n = 1$, then $\mathrm{SL}_1(K) = \{1\}$ is trivial. Lemma 3.4.6 gives the case $n = 2$. For $n > 2$, the theorem follows from the mathematical induction in a similar manner. □

3.4.8. Lemma. The elementary matrices $X_{ij}(a)$, defined in the proof of Theorem 3.4.2, 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. □

3.4.9. Corollary. 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 $\mathrm{SL}_2(\mathbb{F}_3)$.

3.4.10. Remark. $\mathrm{SL}_2(\mathbb{F}_2) \cong \mathrm{PSL}_2(\mathbb{F}_2) \cong S_3$, $\mathrm{SL}_2(\mathbb{F}_3) \ncong S_4$ and $\mathrm{PSL}_2(\mathbb{F}_3) \cong A_4$. However, 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.

3.4.11. Theorem. [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. $\square$

3.4. Exercises. 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.

---

18. Project. 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.

3.5.1. Examples. 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. $\langle a_1, \dots, a_k : a_1^{n_1} = \dots = a_k^{n_k} = e, a_i a_j a_i^{-1} a_j^{-1} = e \text{ if } i \neq j \rangle \cong \mathbb{Z}_{n_1} \times \dots \times \mathbb{Z}_{n_k}$.
4. $\langle a_1, \dots, a_k : a_i a_j a_i^{-1} a_j^{-1} = e \text{ if } i \neq j \rangle \cong \mathbb{Z} \times \dots \times \mathbb{Z}$ (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.

3.5.2. Definition. 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 $i \in \mathbb{N}$.

3.5.3. Definition. 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_1w_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.

3.5.4. Example. 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.

3.5.5. Example. $F_2 = \langle x, y\rangle$. 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. $\{x^i : i \in \mathbb{Z} \smallsetminus \{0\}\}$
3. $\{y^i : i \in \mathbb{Z} \smallsetminus \{0\}\}$
4. $\{x^{i_1}y^{j_1} \dots x^{i_k}y^{j_k} : k > 0, i_r, j_r \in \mathbb{Z} \smallsetminus \{0\}\}$
5. $\{x^{i_1}y^{j_1} \dots x^{i_k}y^{j_k}x^{i_{k+1}} : k > 0, i_r, j_r \in \mathbb{Z} \smallsetminus \{0\}\}$
6. $\{y^{j_1}x^{i_1} \dots y^{j_k}x^{i_k} : k > 0, i_r, j_r \in \mathbb{Z} \smallsetminus \{0\}\}$
7. $\{y^{j_1}x^{i_1} \dots y^{j_k}x^{i_k}y^{j_{k+1}} : k > 0, i_r, j_r \in \mathbb{Z} \smallsetminus \{0\}\}$

3.5.6. Definition. 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.

3.5.7. Theorem. [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 $\theta \circ \Phi = \mathrm{id}_{F[A]}$, 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

3.5.8. Corollary. 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. □

3.5.9. Corollary. 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 $\phi = \mathrm{id}_H$), 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.

3.5.10. Theorem. 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 in the next chapter.

If G is free on a set A, the number of elements in A is called the **rank of** G.

3.5.11. Theorem. Two free groups are isomorphic if and only if they have the same rank.

3.5.12. Theorem. [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).

3.5.13. Example. 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!

3.5.14. Definition. Let $G \xrightarrow{\theta} H \xrightarrow{\phi} K$ be a sequence of groups homomorphisms. We say that it is **exact** at H if $\mathrm{im}\,\theta = \ker \phi$. 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 $\mathrm{im}\,\theta = \ker \phi$.

3.5.15. 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.

3.5.16. Definition. 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.

3.5.17. 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 $w(g_1, g_2) = g_1 g_2 g_1^{-1} g_2^{-1}$.

3.5.18. 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 $i = 1, \ldots, t$.

3.5.19. Definition. 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

3.5.20. Theorem. 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 $F = F[\{x_1, \ldots, x_n\}]$.

3.5.21. Example. 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 $N = F_2'$.

3.5.22. Example. Consider the quaternion group $Q_8 = \langle a, b : a^4 = 1, a^2 = b^2, ba = a^3 b \rangle$ of order eight. We shall determine the structure of Q_8/Q_8'. Since $|a| = 4$ and $ba = a^3 b$, $\langle a \rangle \triangleleft Q_8$ 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 $a^2 = a^{-1}bab^{-1} \in Q_8'$, $\langle a^2 \rangle \subseteq Q_8' \subseteq \langle a \rangle$. 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 $a^2 Q_8' = b^2 Q_8' = Q_8'$. Therefore, $Q_8/Q_8' \cong \mathbb{Z}_2 \times \mathbb{Z}_2$.

3.5.23. Theorem. [von Dyck `/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 3.5.20, $G \cong F/N$, 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 $x_i = x_i N \mapsto h_i$ for all $i \in \{1, \ldots, n\}$. □

3.5.24. Example. 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 $|G| = 6 = |S_3|$, so $G \cong S_3$. $\square$

3.5. Exercises.

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 $x_1, x_2, \ldots, x_r$. Show that each element of F/F' is of the form $(x_1^{m_1}x_2^{m_2}\ldots x_r^{m_r})F'$. 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.

---

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.

4.1.1. Definition. 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 $u \in M$.

4.1.2. 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.

4.1.3. Examples.

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.

4.1.4. Definition. 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.

4.1.5. Definition. 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).

4.1.6. Definition. 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.

4.1.7. 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.

4.1.8. Theorem. [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.

$$\begin{tikzcd}
0 \arrow[r] & \ker f \arrow[r, hook] & M \arrow[r, "f"] \arrow[d, "\pi"'] & N \arrow[r] & \operatorname{coker} f \arrow[r] & 0 \\
 & & M/\ker f \arrow[r, "\cong", "\bar{f}"'] & \operatorname{im} f \arrow[u, "i"'] & &
\end{tikzcd}$$

4.1.9. Theorem. [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.

$$\begin{tikzcd}
0 \arrow[r] & N_1 \cap N_2 \arrow[r] \arrow[d] & N_2 \arrow[r] \arrow[d] & N_2/(N_1 \cap N_2) \arrow[r] \arrow[d, "\cong"] & 0 \\
0 \arrow[r] & N_1 \arrow[r] & N_1 + N_2 \arrow[r] & (N_1 + N_2)/N_1 \arrow[r] & 0
\end{tikzcd}$$

4.1.10. Theorem. [Third Isomorphism Theorem] If $N_2 \leq N_1 \leq N$ are R-modules, then the following diagram is commutative and has exact rows:

$$\begin{tikzcd}
0 \arrow[r] & N_1 \arrow[r] \arrow[d] & N \arrow[r] \arrow[d] & N/N_1 \arrow[r] \arrow[d, "\mathrm{id}"] & 0 \\
0 \arrow[r] & N_1/N_2 \arrow[r] & N/N_2 \arrow[r] & N/N_1 \arrow[r] & 0
\end{tikzcd}$$

That is, $N/N_1 \cong (N/N_2)/(N_1/N_2)$.

