Notes on Abstract Algebra

Jonathan Cui

Ver. 20240617

1 Preliminaries

1.1 Notation

The sets $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ will denote the integers, the rationals, the reals, and the complex numbers respectively. An asterisk removes 0 from the set; that is, $\mathbb{Z}^* = \mathbb{Z}\backslash\{0\}$, $\mathbb{Q}^* = \mathbb{Q}\backslash\{0\}$, $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$, $\mathbb{C}^* = \mathbb{C}\backslash\{0\}$.

For $n \in \mathbb{Z}_{>0}$, we denote with $\mathbb{Z}_n$ the set of integers $\{0, \cdots, n-1\}$, the integers modulo n. Two integers $a, b \in \mathbb{Z}$ are said to be congruent modulo n if $n \mid (a-b)$; that is, n divides $(a-b)$, denoted as $a \equiv b \pmod{n}$. This is defines an equivalence relation on $\mathbb{Z}$ with n equivalence classes, corresponding to the elements of $\mathbb{Z}_n$.

We denote addition and multiplication modulo n by $+_n$ and $\times_n$ respectively. Every $a \in \mathbb{Z}_n$ has a unique additive inverse $n-a$ when $a \neq 0$ and 0 otherwise. When a and n are coprime, a has a unique multiplicative inverse according to Bézout's identity.

1.2 Binary Operations

The first structures we encounter are sets endowed with a binary operations. We formalize this idea as follows.

Definition 1.1. A binary operation over a set S is a map $*\colon S \times S \to S$. A binary operation is sometimes denoted as the pair $(S, *)$ for clarity. We commonly write in infix notation $a * b$ instead of $*(a, b)$ for $a, b \in S$.

When we consider subsets $T \subseteq S$ that have a common structure, it becomes necessary to ensure that the restriction $*|_{T\times T}$ maps to T. This motivates the definition of the closure property.

Definition 1.2. Let $(S, *)$ be a binary operation. A subset $T \subseteq S$ is said to be closed under $*$, or simply $*$-closed, if the image $T * T$ remains a subset of T. This means the restriction $*|_{T\times T}$ is also a binary operation over T, which we commonly denote $(T, *) \subseteq (S, *)$.

For instance, $(\mathbb{Z}, +)$ is a binary operation and the subset $3\mathbb{Z} \coloneqq \{3n \mid n \in \mathbb{Z}\}$ is closed under $+$.

We now define notions of commutativity and associativity.

Definition 1.3. A binary operation $(S, *)$ is said to be commutative if $a * b = b * a$ for all $a, b \in S$. A binary operation $(S, *)$ is said to be associative if $(a * b) * c = a * (b * c)$ for all $a, b, c \in S$.

An important fact is that function composition is always associative. Indeed, given $f\colon C \to D$, $g\colon B \to C$, and $h\colon A \to B$, we have

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

The two always agree, which by definition means that $\circ$ is associative. Of course, the class of all functions is not a set, but the notion of associativity still applies.

Definition 1.4. An element $e \in S$ of a binary operation $(S, *)$ is said to be an identity if $e * a = a * e = a$ for all $a \in S$.

An immediate result of any binary operation, without any additional structure, is that the identity is unique whenever it exists.

Lemma 1.5. A binary operation $(S, *)$ has at most one identity.

Proof. Suppose $e_1, e_2 \in S$ are both identities. Then, $e_2 = e_1 * e_2 = e_1$. □

Definition 1.6. Let $(S, *)$ be a binary operation with identity $e \in S$. An element $a' \in S$ is said to be an inverse of $a \in S$ if $a' * a = a * a' = e$.

An inverse may not exist: $0 \in (\mathbb{R}, \times)$ does not have an inverse even though $1 \in (\mathbb{R}, \times)$ is an identity. The uniqueness of an inverse is not guaranteed, but it is hard to show. Any associative binary operation admitting an inverse will always have at most one inverse for any element: if $a', a'' \in S$ are both inverses of $a \in S$, then $a' = a' * e = a' * a * a'' = e * a'' = a''$. However, a non-associative binary operation can be hard to construct.

2 Groups and Subgroups

2.1 Groups

A group is, in many ways, the largest family of algebraic structures of interest. It requires very few conditions, which we spell out as follows.

Definition 2.1. A group $(G, *)$ is a binary operation with the following properties:

- $*$ is associative;
- $*$ admits an identity $e \in G$;
- Every element of G has an inverse.

A first prototypical example is $(\mathbb{Z}_n, +_n)$ for $n \in \mathbb{Z}_{n>0}$. The identity is 0, and the inverse of a is simply $n - a$. Some other examples include $(\mathbb{Q}^*, \times)$, $(\mathbb{C}, +)$, and $(\{z \in \mathbb{C} : |z| = 1\}, \times)$. These are all examples where the operation **commutes**, which we now define.

Definition 2.2. A group $(G, *)$ is said to be abelian or commutative if $*$ is commutative over G; that is, $a * b = b * a$ for all $a, b \in G$.

An example of a non-abelian group is $\mathrm{GL}_n(\mathbb{R})$, the set of $n \times n$ matrices over $\mathbb{R}$ with a non-zero determinant, under the operation of matrix multiplication. The closure property follows from $|AB| = |A| \cdot |B|$ and the identity is I with the inverse being the matrix inverse. However, the non-commutativity of matrix multiplication is a fact from linear algebra which makes the group non-abelian.

Note that the closure property above is not trivial: it is baked into the definition of a binary operator that $*\colon G \times G \to G$. Checking that the range is within G is necessary in most cases. For a non-example, $(\mathbb{Z}_+, +)$ is not group since it has no identity.

In most problems, it is helpful to use the GL group as an example to visualize how things work. Integers modulo n are also an excellent choice, though it is abelian and hence not as general.

Two important results concern the uniqueness of the identity and of the inverse, which follow immediately from the definition. It is not trivial that we demonstrate these results as we use them implicitly almost all the time.

We first state the uniqueness of the group identity, which is a corollary of the uniqueness of the identity of any binary operation.

Proposition 2.3 (Uniqueness of Group Identity)**.** Suppose $(G, *)$ is a group. Then, there exists a unique identity $e \in G$ that satisfies the definition. We hereafter reserve this notation to denote the identity of a group.

Proof. By Lemma 1.5, there exists a unique $e \in G$ such that $e * a = a * e = a$ for all $a \in G$. This must then be the only choice of e in the definition, which is guaranteed to exist. □

Proposition 2.4 (Uniqueness of Group Element Inverse)**.** Suppose $(G, *)$ in a group and $a \in G$. Then, there exists a unique $a' \in G$ such that $a' * a = a * a' = e$. We hereafter reserve this notation to denote the inverse of a group element.

Proof. Suppose $a', a'' \in G$ are both inverses of a. Then,

$$a' = a' * e = a' * a * a'' = e * a'' = a''.$$

Hence a admits at most one inverse. The existence of such an a is from the definition, which then must be unique. □

A corollary is a formula for the inverse of a product, which is a more general form of $(AB)^{-1} = B^{-1}A^{-1}$ from matrices.

Corollary 2.5. Suppose $(G, *)$ is a group and $a, b \in G$. Then, $(a * b)' = b' * a'$.

Proof. Because $b' * a' * a * b = b' * e * b = e$ and $a * b * b' * a' = a * e * a' = e$, $b' * a'$ is an inverse of $a * b$. The uniqueness of the inverse then guarantees that $b' * a'$ is this inverse. □

We first state the first result on groups, the cancellation laws.

Proposition 2.6 (Left and Right Cancellation Laws)**.** Suppose $(G, *)$ is a group. Then, for all $a, b, c \in G$, $a * b = a * c$ implies $b = c$ and $b * a = c * a$ implies $b = c$.

Proof. First, suppose $a * b = a * c$. Then, $a' * a * b = a' * a * c$, which implies $b = c$. If instead $b * a = c * a$, we similarly have $b * a * a' = c * a * a'$, which also implies $b = c$. □

A corollary is that we can solve all equations like $4 * x = 2$ or $x * 5 = 2$.

Corollary 2.7. Suppose $(G, *)$ is a group. Given $a, b \in G$, the equations $a * x = b$ and $x * a = b$ have unique solutions in $x \in G$, which are $x = a' * b$ and $x = b * a'$ respectively.

Proof. We first consider $a * x = b$. Note that if $x = a' * b$, then $a * x = a * a' * b = b$ indeed. For the other direction, applying the cancellation law (Proposition 2.6) to $a * x = a * a' * b$ implies $x = a' * b$, which we have shown is a valid solution. The same argument follows for $x * a = b \iff x = b * a'$. □

We will now turn to two specific examples: $G_1 = (\mathbb{Z}_2, +_2)$ and $G_2 = (\{1, -1\}, \times)$. The tables are as follows.

$+_2$	0	1
0	0	1
1	1	0

Table 1: The table for G_1

$\times$	1	-1
1	1	-1
-1	-1	1

Table 2: The table for G_2

Note that Tables 1 and 2 are identical if we swap $0, 1$ from G_1 with $1, -1$ from G_2 in that order: the group operations are the same. What that means is that they're essentially the same group, just with elements renamed. This allows us to introduce the concept of isomorphisms, which formalizes this idea.

Definition 2.8. Two groups $(G_1, *_1)$ and $(G_2, *_2)$ are said to be isomorphic, denoted as $(G_1, *_1) \simeq (G_2, *_2)$, if there exists a bijective mapping $f : G_1 \to G_2$ such that $f(a) *_2 f(b) = f(a *_1 b)$ for all $a, b \in G$. Such an f is called a group isomorphism of G_1 and G_2.

This is an equivalence relation: it is reflexive with $f = \iota$, symmetric with $f \leftrightarrow f^{-1}$, and transitive with $f_2 \circ f_1$.

We will show that $\mathbb{Z}$ and $2\mathbb{Z} = \{2n \mid n \in \mathbb{Z}\}$ are isomorphic under $+$. Consider $f : \mathbb{Z} \to 2\mathbb{Z}, n \mapsto 2n$, which is indeed a bijection. For $a, b \in \mathbb{Z}$,

$$f(a) + f(b) = 2a + 2b = 2(a + b) = f(a + b).$$

Therefore, $(\mathbb{Z}, +) \simeq (2\mathbb{Z}, +)$.

All groups with exactly 1 element are isomorphic; so are all groups with exactly 2 and all groups with exactly 3 elements.

2.2 Abelian Examples

Our first abelian example is $(\mathbb{Z}_n, +_n)$ for $n \in \mathbb{Z}_{>0}$, the integers modulo n with modular addition. Specifically, $(\mathbb{Z}_4, +_4)$ has the following table.

$+_4$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

Table 3: The table for $(\mathbb{Z}_4, +_4)$.

We now consider its real counterpart, $(\mathbb{R}_c, +c)$, the reals modulo $c \in \mathbb{R}_{>0}$. The set $\mathbb{R}_c$ is defined as the interval $[0, c)$ with addition

$$a +_c b = \begin{cases} a + b, & \text{if } a + b < c, \\ a + b - c, & \text{if } a + b \geq c. \end{cases}$$

Clearly, the operation is closed in $\mathbb{R}_c$, it is commutative, the identity is 0 and the inverse of a is 0 when $a = 0$ and $c - a$ otherwise. Associativity can be seen manually, though we defer the proof later after the circle group is introduced.