4.1.11. Theorem. 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.

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & N_1 \cap N_2 \arrow[r] \arrow[d] & N_2 \arrow[r] \arrow[d] & N_2/(N_1 \cap N_2) \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & N_1 \arrow[r] \arrow[d] & N \arrow[r] \arrow[d] & N/N_1 \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & N_1/(N_1 \cap N_2) \arrow[r] \arrow[d] & N/N_2 \arrow[r] \arrow[d] & N/(N_1 + N_2) \arrow[r] \arrow[d] & 0 \\
 & 0 & 0 & 0 &
\end{tikzcd}$$

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".

4.1.12. Theorem. [5-Lemma] Suppose the following diagram is commutative and has exact rows.

$$\begin{tikzcd}
A_1 \arrow[r, "\alpha_1"] \arrow[d, "f_1"'] & A_2 \arrow[r, "\alpha_2"] \arrow[d, "f_2"'] & A_3 \arrow[r, "\alpha_3"] \arrow[d, "f_3"'] & A_4 \arrow[r, "\alpha_4"] \arrow[d, "f_4"'] & A_5 \arrow[d, "f_5"'] \\
B_1 \arrow[r, "\beta_1"'] & B_2 \arrow[r, "\beta_2"'] & B_3 \arrow[r, "\beta_3"'] & B_4 \arrow[r, "\beta_4"'] & B_5
\end{tikzcd}$$

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. □

4.1.13. Theorem. [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 $\beta \circ s = \mathrm{id}_N$.
(iii) 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, if k is a field, then

$$0 \longrightarrow k[x]x \longrightarrow k[x] \longrightarrow k[x]/k[x]x \longrightarrow 0$$

splits 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, $a_1 \neq 0$. 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. □

4.1. Exercises.

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 $M_1 \cap M_2 = \{0\}$.
2. 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 where the scalar multiplication is defined by $(r + I)(x + IM) := rx + IM$.
3. Complete the proof of Theorem 4.1.13.
4. (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.

4.2.1. Definition. 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 (external) direct sum of $M_1, \ldots, M_k$ is denoted by $M_1 \oplus \cdots \oplus M_k$ or $\bigoplus_{i=1}^{k} M_i$.

4.2.2. Definition. 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.

4.2.3. Theorem. 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 $M_1, \ldots, M_k$.

4.2.4. Definition. 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

$$\{r_1 x_1 + \cdots + r_k x_k : r_i \in R \text{ and } x_i \in X\}.$$

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**.

4.2.5. Definition. 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 $r_1 x_1 + \cdots + r_k x_k = 0_M$, 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.

4.2.6. 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.

4.2.7. Definition. 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

$$\phi(r_1 x_1 + \cdots + r_n x_n) = r_1 y_1 + \cdots + r_n y_n.$$

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,

4.2.8. Theorem. [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 $\theta \circ i = \alpha$.

$$\begin{tikzcd} X \arrow[r, "i"] \arrow[dr, "\alpha"'] & M \arrow[d, dotted, "\theta"] \\ & N \end{tikzcd}$$

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 $\hom_R(M, N)$.

4.2.9. 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.

4.2.10. Theorem. 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.

4.2.11. Theorem. 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

$$\begin{tikzcd} M \arrow[r, "\phi"] \arrow[d, "\alpha"', "\cong"] & N \arrow[r, "\theta"] \arrow[d, "\beta"', "\cong"] & P \arrow[d, "\gamma"', "\cong"] \\ R^m \arrow[r, "{[\phi]}"] \arrow[rr, bend right, "{[\theta\phi]}"'] & R^n \arrow[r, "{[\theta]}"] & R^p \end{tikzcd}$$

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

4.2.12. Theorem. 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, $m = n$. □

4.2.13. Theorem. 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 4.2.11, we can identify M with $m \times 1$ column vectors and N with $n \times 1$ column vectors and obtain a commutative diagram

$$\begin{tikzcd} M \arrow[r, "\phi"] \arrow[d, "\cong"] & N \arrow[r, "\theta"] \arrow[d, "\cong"] & M \arrow[d, "\cong"] \\ R^m \arrow[r, "{[\phi]}"] \arrow[rr, bend right, "{[\theta][\phi]}"'] & R^n \arrow[r, "{[\theta]}"] & R^m \end{tikzcd}$$

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 4.2.12. □

4.2.14. Definition. 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 $x_1, \dots, x_m$, and hence we have shown:

4.2.15. Corollary. 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 3.5.10 for the finite basis case.

Proof of Theorem 3.5.10 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 3.5 1, G/G' is a free abelian group of rank $|A|$ and $|B|$, respectively. By Corollary 4.2.15, $|A| = |B|$. □

4.2.16. 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.

4.2.17. Example. 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 ${}_RR$.

4.2. Exercises.
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 4.2.17 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 [**?**] for this section.

4.3.1. Definition. Let R be a ring. An R-module P is called **projective** if given any diagram

$$\begin{tikzcd} & P \arrow[d, "f"] \\ M \arrow[r, two heads, "p"'] & N \end{tikzcd}$$

there exists a homomorphism $g : P \to M$ such that

$$\begin{tikzcd} & P \arrow[d, "f"] \arrow[dl, dotted, "g"'] \\ M \arrow[r, two heads, "p"'] & N \end{tikzcd}$$

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,

4.3.2. Theorem. 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 4.2.8, we have:

4.3.3. Theorem. Every free module is projective.

4.3.4. Example. $\mathbb{Q}$ 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.

$$\begin{tikzcd} & \mathbb{Q} \arrow[dl, dotted, "h"'] \arrow[d, "\mathrm{id}_{\mathbb{Q}}"] \\ F \arrow[r, two heads, "g"'] & \mathbb{Q} \end{tikzcd}$$

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

$$\sum_i k b_i x_i = k \sum_i b_i x_i = kh(k^{-1}) = h(1) = \sum_i a_i x_i,$$

so $\sum_i (a_i - kb_i) x_i = 0$. Since X is linearly independent, $a_i = kb_i$ 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.

4.3.5. Theorem. 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

$$\begin{tikzcd} & P \arrow[d, "\mathrm{id}_P"] \\ N \arrow[r, two heads, "g"'] & P \end{tikzcd}$$

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) $\Rightarrow$ (iii). Since any module is a homomorphic image of a free module (Theorem 4.2.8), 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) $\Rightarrow$ (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

$$\begin{tikzcd} & P \arrow[d, "f"] \\ M \arrow[r, two heads, "q"'] & N \end{tikzcd}$$

Combining the two diagrams, we obtain

$$\begin{tikzcd} 0 \arrow[r] & P' \arrow[r, "i"] & F \arrow[r, shift left, "p"] \arrow[dr, "fp"'] & P \arrow[l, shift left, "i'"] \arrow[r] \arrow[d, "f"] & 0 \\ & & M \arrow[r, two heads, "q"'] & N & \end{tikzcd}$$

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

$$\begin{tikzcd} & P \arrow[dl, "gi'"'] \arrow[d, "f"] \\ M \arrow[r, two heads, "q"'] & N \end{tikzcd}$$

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.

4.3.6. Corollary. 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.

4.3.7. Definition. An R-module Q is called **injective** if given any diagram of homomorphisms

$$\begin{tikzcd} 0 \arrow[r] & N \arrow[r, "i"] \arrow[d, "f"'] & M \\ & Q & \end{tikzcd}$$

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 4.3.5), 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.

4.3.8. Theorem. [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 $I = \operatorname{ann}(x + M')$ 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. $\square$

For certain "nice" rings, the concept of injectivity of modules is closely related to the simpler notion of divisibility, which we proceed to define.

4.3.9. Definition. 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

4.3.10. Theorem.
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.

4.3.11. Examples.
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 4.3.10.

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.

4.3.12. Lemma. 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 4.3.10 is the next corollary.

4.3.13. Corollary. 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.

4.3.14. Lemma. 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

$$\begin{tikzcd}
\hom_{\mathbb{Z}}(N, Q) \arrow[r, "\psi_N"] \arrow[d, "\bar{f}"'] & \hom_R(N, \hom_{\mathbb{Z}}(R, Q)) \arrow[d, "f^*"] \\
\hom_{\mathbb{Z}}(N', Q) \arrow[r, "\psi_{N'}"'] & \hom_R(N', \hom_{\mathbb{Z}}(R, Q))
\end{tikzcd}$$

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. $\square$

The foregoing lemma completes the proof of the embedding theorem.

4.3.15. Theorem. 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.

4.3.16. Theorem. 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) $\Rightarrow$ (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

$$\begin{tikzcd} 0 \arrow[r] & N' \arrow[r, "j"] \arrow[d, "f"'] & N \\ & Q & \end{tikzcd}$$

Since M is injective, we can enlarge this to a commutative diagram

$$\begin{tikzcd} 0 \arrow[r] & N' \arrow[r, "j"] \arrow[d, "f"'] & N \arrow[dl, dotted, "p'g"'] \arrow[ddl, dotted, "g"] \\ & Q \arrow[d, shift right, "i"'] & \\ & M \arrow[u, shift right, "p'"'] & \end{tikzcd}$$

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. $\square$

4.3. Exercises.

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 4.3.16 by dualizing the proof of Theorem 4.3.5.
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?

---

19. Project. (Injective hull) It is possible to prove a sharper result than Theorem 4.3.15, 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.

4.4.1. Theorem. 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

4.4.2. Theorem. 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:

4.4.3. Corollary. 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

$$\mathrm{GL}_n(R) = \{A \in M_n(R) : \det A \text{ is a unit in } R\}.$$

4.4.4. Theorem. 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. □

4.4.5. 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 $\gcd(x, y) = 1$.

4.4.6. Theorem. 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 $Ra \subset Ra_1 \subset Ra_2 \subset \dots$ 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 $\bar{P} = P_m P_{m-1} \dots P_1$ and $\bar{Q} = Q_1 \dots Q_{m-1} Q_m$. 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. □

4.4.7. Theorem. 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 $d_1 \mid d_2, \dots, d_{r-1} \mid d_r$.

Proof. We shall prove this theorem by induction on (m, n).
Case I. $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$ is a $1 \times n$ matrix. The proof of Theorem 4.4.6 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 $Q_1 \in \mathrm{GL}_2(R)$

$$\begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} Q_1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}_{n \times n} = \begin{bmatrix} \gcd(a_1, a_2) & 0 & a_3 & \dots & a_n \end{bmatrix}.$$

Then a succession of such right multiplications together with elementary transformations (or a suitable induction) show that we can obtain

$$AQ = \begin{bmatrix} \gcd(a_1, \dots, a_n) & 0 & \dots & 0 \end{bmatrix} \quad \text{for some } Q \in \mathrm{GL}_n(R)$$

as follows:

$$\begin{aligned}\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} &\sim \begin{bmatrix} \gcd(a_1,a_2) & 0 & a_3 & \dots & a_n \end{bmatrix} \\ &\sim \begin{bmatrix} \gcd(a_1,a_2) & a_3 & 0 & a_4 & \dots & a_n \end{bmatrix} \\ &\sim \begin{bmatrix} \gcd(a_1,a_2,a_3) & 0 & 0 & a_4 & \dots & a_n \end{bmatrix} \\ &\sim \dots \sim \begin{bmatrix} \gcd(a_1,\dots,a_n) & 0 & 0 & \dots & 0 \end{bmatrix}.\end{aligned}$$

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 $\begin{bmatrix} a_{11} & \dots & a_{1n} \end{bmatrix} Q_1 = \begin{bmatrix} a & 0 & \dots & 0 \end{bmatrix}$ where $a = \gcd(a_{11}, \dots, a_{1n})$ by Case I. Then

$$AQ_1 = \begin{bmatrix} a & 0 & \dots & 0 \\ X & & Y & \\ & & & \end{bmatrix}_{m \times n}$$

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 $e_1 \mid e_2, \dots, e_{s-1} \mid e_s$. Then

$$\begin{aligned}\begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} AQ_1 \begin{bmatrix} 1 & \\ & Q_2 \end{bmatrix} &= \begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} \begin{bmatrix} a & 0 \\ X & Y \end{bmatrix} \begin{bmatrix} 1 & \\ & Q_2 \end{bmatrix} = \begin{bmatrix} a & 0 \\ P_2 X & P_2 Y Q_2 \end{bmatrix} \\ &= \begin{bmatrix} a & 0 & & \cdots & 0 \\ a_2 & e_1 & & & \\ a_3 & & \ddots & & \\ \vdots & & & e_s & \\ a_m & & & & \end{bmatrix}.\end{aligned}$$

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 $b = \gcd(a, a_3, \dots a_m)$.

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 4.4.6, 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 $e = \gcd(a_2, e_1, b) = \gcd(a_2, e_1, b,$ entries of $Z) = \gcd($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 $\leq n$.

4.4.8. Theorem. 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 4.3.5. □

4.4.9. Theorem. 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 4.4.8. □

4.4.10. Theorem. 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 $M = Rx_1 + \ldots Rx_n$ is free of finite rank n and N is a submodule of M. By Theorem 4.4.9, 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

$$[\bar{N}] = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix}.$$

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 4.4.7, 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 $d_1 \mid d_2, \ldots, d_{k-1} \mid d_k$.

$[\bar{N}][Q]$ 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 $d_1 \mid d_2, \ldots, d_{k-1} \mid d_k$. □

4.4.11. Example. 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*.

4.4.12. Definition. 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.

4.4.13. Theorem. 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 $\tau(M/\tau(M)) = 0$. □

4.4.14. 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.

4.4.15. Theorem. 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. □

4.4.16. Theorem. 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 4.4.15, 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 4.4.15, 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 $i=1,\dots,k$.

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 $i=1,\dots,k$ and the theorem is proved. □

4.4.17. Theorem. 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, $d_1\mid d_2,\dots,d_{k-1}\mid d_k$ and $e_1\mid e_2,\dots,e_{m-1}\mid e_m$. Then $r=s$, $k=m$ and $R/Rd_i\cong R/Re_i$ for all $i=1,\dots,k$.

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 4.2.13. Finally, Theorem 4.4.16 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 $i=1,\dots,k$. □

4.4. Exercises.

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 $d=p_1^{n_1}\dots p_k^{n_k}$ where $p_1,\dots,p_k$ are distinct primes and $n_1,\dots,n_k>0$, 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 2.4.26 (every PID is a UFD), the fact that "every ideal of R is a principal ideal" 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.

4.5.1. Definition. 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, $s_k = s_{k+1} = \ldots$ 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..

4.5.2. Theorem. 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 $I_1 \subset I_2 \subset \ldots$ eventually stops, that is $I_k = I_{k+1} = \ldots$ 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.