The next example is $U = \{z \in C : |z| = 1\}$, the complex unit circle. This is made into a group with the multiplication $\times$ on $\mathbb{C}$. The closure property follows from $|wz| = |w| \cdot |z|$, the identity is 1, and the inverse is complex conjugation.

An interesting fact is that U is isomorphic to $\mathbb{R}_c$. To see this, we first show that any $\mathbb{R}_c$ is isomorphic to any $\mathbb{R}_d$ for $c, d \in \mathbb{R}_{>0}$.

Proposition 2.9. For any $c, d \in \mathbb{R}_{>0}$, $(\mathbb{R}_c, +_c) \simeq (\mathbb{R}_d, +_d)$.

Proof. Define $f : \mathbb{R}_c \to \mathbb{R}_d, x \mapsto x/c \cdot d$, which is clearly a bijection. For any $a, b \in \mathbb{R}_c$, we have

$$\begin{aligned} f(a) +_d f(b) &= \begin{cases} a/c \cdot d + b/c \cdot d, & \text{if } a/c \cdot d + b/c \cdot d < d, \\ a/c \cdot d + b/c \cdot d - d, & \text{otherwise} \end{cases} \\ &= \begin{cases} (a + b)/c \cdot d, & \text{if } a + b < c, \\ (a + b)/c \cdot d - d, & \text{otherwise} \end{cases} \\ &= \left(\begin{cases} a + b, & \text{if } a + b < c, \\ a + b - c, & \text{otherwise} \end{cases} \right) \Big/ c \cdot d \\ &= f(a +_c b). \end{aligned}$$

Hence, $(\mathbb{R}_c, +_c) \simeq (\mathbb{R}_d, +_d)$. □

We now show that $U \simeq \mathbb{R}_{2\pi}$, which then shows the intended by transitivity.

Proposition 2.10. $\mathbb{R}_{2\pi} \simeq U$.

Proof. Define $f : \mathbb{R}_{2\pi} \to U, \theta \mapsto e^{i\theta}$, which is clearly a bijection. For any $\theta, \phi \in \mathbb{R}_{2\pi}$, we have

$$f(\theta) \cdot f(\phi) = e^{i\theta} \cdot e^{i\phi} = e^{i(\theta+\phi)} = e^{i(\theta +_{2\pi} \phi)} = f(\theta +_{2\pi} \phi),$$

which shows $\mathbb{R}_{2\pi} \simeq U$. □

Our last example is $U_n = \{z \in \mathbb{C} : z^n = 1\} = \{e^{2\pi i/n} \mid n \in \mathbb{Z}_n\}$ for $n \in \mathbb{Z}_{>0}$, with the multiplication operator. This definition alone suffices to show that $U_n \simeq \mathbb{Z}_n$.

2.3 Non-Abelian Examples

Hereafter, we may omit the group operation by writing ab instead of $a * b$. The inverse of an element $a \in G$ is also denoted as a'. We also define a^n for $n \in \mathbb{Z}$ by

$$a^n := \begin{cases} \overbrace{a \cdots a}^{n \text{ copies}}, & \text{if } n > 0, \\ e, & \text{if } n = 0, \\ (a^{-1})^n = (a^n)^{-1}, & \text{if } n < 0. \end{cases}$$

We often refer to the number of elements of a group, which we give a special name.

Definition 2.11. Suppose $(G, *)$ is a group. Then, the order of G is the cardinality of G, denoted as $|G|$.

We will now turn to an elementary investigation on permutations, which are simply shuffling or renaming the letters. For example

$$1 \mapsto 2 \qquad 2 \mapsto 4 \qquad 3 \mapsto 3 \qquad 4 \mapsto 2$$

is a permutation of $\{1, 2, 3, 4\}$.

Definition 2.12. A permutation of a set A is a bijection from A to A. The collection of all permutations of A is denoted as S_A.

Because the composition of bijections is a bijection, we have a way to "chain" permutations.

Definition 2.13. Suppose $\sigma, \tau \in S_A$ are permutations of A. Then, the composition or multiplication of σ and τ, denoted as $\sigma \circ \tau$ or simply $\sigma\tau$, is the permutation $\sigma \circ \tau$.

In a standard notation, we write the example above instead with a $2 \times n$ array where the first row is the elements and the second row is where each element goes to:

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

Recall that compositions merge from the right. This gives an easy way to compute the composition of permutations. For instance,

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

We can write out the top row of the RHS, which is $1, \cdots, 5$. For each bottom entry, simply track where they'd go. $1 \mapsto 3 \mapsto 5$, so we put 5 under 1. Similarly, $2 \mapsto 5 \mapsto 1$, so we put 1 under 2. We can check that the bottom row is indeed a permuted rewriting of the top row to ensure our calculation was correct.

These results suggest a nice structure of permutations under compositions, which we show is a group.

Theorem 2.14. *Let A be a nonempty set. Then, $(S_A, \circ)$ is a group.*

Proof. First, bijections are closed under function composition, so S_A is closed under $\circ$. Further, composition is also associative. The identity map $\iota\colon A \to A, a \mapsto a$ is indeed a permutation, which is an identity for the group because $\iota\sigma = \sigma\iota = \sigma$ for all $\sigma \in S_A$. Lastly, the function inverse σ^{-1} exists for each $\sigma \in S_A$ and $\sigma^{-1}\sigma = \sigma\sigma^{-1} = \iota$, so this is also the group inverse. □

Note that the permutation group cares only about the set size: $S_A = S_B$ if $|A| = |B|$. We will therefore use canonically $\{1, \cdots, n\}$ for A.