4.5.3. Definition. If one of the above 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 $I \neq (f_1, \ldots, f_k)$, pick $f_{k+1} \in I \smallsetminus (f_1, \ldots, f_k)$. Then by the a.c.c. (i), the chain of ideals

$$(f_1) \subset (f_1, f_2) \subset \cdots \subset (f_1, \ldots, f_k) \subset \cdots$$

must break off at some stage, and this can only happen if $(f_1, \ldots, f_k) = I$ 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. □

4.5.4. 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.

4.5.5. Example. $\mathbb{Z}$ 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.

4.5.6. Examples. 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,
$$\begin{aligned} A_1 &= \Big\{ f(x,y) = \sum a_{ij} x^i y^j : i, j \geq 0 \text{ and } i > 0 \text{ if } j \neq 0 \Big\} \\ &= k[x, xy, xy^2, \dots, xy^n, \dots] \subset k[x, y]. \end{aligned}$$
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,
$$A_2 = \Big\{ f(x,y) = \sum a_{ij} x^i y^j : i \geq 0, \text{ and } j \geq 0 \text{ if } i = 0 \Big\} = k[x, y, x/y, x/y^2, \dots, x/y^n, \dots].$$
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.

4.5.7. Definition. 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 $M_1 \subset M_2 \subset \dots \subset M_k \subset \dots$ 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.

4.5.8. Theorem. 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 $M_1 \subset M_2 \subset \dots$ is an increasing chain of submodules of M; then identifying $\alpha(L)$ with L and taking intersection gives a chain

$$L \cap M_1 \subset L \cap M_2 \subset \dots$$

of submodules of L and applying β gives a chain $\beta(M_1) \subset \beta(M_2) \subset \dots$ 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: □

4.5.9. Lemma. 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:

4.5.10. Corollary. 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.

4.5.11. Corollary.
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.

4.5.12. Theorem. [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 $k \leq n$, the ideal $J_k \subset R$ is finitely generated, say $J_k = (a_{k,1}, \dots, a_{k,r_k})$; and by definition of J_k, for each $a_{k,j}$ with $1 \leq j \leq r_k$ there is a polynomial $f_{k,j} \in I$ of degree k having the leading coefficient $a_{k,j}$. This allows us to write down a finite set

$$S = \{f_{k,j}\}_{0 \leq k \leq n, 1 \leq j \leq r_k}$$

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. □

4.5.13. Corollary. 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 4.5.11, (4), $R[x_1, \dots, x_n]$ is Noetherian implies that so is $R[x_1, \dots, x_n]/I$. □

4.5. Exercises.

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".

4.6.1. Definition. 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.

4.6.2. Theorem. 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 $N = \sum_i Rm_i$. □

4.6.3. 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.

4.6.4. Definition. A ring whose ideals satisfy the descending chain condition (d.c.c.), i.e., whenever $I_1 \supseteq I_2 \supseteq \ldots$ 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 4.5.2, we have the following theorem.

4.6.5. Theorem. 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 4.5.8, we have:

4.6.6. Theorem. 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.

4.6.7. Lemma. 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}, $i = 0, 1, \ldots, m-1$, 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 4.5.2, M and R/M is Noetherian. Thus, R/M and M are Artinian by Exercise 4.6 (2). Hence, it follows from Theorem 4.6.6 that R is Artinian. The converse is proved in the same way. □

4.6.8. Lemma. 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.

4.6.9. Theorem. 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 $M_1, \ldots, M_n$ 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 $i \in \{1, \ldots, n\}$.
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 4.6.5, $\mathscr{S}$ 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 4.6.8, $M_i \subseteq M$ 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 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

$$\bigcap_{i=1}^{n} M_i^m \subseteq \left(\bigcap_{i=1}^{n} M_i\right)^m = J^m = \{0\}.$$

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 4.6.7. □

4.6.10. Corollary. Every finite commutative ring is a direct product of a finite number of local rings.

4.6.11. Example. 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 $i \in \{1, 2, \ldots, r\}$.

4.6. Exercises.

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 $(R \times S)^\times = R^\times \times S^\times$.

---

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 Preliminary Results

5.1.1. Definition. Let F be a field. The intersection of all subfields of F is the smallest subfield of F, called the **prime field** of F.

Recall that the characteristic of a field is 0 or a prime p.

5.1.2. Theorem. Let F be a field with the prime subfield P and 1_F denote the identity of F. Then
(1) If $\operatorname{char} F = p$, a prime, then $P = \{n \cdot 1_F : n = 0, 1, \ldots, p-1\} \cong \mathbb{Z}_p$.
(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 $P = \operatorname{im}\varphi \cong \mathbb{Z}/p\mathbb{Z} \cong \mathbb{Z}_p$.
(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. $\square$

5.1.3. Definition. A field K is said to be an **extension** of a field F if F is a subring of K.

5.1.4. Definition. 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 $[\mathbb{R} : \mathbb{Q}] = |\mathbb{R}|$).

5.1.5. Theorem. 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 $\{\alpha_i\}_{i\in I}$ be a basis of L over K. Every element of L can be written uniquely as a 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 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 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 so $[L : F] = |I \times J| = [L : K][K : F]$. □

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.

5.1.6. Theorem. Let K be a field extension of a field F and let $u \in K$. Then EITHER
(a) $[F(u) : F]$ is infinite 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 $F[u] = F(u)$. □

5.1.7. Remarks.
1. If $g(t) = g_0 + g_1t + \cdots + g_{n-1}t^{n-1} + t^n$, 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 $g(t) = g_0 + g_1t + t^2 \in \mathbb{R}[t]$. 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.

5.1.8. Definition. 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** 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 $\mathbb{Q}$; it can be shown that most of real numbers are in fact transcendental over $\mathbb{Q}$ (see Exercises). Theorem 5.1.6 yields characterizations of algebraic and transcendental elements:

5.1.9. Corollary. 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.

5.1.10. Corollary. 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 $g(t)$ divides $f(t)$; $F(u) \cong F[t]/(g(t))$; and $[F(u) : F] = \deg g(t)$.

5.1.11. Definition. 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 $\deg g(t)$.

5.1.12. Definition. 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 algebraic extension.

5.1.13. Definition. An extension field E of a field F is said to be a **simple extension** of F if $E = F(\alpha)$ for some $\alpha \in E$.

5.1.14. Theorem. 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 5.1.10. Thus,

$$[F(u) : F] \leq [E(u) : F] = [E(u) : E][E : F] < \infty.$$

Hence, u is algebraic over F. $\square$

5.1.15. Corollary. 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 $\leq mn$. In other words,

$$A = \{u \in K : u \text{ is algebraic over } F\}$$

is a subfield of K and is an algebraic extension over F.

Proof. By Corollary 5.1.10, $[F(a) : F] = m$ 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 5.1.5, $[F(a, b) : F] = [F(a)(b) : F] = [F(a)(b) : F(a)][F(a) : F] \leq mn$. 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 $\leq mn$. □

5.1. Exercises.

1. 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 $a, b, c \in \mathbb{Q}$.
2. 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.
3. 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) $\sqrt{v}$
4. Let $E = F(u)$, u transcendental and let $K \neq F$ a subfield of E/F. Show that u is algebraic over K.
5. (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 $\mathbb{R} \smallsetminus A$ is uncountable.
6. Let K be a field. A map $D : K \to K$ is called a **derivation** if $D(u + v) = D(u) + D(v)$ and $D(uv) = uD(v) + D(u)v$ for all $u, v \in K$.
   (a) Show that $D(1) = D(0) = 0$, $D(x - y) = D(x) - D(y)$ and that the set of element $x \in K$ such that $D(x) = 0$ forms a subfield M of K.
   (b) If $[K : M]$ is finite, prove that $\operatorname{char} K = p > 0$ and for every $u \in K$ there are $m \in M$ and $i \in \mathbb{N} \cup \{0\}$ such that $u^{p^i} - m = 0$.
7. 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 $[E_1E_2 : F] \leq [E_1 : F][E_2 : F]$.

5.2 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 $x - r$.

5.2.1. Definition. 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.

5.2.2. Theorem. 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. □

5.2.3. Theorem. [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 5.1.10. □

5.2.4. Corollary. 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 5.2.3, 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 $p(t)$. □

5.2.5. Definition. 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.

5.2.6. Theorem. [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 5.2.4, 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. □

5.2.7. Corollary. 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 5.2.6 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 $[E : F] \le n!$. Hence, $E = F(r_1, \dots, r_n)$ is a desired field. □

5.2.8. Examples. (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 $x^p - 1 = (x - 1)(x^{p-1} + x^{p-2} + \dots + x + 1)$ and we know that $x^{p-1} + x^{p-2} + \dots + x + 1$ is irreducible in $\mathbb{Q}[x]$. Let $E = \mathbb{Q}(z)$ where $z = x + (x^{p-1} + x^{p-2} + \dots + x + 1)$ in $\mathbb{Q}[x]/(x^{p-1} + x^{p-2} + \dots + x + 1)$. We have $1, z, \dots, z^{p-1}$ 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 $\mathbb{Z}/(p)$.

5.2.9. Theorem. [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

$$\begin{tikzcd}
F \arrow[r] \arrow[dddd, "\eta"'] & F[r] \arrow[r] & E \\
& F[x]/\hat{f}(x)F[x] \arrow[u, "i"] \arrow[d, "\hat{\eta}"] & \\
& F_1[x]/\hat{f}_1(x)F_1[x] \arrow[d, "j"] & \\
F_1 \arrow[r] & F_1[r_1] \arrow[r] & E_1.
\end{tikzcd}$$

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. □

5.2.10. Theorem. Assume $f(x)$ has no multiple factors as an element of $F[x]$. Under the hypothesis of Theorem 5.2.9, 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 5.2.9, 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 $e \le d$ 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 5.2.9 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 $\hat{\eta}_j(r) = r'_j$.

$$\begin{tikzcd}
F \arrow[r] \arrow[d, "\eta"'] & F[r] & \\
F_1 \arrow[r] & F_1[r'_j] \arrow[r, hook] & E_1
\end{tikzcd}$$

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

$$\text{the number of roots of } \hat{f}(x) = e \le d = [F[r] : F].$$

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,

$$\begin{aligned}
&\text{the number of extensions of } \eta : F \to F_1 \text{ to } \eta^* : E \to E_1 \\
&\qquad\qquad \le e[E : F[r]] \le [F[r] : F][E : F[r]] = [E : F].
\end{aligned}$$

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

$$\begin{array}{ccc} F[r] & \longrightarrow & E \\ {\scriptstyle \hat{\eta}_j}\downarrow & & \\ F_1[r'_j] & \longrightarrow & E_1 \end{array}$$

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. □

5.2.11. 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

$$f(x) = x^p - a = x^p - r^p = (x-r)^p.$$

5.2. Exercises.

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 $a \in F$.

5.3 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.

5.3.1. Theorem. 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. $\square$

We recall that (Corollary 5.1.15) 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.

5.3.2. Definition. 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 5.6.10), 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 5.2.9) 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.

5.3.3. Lemma. 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.

5.3.4. Corollary. 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.

5.3.5. Theorem. 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.

5.3.6. Definition. 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 5.2.9.

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.

5.3.7. Theorem. 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

5.3.8. Theorem. 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 $\bar{F}/F$.

5.3. Exercises.
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.]

5.4 Multiple Roots and Separability

Recall the following facts from Subsection 2.6.2 about the multiple roots.

5.4.1. Definition. 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 $f(x)$.

5.4.2. Definition. If $f(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$, we define $f'(x) \in R[x]$, the **derivative of** $f(x)$, to be

$$f'(x) = a_1 + a_2x + \cdots + na_nx^{n-1}.$$

We record the straightforward properties of the derivative of polynomials in the next lemma.

5.4.3. Lemma. 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 $n \in \mathbb{N}$.

5.4.4. Theorem. 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 $f(x)$. □

5.4.5. Definition. 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.

5.4.6. Examples.
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 $f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n$. Then $f'(x) = a_1 + 2a_2x + \cdots + nx^{n-1}$. 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:

5.4.7. Theorem. Let F be a field of characteristic zero. Then every polynomial $f(x) \in F[x]$ is separable.

5.4.8. Definition. 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.

5.4.9. Remark. Suppose F is a field (or even a commutative ring) of characteristic $p > 0$. Then the identities

$$(ab)^p = a^pb^p \quad \text{and} \quad (a + b)^p = a^p + b^p$$

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**.

5.4.10. Theorem. 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 $1 \le i \le p - 1$. Thus, if $x^p - a$ has a proper factor over F, then $r^i \in F$ for some $1 \le i \le p - 1$. But then r^i and $r^p = a \in F$, so $r \in F$ since $(i, p) = 1$. Hence, $a = r^p \in F^p$. □

5.4.11. Theorem. 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 $a \in F \smallsetminus F^p$. By Theorem 5.4.10, $x^p - a$ 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 5.4.4, 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 $F \neq F^p$. □

5.4.12. Corollary. 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 5.4.11. □

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**.

5.4.13. Theorem. 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 5.2.2, 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. □

5.4.14. Theorem. [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, $F(\beta, \gamma) = F(\alpha)$. □

5.4. Exercises.

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 $m \in \mathbb{N}$.

5.5 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.

5.5.1. Examples. (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 5.5 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 $\operatorname{Aut} F$ is in general very complicated and probably impossible to describe. Galois theory proceeds in a different direction. One takes a subgroup H of $\operatorname{Aut} F$—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

$$\begin{aligned} &1 \subseteq K \subseteq H \\ &F \supseteq F^K \supseteq F^H. \end{aligned}$$

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".

5.5.2. Definition. 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 $\mathrm{Inv}\, G$.

5.5.3. Theorem. (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 5.2.10 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.