Definition 2.15. Let $n \in \mathbb{Z}_{>0}$ and let $A = \{1, \cdots, n\}$. The symmetry group on n letters, denoted as S_n, is defined as the group S_A of permutations of A.

Note that A may be infinite but n is finite for our investigation. Elementary combinatorics implies that $|S_n| = n!$.

In general, S_n is not an abelian group! For $\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$ and $\tau = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$, we have

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

An interesting result is that any group with at most four elements is abelian. Later, we will see that $\mathbb{Z}_5$ is the only group of order 5 up to isomorphism, which is abelian. Importantly, S_3 is the smallest non-abelian group.

We now turn to a more compact way to specify a permutation, called the disjoint cycle notation. For example, consider $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 6 & 2 & 5 & 1 \end{pmatrix}$. In this notation, we begin by writing "(1,". With $\sigma(1) = 3$, we will add a 3 to arrive at "(1, 3,". Finally, $\sigma(3) = 6$, so we have "(1, 3, 6". Because $\sigma(6) = 1$ goes back to the start, we will close the parentheses to get to "(1, 3, 6)". We call this a cycle since we shift each element according to this cycle.

We haven't mentioned 2 and 4 yet, for which we need a new pair of parentheses. We opt for the smaller 2 first, so we have "(1, 3, 6)(2,". Because $\sigma(2) = 4$, we put 4 and since $\sigma(4) = 2$, we will close the parentheses for the final output $(1, 3, 6)(2, 4)$. A pigeonhole argument can be applied inductively to show that every permutation can be put into this form.

Sometimes we have a parentheses of a singleton. For example, $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$ will have $(1, 3)(2)$. Because parentheses of a singleton is equivalent to not doing anything, we may omit them. This allows us to view $(1, 3)$ as a permutation in its own right, where unspecified elements will not change.

Definition 2.16. A k-cycle in S_n, where $1 \le k \le n$, is a sequence of k elements $(a_0, \cdots, a_{k-1})$ that represents a permutation $\sigma \in S_n$ by

$$\sigma(a) = \begin{cases} a_{(i+1) \bmod k}, & \text{if } a = q_i \text{ for some } i \in \{0, \cdots, n-1\}, \\ a & \text{otherwise.} \end{cases}$$

This means we can view $(1, 3, 6)(2, 4)$ as a product of a 3-cycle and a 2-cycle. Indeed, because a number appears at most once in the *disjoint* cycle notation, at most one factor in the product will map to its input to a different number, exactly how the notation is to be interpreted.

Definition 2.17. The disjoint cycle notation of a permutation σ is the formal product of a sequence of cycles where no number is repeated such that the product evaluates to σ.

We provide a canonical (unique) way of writing the disjoint cycle notation of any permutation, where we'll always begin with the smallest unused number, defined as follows.

We'll use the shorthand $[n] \coloneqq \{1, \cdots, n\}$. Given a permutation $\sigma \in S_n$, where $n \in \mathbb{Z}_{>0}$,

- Initialize `prev` $\leftarrow \varnothing$ and `out` $\leftarrow \iota$;
- While `prev` $\neq [n]$:

- Let $i \leftarrow \min[n] \backslash \texttt{prev}$;
- Redefine $\texttt{out} \leftarrow \texttt{out} \cdot (i, \sigma(i), \cdots, \sigma^{k-1}(i))$, where $k \leftarrow \min\{k > 0 \mid \sigma^k(i) = i\}$;
- Update $\texttt{prev} \leftarrow \texttt{prev} \cup \{i, \cdots, \sigma^k(i)\}$;

- Return out;

Proposition 2.18. Every permutation $\sigma \in S_n$ can written uniquely in the canonical disjoint cycle notation.

Proof. Within the while loop, because $\texttt{prev} \subsetneq [n]$, we have $[n] \backslash \texttt{prev} \neq \emptyset$, so its minimum i exists. Subsequently, note that the mate $\mathbb{Z}_{\geq 0} \to [n]$ by $p \mapsto \sigma^p(i)$ cannot be injective as its domain has strictly larger cardinality than its codomain. Therefore, $\sigma^s(i) = \sigma^t(i)$ for some $s > t \geq 0$. This means $\sigma^{s-t}(i) = \iota(i) = i$ by the cancellation law of the group S_n (Proposition 2.6), where $s - t > 0$. Such a minimal $k = s - t$ then also exists. As $k > 0$, each execution of while loop will create a cycle of length at least 1, and thus increments $|\texttt{prev}|$ by at least $k + 1 = 1$. Therefore, the loop cannot be executed more than n times. The algorithm has been shown to terminate without error.

To show the correctness of the algorithm, observe that each iteration of the loop creates a k-cycle $(i, \cdots, \sigma^{k-1}(i))$. We claim that these cycles are disjoint; that is, no two of the cycles contain any shared elements. Suppose the contrary that $\sigma^{p_1}(i_1) = \sigma^{p_2}(i_2)$, where $0 \leq p_1 < k_1$ and $0 \leq p_2 < k_2$ correspond to the k_1- and k_2-cycles starting with i_1 and i_2 respectively. We will assume without loss of generality that $i_1 < i_2$; that is, the k_1-cycle is considered by the algorithm before the k_2-cycle. Note that $\sigma^{k_2 - p_2}(\sigma^{p_2}(i_2)) = i_2$. Then, $i_2 = \sigma^{k_2-p_2}(\sigma^{p_1}(i_1)) = \sigma^{(k_2-p_2+p_1) \bmod k_1}(i_1)$, which is a contradiction since $i_2 \notin \texttt{prev}$ by the time the k_2-cycle is considered. Now, to show that the product of cycles in out agrees with σ, observe that if $x \in [n]$, then x exists in exactly one cycle starting with some $i \in [n]$, and $x = \sigma^p(i)$. All other factors can be ignored as applying them does not modify i. Let $k > 0$ be the length of this cycle. Then, by construction, $\sigma(i) = \sigma^{p+k1}(i)$, so the output of the k-cycle—and thus of the disjoint cycles—agrees with the evaluation of σ. □

The example above provides a procedure that turns a permutation into the disjoint cycle notation. We will omit a formal argument because it is mostly technical. A note is that disjoint cycles commute but cycles in general do not.