5.5.4. Lemma. 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

5.5.5. Lemma. [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

$$u_1 x_1 + \cdots + u_{n+1} x_{n+1} = 0,$$

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

$$M = \begin{bmatrix} x_1 & x_2 & \dots & x_{n+1} \\ g_2(x_1) & g_2(x_2) & \dots & g_2(x_{n+1}) \\ \vdots & & & \\ g_n(x_1) & g_n(x_2) & \dots & g_n(x_{n+1}) \end{bmatrix}.$$

This matrix has rank $\leq n$, so there is a nonzero $(n+1) \times 1$ vector $\vec{v} = (v_1, \dots, v_{n+1})^t$ 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 $x_1, \dots, x_{n+1}$, 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,

$$M\vec{v} = \vec{0}_{(n+1)\times 1} \quad \text{where} \quad \vec{v} = (v_1, \dots, v_{r-1}, 1, 0, \dots, 0)^t.$$

Claim. *If $h \in G$ and $h(\vec{v}) = (h(v_1), \dots, h(v_{r-1}), 1, 0, \dots, 0)^t$, 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 $h \in G$

$$\begin{aligned} \vec{v} - h(\vec{v}) &= (v_1, \dots, v_{r-1}, 1, 0, \dots, 0)^t - (h(v_1), \dots, h(v_{r-1}), 1, 0, \dots, 0)^t \\ &= (\overbrace{*, \dots, *}^{r-1}, 0, \dots, 0)^t. \end{aligned}$$

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 $i = 1, \dots, r-1$, we have $h(v_i) = v_i$. Thus, all the v_i lie in $E^G = F$ and $(u_1, \dots, u_{n+1}) = (v_1, \dots, v_{r-1}, 0, \dots, 0)$ 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.

5.5.6. Definition. 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

5.5.7. Theorem. 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 5.5.4, $|G| = [E : F]$ 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

5.5.8. Corollary. 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., $E = F[r_1, \ldots, r_n] = F(r_1, \ldots, r_n)$.

5.5.9. Example. (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 5.5.7, E is a Galois extension of F with Galois group $\mathrm{Gal}(E/F) = S_n$ and dimension $|S_n| = n!$.

5.5.10. Example. 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

$$\begin{tikzcd}
 & F(x_1) \arrow[dr, dash, "2"] & \\
F = K(e_1, e_2, e_3) \arrow[ur, dash, "3"] \arrow[r, dash, "3"] \arrow[dr, dash, "3"'] & F(x_2) \arrow[r, dash, "2"] & K(x_1, x_2, x_3) = E \\
 & F(x_3) \arrow[ur, dash, "2"'] &
\end{tikzcd}$$

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.

$$\begin{tikzcd}
 & & S_3 \arrow[dll, dash, "3"'] \arrow[dl, dash, "3"'] \arrow[dr, dash, "3"] \arrow[drr, dash, "2"] & & \\
\langle(12)\rangle \arrow[drr, dash, "2"'] & \langle(13)\rangle \arrow[dr, dash, "2"] & & \langle(23)\rangle \arrow[dl, dash, "2"'] & A_3 \arrow[dll, dash, "3"] \\
 & & \{(1)\} & &
\end{tikzcd}$$

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 $\sigma \in S_3 \smallsetminus A_3$.

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

$$\begin{tikzcd}
 & & F = E^{S_3} \arrow[dll, dash, "3"'] \arrow[dl, dash, "3"'] \arrow[dr, dash, "3"] \arrow[drr, dash, "2"] & & \\
F[x_3] = E^{\langle(12)\rangle} \arrow[drr, dash, "2"'] & F[x_2] = E^{\langle(13)\rangle} \arrow[dr, dash, "2"] & & F[x_1] = E^{\langle(23)\rangle} \arrow[dl, dash, "2"'] & F[\Delta] = E^{A_3} \arrow[dll, dash, "3"] \\
 & & E = E^{\{(1)\}} & &
\end{tikzcd}$$

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.

5.5.11. Theorem. [Fundamental Theorem of Galois Theory] Let E be a finite dimensional Galois extension of a field F (i.e., the conditions of Theorem 5.5.7 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 5.5.7 to H in place of G we see that $\mathrm{Gal}(E/E^H) = H$. By Lemma 5.5.4, $|H| = |\mathrm{Gal}(E/E^H)| = [E : E^H]$. 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 5.5.7 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 5.5.7. 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). $\square$

5.5.12. Example. 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

$$\begin{tikzcd}[column sep=small]
 & \{\mathrm{id}_E\} \arrow[dl, dash] \arrow[d, dash] \arrow[dr, dash] & \\
\langle\tau_2\rangle \arrow[dr, dash] & \langle\tau_3\rangle \arrow[d, dash] & \langle\tau_4\rangle \arrow[dl, dash] \\
 & G &
\end{tikzcd}
\qquad
\begin{tikzcd}[column sep=small]
 & E \arrow[dl, dash] \arrow[d, dash] \arrow[dr, dash] & \\
\mathbb{Q}[\sqrt{2}] = E^{\langle\tau_2\rangle} \arrow[dr, dash] & \mathbb{Q}[\sqrt{3}] = E^{\langle\tau_3\rangle} \arrow[d, dash] & \mathbb{Q}[\sqrt{6}] = E^{\langle\tau_4\rangle} \arrow[dl, dash] \\
 & \mathbb{Q} &
\end{tikzcd}$$

5.5. Exercises.

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 $[E : F]$.

5.6 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.

5.6.1. Theorem. 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 $\mathrm{Gal}(E/F)$. □

5.6.2. 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.

5.6.3. Example. 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 $gs = t$.

5.6.4. 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.

5.6.5. Theorem. 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 5.2.9. 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 5.2.9, 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. □

5.6.6. 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

$$\begin{array}{ccc} E & & \{\mathrm{id}_E\} \\ \Big| & & \Big| \\ F[r] & & \mathrm{Gal}(E/F[r]) = H \\ \Big|\, n & & \Big|\, n \\ F & & \mathrm{Gal}(E/F). \end{array}$$

Thus, $F[r] = E^H$ where H is a subgroup of index n in G.

The basic Theorems 5.2.6 and 5.2.9 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.

5.6.7. Theorem. 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

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

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

5.6.8. Theorem. Every complex number has a square root.

5.6.9. Theorem. 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.

5.6.10. Theorem. [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 5.5.7.

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

$$\begin{tikzcd}
 & E \arrow[dl, dash] \arrow[dr, dash, "2^a"] & \\
\mathbb{R}[i] = \mathbb{C} \arrow[dr, dash, "2"'] & & E^P \arrow[dl, dash, "m"] \\
 & \mathbb{R} &
\end{tikzcd}$$

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 5.6.7. 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

$$\mathbb{R} \overset{2}{\text{———}} \mathbb{C} = E^H \overset{2}{\text{———}} E^K \text{———} E.$$

Thus, $[E^K : \mathbb{C}] = 2$ which contradicts Theorem 5.6.9. 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.

5.7 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 5.4.13).

5.7.1. Theorem. $k^\times$ is cyclic of order $q - 1$.

Some immediate consequences of the above theorem are as follows.

5.7.2. Corollary. The field k consists of the solutions to $x^q - x = 0$ in an algebraic closure of $\mathbb{Z}/p\mathbb{Z}$ containing k.

5.7.3. Corollary. 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 $\mathbb{Z}/p\mathbb{Z}$.

5.7.4. Corollary. For each positive divisor r of $q - 1 (= |k^\times|)$ there are exactly $\phi(r)$ elements in $k^\times$ of order r.

5.7.5. Corollary. 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 5.2.9, 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 $\mathbb{Z}/p\mathbb{Z}$.

5.7.6. Corollary. Given any positive integer d, there exists an irreducible polynomial of degree n over $\mathbb{F}_p$.

Proof. By Corollary 5.7.3, $\mathbb{F}_{p^d} = \mathbb{F}_p[\alpha]$ 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 5.7.2 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 $|\mathrm{Gal}(k_n/k)| \le [k_n : k] = n$, $\mathrm{Gal}(k_n/k) = \langle\sigma\rangle$ and so the field k_n is Galois over k. We record this in

5.7.7. Theorem. 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

5.7.8. Lemma. 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 $r \ge 2$ and τ_i are distinct. Let $y \in E$ be such that $\tau_1(y) \ne \tau_2(y)$. 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. □

5.7.9. Theorem. [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 5.7.8. 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. □

5.7.10. 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.

5.7.11. Definition. 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

5.7.12. Theorem. 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
$$a_1 = -\mathrm{Tr}_{k_n/k}(z) \quad \text{and} \quad a_n = (-1)^n\mathrm{N}_{k_n/k}(z).$$

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). □

5.7. Exercises. 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 $\mathrm{Tr}_{k_n/k}(k_n) = k\mathrm{Tr}_{k_n/k}(z)$.]

---

20. Project. (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 1.5.18 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.

5.8 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.

5.8.1. Theorem. 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 5.4.13. 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 $\mathrm{Aut}\, A = \mathrm{Aut}\, \mathbb{Z}/(n)$. □

5.8.2. Definition. 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

$$\begin{aligned}\langle x \rangle = \langle y \rangle &\iff o(x) = o(y)\\ &\iff \theta(x) = y \text{ for some } \theta \in \mathrm{Aut}\, \mathbb{Z}/(n)\\ &\iff x \text{ and } y \text{ lie in the same orbit under the action of } \mathrm{Aut}\, \mathbb{Z}/(n).\end{aligned}$$

5.8.3. Definition. 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

$$\omega = e^{2\pi i/n} = \cos(2\pi/n) + i\sin(2\pi/n) \text{ and } \omega^t = e^{2\pi i t/n} = \cos(2\pi t/n) + i\sin(2\pi t/n).$$

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 $\phi(n)$.

5.8.4. Definition. 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

$$x^n - 1 = \prod_{d|n} \Phi_d(x). \tag{5.8.1}$$

5.8.5. Remark. The formula (5.8.1) 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 (5.8.1) 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

$$\begin{aligned}
\Phi_4(x) &= \frac{x^4 - 1}{\Phi_1(x)\Phi_2(x)} = x^2 + 1 \\
\Phi_6(x) &= \frac{x^6 - 1}{\Phi_1(x)\Phi_2(x)\Phi_3(x)} = x^2 - x + 1 \\
\Phi_{12}(x) &= \frac{x^{12} - 1}{\Phi_1(x)\Phi_2(x)\Phi_3(x)\Phi_4(x)\Phi_6(x)} = x^4 - x^2 + 1.
\end{aligned}$$

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

5.8.6. Theorem. 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

$$x^n - \bar{1} = \bar{h}(x)\bar{k}(x)\bar{g}(x) \tag{5.8.2}$$

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 (5.8.2) 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 $\mathbb{Q}[x]$. $\square$

As an immediate consequence of Theorem 5.8.6, we get

5.8.7. Theorem. 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 5.8.1, $\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q})$ 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 5.8.7 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.

5.8.8. Theorem. [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.

5.8.9. Example. If $\omega = e^{2\pi i/72}$ is a primitive 72nd root of unity, then

$$\mathrm{Gal}(\mathbb{Q}[\omega]/\mathbb{Q}) \cong U_{72} \cong \mathbb{Z}/(2) \times \mathbb{Z}/(2) \times \mathbb{Z}/(6).$$

5.8.10. Definition. 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.

5.8.11. Theorem. 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. □

5.8.12. Remark. A deep theorem of Kronecker and Weber says that the converse of Theorem 5.8.11 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."*

5.8.13. Example. Let $\omega = e^{2\pi i/71}$ be a primitive 71st 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 $\mathrm{char}\, F \not| n$.

5.8.14. Theorem. 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, $\operatorname{Gal}(E/F) = \langle \alpha \rangle \cong \mathbb{Z}/(d)$. $\square$

For the sake of clarity, we reformulate Theorem 5.8.14 slightly to emphasize the case where $f(x)$ is irreducible, which is the important one.

5.8.15. Theorem. 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 $\operatorname{Gal}(E/F) \cong \mathbb{Z}/(n)$.

5.8.16. Example. 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

$$\varphi(\omega) = \omega^i \quad \text{and} \quad \varphi(r) = \omega^j r$$

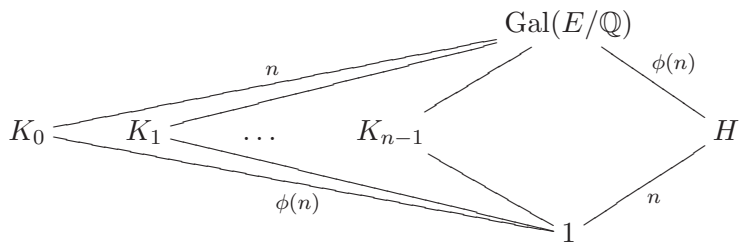
for some $1 \le i \le n-1$ such that $\gcd(i,n) = 1$ and $0 \le j \le n-1$. The choice of i and j completely determines φ and it turns out that all of the above choices do determine automorphisms of E. Thus,

$$|\mathrm{Gal}(E/\mathbb{Q})| = n \cdot \phi(n).$$

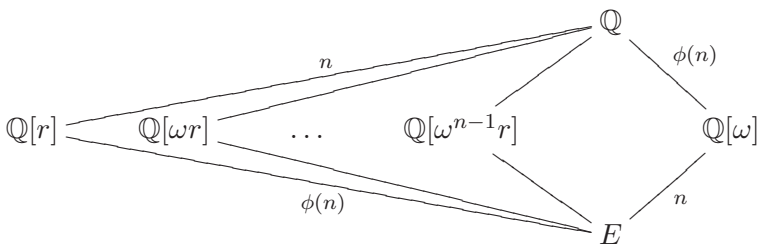
To describe $\mathrm{Gal}(E/\mathbb{Q})$ more precisely, let $\mathbb{Q}[\omega] = E^H$, and for $0 \le j \le n-1$, let $\mathbb{Q}[\omega^j r] = E^{K_j}$. Since $\mathbb{Q}[\omega]$ is Galois over $\mathbb{Q}$, H is normal in $\mathrm{Gal}(E/\mathbb{Q})$. Moreover, by Theorem 5.8.15, $H = \mathrm{Gal}(E/\mathbb{Q}[\omega]) = \langle \tau \rangle \cong \mathbb{Z}/(n)$ 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.

5.8.17. Theorem. 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.

5.8.18. Theorem. [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 1.4.14) for the multiplicative group $D \smallsetminus \{0\}$ is

$$q^n - 1 = |F \smallsetminus \{0\}| + \sum_{i=1}^{s} [D \smallsetminus \{0\} : C(d_i) \smallsetminus \{0\}] = q - 1 + \sum_{i=1}^{s} \frac{q^n - 1}{q^{m_i} - 1},$$

where $d_1, d_2, \ldots, d_s$ represent the conjugacy classes of $D \smallsetminus \{0\}$ which contains more than one element and $|C(d_i)| = q^{m_i}$ for some $m_i < n$ for all i. Because each $(q^n - 1)/(q^{m_i} - 1) = [D \smallsetminus \{0\} : C(d_i) \smallsetminus \{0\}]$ 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. □

5.8.19. Definition. 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 5.8.20, 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.