Lastly, a benefit of the disjoint cycle notation is the ease of taking inverses: $(1, 4, 3, 5)^{-1} = (5, 3, 4, 1)$—just write the sequence inside in reverse.

Our next topic of investigation is the dihedral group, a subgroup of symmetries on the vertices of an n-gon.

Definition 2.19. Suppose $n \in \mathbb{Z}_{\geq 3}$. Define $P_n = \mathbb{Z}_n$ as vertices $e^{2\pi p/n} \in \mathbb{C}$ on the complex unit circle for each $p \in P_n$. P_n is made into a simple undirected graph where the set of edges is $E = \{\{p, (p + 1) \bmod n\} \mid p \in P_n\}$.

This allows us to fix a positioning and labeling of the polygon:

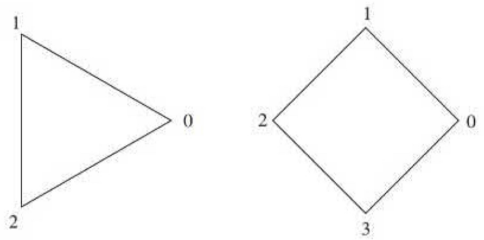

Figure 1: The regular polygons P_3 and P_4, where 0 always lies on $1 + 0i$.

We define the dihedral group D_n as the permutation of P_n that preserves the edges.

Definition 2.20. Let $n \in \mathbb{Z}_{\geq 3}$. The dihedral group D_n is defined as the collection of all permutations σ on P_n such that $\{i, j\} \in E \iff \{\sigma(i), \sigma(j)\} \in E$ for all $i, j \in P_n$, where E is the set of edges of P_n.

Just because we say that its a group doesn't mean it actually is!

Theorem 2.21. *The dihedral group D_n with composition is a group for all $n \in \mathbb{Z}_{\geq 3}$.*

Proof. We first demonstrate closure. Suppose $\sigma, \tau \in D_n$. Then, for all $i, j \in P_n$,

$$\{i, j\} \in E \iff \{\tau(i), \tau(j)\} \in E \iff \{\sigma(\tau(i)), \sigma(\tau(j))\} \in E.$$

Thus, $\sigma\tau \in D_n$.

The associativity of the group operation is from function composition. Clearly, the identity map ι is in the dihedral group and serves as the identity because $\iota\sigma = \sigma\iota = \sigma$ for all $\sigma \in D_n$. Finally, the permutation inverse is the group inverse. It is obvious that $\sigma^{-1}\sigma = \sigma\sigma^{-1} = \iota$. It remains to show that the permutation inverse is a group element. Indeed, for any $I, J \in D_n$, by setting $i = \sigma^{-1}(I)$ and $j = \sigma^{-1}(J)$, we have

$$\{\sigma^{-1}(I), \sigma^{-1}(J)\} \in E \iff \{\sigma(\sigma^{-1}(I)), \sigma(\sigma^{-1}(J))\} \in E \iff \{I, J\} \in E.$$

The proof is now complete. □

We define two basic elements in D_n.

Definition 2.22. Let $n \in \mathbb{Z}_{\geq 3}$. Define $\underline{\rho} \in D_n$ by $\rho(k) \coloneqq (k+1) \bmod n$, which rotates the n-gon by $2\pi/n$ radians.

Because $(P_n, +_n) = (\mathbb{Z}_n, +_n)$ is a group, the modular addition in ρ must be cancelable, so ρ is bijective. Further, by the definition of E, $\{i, j\} \in E \iff \{i +_n 1, j +_n 1\} \in E \iff \{\rho(i), \rho(j)\} \in E$, so $\rho \in D_n$.

The other element flips the n-gon with respect to the x-axis.

Definition 2.23. Let $n \in \mathbb{Z}_{\geq 3}$. Define $\underline{\mu} \in D_n$ by $\mu(k) \coloneqq (-k) \bmod n$, which flips vertices vertically.

The dihedral group is not abelian by just considering these two elements:

$$(\mu\rho)(0) = \mu(1) = n - 1 \qquad \neq \qquad (\rho\mu)(0) = \rho(0) = 1.$$

So from definition, we know that $D_n \subseteq S_n$. But what does it consist of? The following statement shows that D_n is simply "generated" ρ and μ.

Proposition 2.24. Let $n \in \mathbb{Z}\geq 3$. Then, $|D_n| = 2n$ and

$$D_n = \{\iota, \rho, \cdots, \rho^{n-1}, \quad \mu, \mu\rho, \cdots, \mu\rho^{n-1}\}.$$

Proof. Note that any $\sigma \in D_n$ will map P_n to P_n. First, consider $\sigma(0)$, of which there are n possibilities free in P_n. Because $\{0, 1\} \in E$, we must have $\{\sigma(0), \sigma(1)\} \in E$. By construction of the graph/polygon, this means $\sigma(1)$ is either $\sigma(0) +_n 1$ or $\sigma(0) -_n 1$. Now, there will be only one possibility for where to map any other vertices $2, \cdots, n$. So $|D_n| \leq 2n$.

Now, we show that $\{\iota, \rho, \cdots, \rho^{n-1}, \quad \mu, \mu\rho, \cdots, \mu\rho^{n-1}\} \subseteq D_n$ are distinct. There are three possibilities.

- If we choose two from the first half, then they can't be equal. Note that $\rho^i(0) = i$ and $\rho^j(0) = j$ for $i, j \in P_n$, so $i \neq j$ implies $\rho^i \neq \rho j$;
- If we choose two from the second half, then they can't be equal either. We can cancel μ because μ is a group element. They same argument as the first item ensues;
- Lastly, if we choose one from each half by taking ρ^i and $\mu\rho^j$ where $i, j \in P_n$. Note that $\rho^i(0), \cdots, \rho^i(n-1)$ are in a counterclockwise order on P_n whereas $\mu\rho^j(0), \cdots, \mu\rho^j(n-1)$ are clockwise. This means the functions $\rho^i \neq \mu\rho^j$.

Therefore, D_n has size at most $2n$, and we have found $2n$ distinct elements which must encompass all of D_n. The proof is now complete. □

Some useful facts are as follows:

- $\rho^i = \rho^{i \bmod n}$;
- $\mu^2 = \iota$;
- $\mu\rho^i = \rho^{-i}\mu$.

2.4 Subgroups

We have just seen how D_n is a subset of S_n and is a group in its own right. We should expect this inclusion to provide additional structure that allows us to quantify how they interact. We formalize this notation with the concept of subgroups.

Definition 2.25. Let $(G, *)$ be a group. A subset $H \subseteq G$ is said to be a subgroup, denoted as $H \leq G$, if $(H, *|_{H\times H})$ is a group.

This succinct definition entails a couple things, actually:

- First, H must be closed under $*$. If $a * b \notin H$ for some $a, b \in H$, then $*$ would not even be a well-defined binary operation on H to begin with;
- It must include the identity. For example, $\mathbb{Z}_{>0} \subset (\mathbb{Z}, +)$ is closed under $+$, but it isn't a (sub)group since it doesn't have an identity;
- It must include all the inverses of elements of H. Another similar example is $\mathbb{Z}_{\geq 0} \subseteq (\mathbb{Z}, +)$. It is closed under $+$ and has an identity. But no positive integer here has an additive inverse: it is not a group.

It turns out there are so many things to consider when we simply want to decide if a subset matches the definition of a subgroup. Are there easier characterizations? We answer this question with the following useful facts.

Proposition 2.26. Let $(G, *)$ be a group. A non-empty subset $H \subseteq G$ is a subgroup of G if and only if H is closed under $*$ and taking the inverse; that is, $a * b \in H$ and $a_G^{-1} \in H$ for all $a, b \in H$.

One thing to note, before we prove this, is that the identity of H must be inherited from G. As groups, H and G both have an identity unique to itself. Because $H \subseteq G$, this means that G's identity is also H's identity by restriction. Similarly, the inverse argument is the same. The notation a_G^{-1} means that we're referring to the inverse of a^{-1} in G, but we'll see that it's the same in H so we don't need this extra notation.

Proof. The $\Rightarrow$ direction is evident. For the other direction, the closure property implies that $*$ is a binary operation on H after restriction, and associativity is preserved under this restriction. Let $a, b \in G$ be arbitrary. Because $a \in H$ and $a_G^{-1} \in H$, $e_G = a * a_G^{-1} \in H$. Because $e_G * a = a * e_G = a$, e_G is an identity of H. Therefore, $(H, *|_{H\times H})$ is a group, so $H \leq G$. □

For finite groups, it turns out we can even remove the requirement that the inverse exists. This is obviously not true in general for infinite groups, like the example $\mathbb{Z}_{>0} \not\leq \mathbb{Z}$ under addition. But the finitude allows us to apply a pigeonhole argument that solve this problem. We'll basically prove the pigeonhole principle along the way to make every step clear.

Proposition 2.27. Let $(G, *)$ be a group and suppose $H \subseteq G$ is non-empty and finite. Then, $H \leq G$ if and only if H is $*$-closed.

Proof. The $\Rightarrow$ direction is obvious. For the $\Leftarrow$ direction, let $n = |H|$ and choose an arbitrary $a \in H$. Consider the map from $\mathbb{Z}_+$ to H by $n \mapsto a^n$. Because the domain has strictly larger cardinality than the range, the map cannot be injective. Thus, $a^i = a^j$ for some $0 < i < j$. This implies $a \cdot a^{j-i+1} = a^{j-i+1} \cdot a = e$, so $a^{-1} = a^{j-i+1} \in H$. By Proposition 2.26, H is a subgroup of G. □

Let's turn to some examples. As we have seen, $\mathbb{Z} \leq \mathbb{Q} \leq \mathbb{R} \leq \mathbb{C}$ as additive groups. Another interesting example is consider

the addition operation on vector spaces. Because $C^\infty(\mathbb{R})$ is a subspace of $C^0(\mathbb{R})$, we have $C^\infty(\mathbb{R}) \leq C^0(\mathbb{R})$.

What about finite structures? $\mathbb{Z}_4$ has only two proper subgroups, $\{0, 2\}$ and $\{0\}$. But the Klein-4 group, which we denote as $V = \mathbb{Z}_2 \times \mathbb{Z}_2$ with component-wise addition of $+_2$, has four: $\{(0,0),(0,1)\}$, $\{(0,0),(1,0)\}$, $\{(0,0),(1,1)\}$, and the trivial $\{(0,0)\}$.