5.8.20. Theorem. [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 5.8.14 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 3.2.10 (2). □

We make two remarks without proof.

1. The converse of Theorem 5.8.20 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 5.8.20 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

5.8.21. Corollary. The general polynomial of degree $n \geq 5$ over $\mathbb{Q}$ is not solvable by radical.

5.8. Exercises. 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 $k \neq 0$.

---

21. Project. (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 5.8.20).
(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.

5.9 Normal Bases

Let E be an extension field of a field F. We have known from Lemma 5.7.8 that the automorphism in $\mathrm{Gal}(E/F)$ are E-linearly independent F-linear transformations.

5.9.1. Theorem. If E/F is a finite Galois extension with Galois group $G = \{1, \sigma_2, \dots, \sigma_n\}$. Then $\{u_1, u_2, \dots, u_n\}$ is a basis for E/F if and only if

$$\det \begin{bmatrix} u_1 & u_2 & \dots & u_n \\ \sigma_2(u_1) & \sigma_2(u_2) & \dots & \sigma_2(u_n) \\ \vdots & & & \\ \sigma_n(u_1) & \sigma_n(u_2) & \dots & \sigma_n(u_n) \end{bmatrix} \neq 0.$$

Proof. Call the above matrix M and suppose that $\det M = 0$. Since $M \in M_n(E)$, there are $\alpha_1, \alpha_2, \dots, \alpha_n \in E$, not all zero, such that

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_n \end{bmatrix} M = \vec{0}.$$

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 $u_1, u_2, \dots, u_n$. Since $1, \sigma_2, \dots, \sigma_n$ are linearly independent over K, Lemma 5.7.8 says that $\theta \neq 0$, so we have a contradiction.

Conversely, if $u_1, u_2, \dots, u_n$ are not a basis for E/F, then there are $\beta_1, \beta_2, \dots, \beta_n \in F$, 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 $M \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_n \end{bmatrix}^T = \vec{0}$. 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.

5.9.2. Theorem. Let F be an infinite field and $F \subseteq E$. If $f(x_1, \dots, x_n)$ is a nonzero polynomial in $E[x_1, \dots, x_n]$, then there exist $\alpha_1, \dots, \alpha_n \in F$ such that $f(\alpha_1, \dots, \alpha_n) \neq 0$.

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

$$f(x_1, \dots, x_{n+1}) = f_0(x_1, \dots, x_n) + f_1(x_1, \dots, x_n)x_{n+1} + \cdots + f_t(x_1, \dots, x_n)x_{n+1}^t.$$

Since $f \neq 0$, at least one of $f_0(x_1, \dots, x_n), \dots, f_t(x_1, \dots, x_n)$ is nonzero, so there are $\alpha_1, \dots, \alpha_n \in F$ such that $f(\alpha_1, \dots, \alpha_n, x_{n+1}) \neq 0$ in $E[x_{n+1}]$. By the one variable case, there is $\alpha_{n+1} \in F$ such that $f(\alpha_1, \dots, \alpha_n, \alpha_{n+1}) \neq 0$. □

5.9.3. Theorem. Let F be an infinite field and E/F Galois with Galois group $G = \mathrm{Gal}(E/F) = \{1, \sigma_2, \dots, \sigma_n\}$. Suppose that $0 \neq f(x_1, \dots, x_n) \in F[x_1, \dots, x_n]$ where $x_1, \dots, x_n$ are indeterminates over F. Then there exists $u \in E$ such that $f(u, \sigma_2(u), \dots, \sigma_n(u)) \neq 0$.

Proof. Let $\{u_1, \dots, u_n\}$ be a basis for E/F. By Theorem 5.9.1, the matrix

$$M = \begin{bmatrix} u_1 & u_2 & \dots & u_n \\ \sigma_2(u_1) & \sigma_2(u_2) & \dots & \sigma_2(u_n) \\ \vdots & & & \\ \sigma_n(u_1) & \sigma_n(u_2) & \dots & \sigma_n(u_n) \end{bmatrix} \in M_n(E)$$

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 5.9.2, 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.

5.9.4. Definition. 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.

5.9.5. Theorem. [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 5.7). Let $u \in E$. By Theorem 5.9.1, $\{\sigma_1(u), \sigma_2(u), \ldots, \sigma_n(u)\}$ 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 5.9.3, 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. □

5.9. Exercises.
1. Determine a normal basis for the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$ by using the Galois group in Example 5.5.12.
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.

5.10 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.

5.10.1. Definition. 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.

5.10.2. 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.

5.10.3. Example. 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 $f(y^2, yz, z^2) = 0$).

5.10.4. Definition. A field extension E is **algebraic over a field** 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.

5.10.5. Theorem. 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 $F(X)$.

$$\begin{array}{c} E \\ \Big| \text{ algebraic} \\ F(X) \\ \Big| \text{ purely transcendental} \\ F \end{array}$$

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 $F(X)$. □

5.10.6. 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 5.10.11). However, we shall see shortly that the number of elements in the set X is independent of particular set X.

5.10.7. Definition. Let F be a subfield of E. A maximal algebraically independent (over F) subset of E is called a **transcendence basis** for E/F.

5.10.8. Remark. By Theorem 5.10.5, 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 $E = F(B)$.

5.10.9. Theorem. Let F be a subfield of E. Then any two transcendence bases for E/F have the same cardinality.

5.10.10. Definition. 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 $F(x_1,\dots,x_n)$ of the polynomial ring $F[x_1,\dots,x_n]$ in indeterminates $x_1,\dots,x_n$. Even though these fields look quite innocent, there are difficult and unsolved problems particularly on the nature of the subfields of $F(x_1,\dots,x_n)/F$. 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

$$(a_n - ub_n)t^n + (a_{n-1} - ub_{n-1})t^{n-1} + \cdots + (a_0 - ub_0) = 0 \tag{5.10.1}$$

and $a_n - ub_n \neq 0$ since either $a_n \neq 0$ or $b_n \neq 0$ and $u \notin F$. Thus, (5.10.1) shows that t is algebraic over $F(u)$ and $[F(t) : F(u)] \leq n$. We prove the following more precise result.

5.10.11. Theorem. 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.

5.10.12. Theorem. [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 5.10.11, $[E : F(u)] = m$. 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

$$g(x) - uh(x) = q(x)f(x). \tag{5.10.2}$$

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 $1 \le i \le n$. 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 $\ge m$. Now replace u in (5.10.2) 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 $\le m$ and that of $f(x,t)$ is $\ge m$, 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 $K = F(u)$. □

5.10. Exercises.

1. Prove that there is no intermediate field K with $\mathbb{Q} \subseteq K \subsetneqq \mathbb{C}$ 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.

---

22. Project. (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 $K \subsetneqq L \subseteq K(x_1, \ldots, x_n)$ 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 $\mathbb{C} \subsetneqq L \subseteq \mathbb{C}(x_1, \ldots, x_n)$ 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 $\mathbb{C} \subsetneqq L \subsetneqq \mathbb{C}(x_1, x_2, x_3)$ such that $L/\mathbb{C}$ has transcendence degree 3 but cannot be generated by three elements.