In our previous proof, we leveraged the finitude of H by considering the set of all elements $a, a^2, a^3, \cdots$. Do they form a subgroup? Well, they're clearly closed, and we've seen that there're always inverses—provided that the group is finite. But what about infinite groups? In $(\mathbb{Z}, +)$, we can consider $2, 4, 6, 8, \cdots$. This is obviously not a group: no inverses are included!

We can extend our construction so that this is always a group, by considering the following.

Definition 2.28. Suppose G is a group and $g \in G$. For any integer $n \in \mathbb{Z}$, we define

$$g^n = \begin{cases} \overbrace{g \cdots g}^{n \text{ copies}}, & \text{if } n > 0, \\ e, & \text{if } n = 0, \\ (g^{-1})^n = (g^n)^{-1}, & \text{if } n < 0. \end{cases}$$

The cyclic subgroup of G generated by g, denoted as $\langle g \rangle$, is defined as the set

$$\{a^n \mid n \in \mathbb{Z}\}.$$

Let's show that this is a group, combining our insights from before for a formal argument. Keep in mind that while closure remains obvious, we need to explicit demonstrate the existence of inverses. As we've seen, $\langle g \rangle$ may be infinite, so Proposition 2.27 does not apply.

Proposition 2.29. Suppose G is a group and $g \in G$. Then, $\langle g \rangle \leq G$.

Proof. Let $g^m, g^n \in \langle g \rangle$. Then, $g^m \cdot g^n = g^{m+n} \in \langle g \rangle$. Now, by construction, $(g^n)^{-1} = g^{-n} \in \langle g \rangle$. □

In the context of the dihedral group, one can see that $\langle \rho \rangle = \{\iota, \rho, \cdots, \rho^{n-1}\}$ and $\langle \mu \rangle = \{\iota, \mu\}$. But note that ρ is not the only element in D_n that generates this subgroup. Consider D_5 and $\langle \rho^2 \rangle$, which turns out to be equal to ρ! At the same time, however, $\langle \rho^2 \rangle \leq \langle \rho \rangle$ in D_4.

2.5 Cyclic Groups

In this section, we'll get to delve into the interesting structure of cyclic (sub)groups more. We first extend the concept of cyclicness from subgroups to groups.

Definition 2.30. A group G is said to be a cyclic group if some element of G generates G; that is, if there exists $g \in G$ such that $\langle g \rangle = G$.

For example, $\mathbb{Z}$ is cyclic, because $\langle 1 \rangle = \{n \mid n \in \mathbb{Z}\} = \mathbb{Z}$, but $\mathbb{Q}$ is not since any $g \in G$ means every pair of numbers in $\langle g \rangle$ is separated apart by at least g, while this cannot be the case for rationals. Of course, since every cyclic group is at most countably infinite, any uncountable group like $(\mathbb{R}, +)$ or $(\mathbb{C}^*, \times)$ cannot be cyclic.

Our first result is that cyclic groups are abelian, for the simple reason that powers commute. Think matrices: this is exactly how we define matrix powers.

Proof. A cyclic group is abelian. □

Proof. Suppose $G = \langle g \rangle$ is a cyclic group, where $g \in G$. For any $g^n, g^m \in G$, we have

$$g^n \cdot g^m = g^{n+m} = g^{m+n} = g^m \cdot g^n.$$

Therefore, G is abelian. □

Nothing too fancy here. We'll now take a look at the subgroups of a cyclic group. It turns out that they're also cyclic. While intuitive, there's nothing obvious about this fact, which we prove as follows.

Proposition 2.31. Let G be a cyclic group and $H \leq G$. Then H is cyclic.

Proof. If $H = \{e\}$, then $H = \langle e \rangle$. Otherwise, there is some element $g^k \in H \backslash \{e\}$ where $k \neq 0$. Then, $|k| > 0$ and $g^{|k|} \in H$. Let n be the smallest positive integer such that $g^n \in H$, which is now guaranteed to exist. We claim that $H = \langle g^n \rangle$.

It is obvious from induction that $\langle g^n \rangle$ in H. Now let $h \in H$, where we take $h = g^m$. Let $q = \lfloor m/n \rfloor$ and $r = m \bmod n$. Then,

$$h = g^m = (g^n)^q \cdot g^r.$$

Because $(g^n)^q \in H$, we can move it to the LHS to conclude that $g^r \in H$, but $r < n$. Therefore, the only possibility is $r = 0$, and hence $h = g^m = (g^n)^q$. □

Before we move forward, we state a useful fact from discrete math.

Proposition 2.32. Suppose $r \in \mathbb{Z}_{>0}$ is coprime to $s \in \mathbb{Z}_{\geq 0}$. Then, for all $n \in \mathbb{Z}$ where $r \mid sn$, $r \mid n$.

Proof. Fix $a, b \in \mathbb{Z}$ such that $ar + bs = 1$, or $arn + bsn = n$. Obviously $r \mid arn$. Also, because $r \mid sn$, $r \mid bsn$. Then, $r \mid (arn + bsn)$, or $r \mid n$. □

Another useful fact that captures the order of a cyclic group is stated.

Proposition 2.33. Let G be a cyclic group generated by $g \in G$. Then, G is finite if and only if $g^n = e$ for some $n \in \mathbb{Z}_{>0}$. In this case, the smallest such n is $|G|$.

We can now describe exactly what cycle groups are.

Proposition 2.34. Suppose G is a cyclic group. If G is finite, then $G \simeq (\mathbb{Z}_{|G|}, +_{|G|})$; otherwise, $G \simeq (\mathbb{Z}, +)$.

Proof. Suppose $G = \langle g \rangle$ is finite with order $n \in \mathbb{Z}_{>0}$. Define $\phi \colon \mathbb{Z}_n \to G$ by $\phi(k) = g^k$, which we claim is an isomorphism. To see injectivity, let $g^i = g^j$ where $0 \leq i \leq j < n$. Then, $g^{j-i} = e$ for $0 \leq j - i < n$, which is only possible when $i = j$. For surjectivity, we must show $\{e, \cdots, g^{n-1}\} = G$. Inclusion in the $\subseteq$ direction is evident. For the $\supseteq$ direction, consider an arbitrary $m \in \mathbb{Z}$ with quotient $q = \lfloor m/n \rfloor$ and remainder $r = m \bmod n$. Then,

$$g^m = (g^n)^q \cdot g^r = e^q \cdot g^r = g^r \in \mathbb{Z}_n.$$

This establishes that ϕ is bijective. Now, for all $i, j \in \mathbb{Z}_n$

$$\phi(i) \cdot \phi(j) = g^i \cdot g^j = g^{i+j} = g^{(i+j) \bmod n} = \phi(i +_n j),$$

so $G \simeq (\mathbb{Z}_n, +_n)$.

Suppose now that $G = \langle g \rangle$ is infinite. We have similarly a map $\phi \colon \mathbb{Z} \to G$ by $\phi(k) = g^k$. To see injectivity, suppose $g^i = g^j$ where $i < j$. Then, $g^{j-i} = e$ which contradicts the infinitude of G. Surjectivity is from construction. Finally, we note that ϕ is an isomorphism by

$$\phi(i) \cdot \phi(j) = g^i \cdot g^j = g^{i+j} = \phi(i + j).$$

Therefore, $G \simeq (\mathbb{Z}, +)$. □

We've seen that generators are not unique. For example, $\langle 1\rangle = \langle 7\rangle = \mathbb{Z}_{10}$ in the additive group. It's obvious why 1 generates the whole group, and it's not too bad either for 7—because 7 and 10 are such that adding 7's repeatedly gives you a shifting that goes through all 10 possibilities:

$$0, 7, 14, 21, 28, 35, 42, 49, 56, 63, \cdots,$$

before it starts looping back to 70. In contrast, if we take 6, we see that we'll only loop through 5 of the 10 numbers in $\mathbb{Z}_10$, namely

$$0, 6, 12, 18, 24,$$

after which we have 30, which loops back to 0.

Is it because 7 is a prime? That's a good direction to go in: indeed, any $(\mathbb{Z}_n, +_n)$ can be generated by any prime less than n. But that's not all: one can verify easily that $\langle 9\rangle = \mathbb{Z}_{10}$ too.

The next result makes this idea more precise, pointing out that $\langle a\rangle = \mathbb{Z}_n$ exactly when a is coprime to n. Since we already know that cycle groups are just additive integral groups, we can rely completely on $\mathbb{Z}_n$ by considering a number s as an element g^s and vice versa. While every $0 \le s < n$ corresponds to exactly all group elements, a group element will have many corresponding s's when the range is not specified: they are congruent modulo n.

Note the specific way in which we framed the following Proposition. In the context of $G \simeq \mathbb{Z}_10$, this is saying that there's a unique subgroup of G of size 10, which can be found by $\langle g\rangle^{10/1}$. While the subgroup is unique, its generators are not: the generators of G are precisely elements g^s where $\gcd(s, n) = \gcd(1, 10) = 1$.

Proposition 2.35. Suppose G is a cyclic group of finite order n generated by $g \in G$. Then, for any positive integer $d \in \mathbb{Z}_{>0}$ such that $d \mid n$, there exists a unique subgroup $H \le G$ of order d, which is $H = \langle g^{n/d}\rangle$. Further, for any $s, t \in \mathbb{Z}_{>0}$, $\langle g^s\rangle = \langle g^t\rangle$ if and only if $\gcd(n, s) = \gcd(n, t)$.

Proof. We first show that $\langle g^s\rangle = n/\gcd(n, s)$, where $s \in \mathbb{Z}_{>0}$. Define $D \in \mathbb{Z}_{>0}$ as the smallest positive integer such that $(g^s)^D = e$. Such a D must exist since the map $\mathbb{Z}_{\ge 0} \to G$ by $D \mapsto (g^s)^D$ has a domain strictly larger than its codomain, which cannot be injective. Thus $(g^s)^{D_1} = (g^s)^{D_2}$ for some $D_1 > D_2 \ge 0$, so $(g^s)^{D_1 - D_2} = e$, where $D_1 - D_2 \ge 1$.

Because $(g^s)^D = g^{sD} = e$, we conclude that $sD \equiv 0 \pmod n$ by the isomorphism with $\mathbb{Z}_n$, so $n \mid sD$. Let $m = \gcd(n, s)$, so integers n/m and s/m are coprime and we have $n/m \mid (s/m)D$. By Proposition 2.32, we have $n/m \mid D$. Since D is smallest such number, $D = n/m$; that is, $\langle g^s\rangle = n/\gcd(n, s)$.

Now consider any positive integer d which divides n. Note that when $s = n/d$, we have $\gcd(n, s) = n/d$ and $\langle g^s\rangle = n/(n/d) = d$. The existence as promised has been demonstrated.

We conclude the proof by showing the last item. Suppose now that $\langle g^s\rangle = \langle g^t\rangle$; that is, $s \equiv t \pmod n$. By Euclid's gcd algorithm,

$$\gcd(n, s) = \gcd(n, s \bmod n) = \gcd(n, t \bmod n) = \gcd(n, t).$$

Conversely, if the equation above is given, then $s \equiv t \pmod n$, and the isomorphism between G and $\mathbb{Z}_n$ establishes $\langle g^s\rangle = \langle g^t\rangle$. The proof is complete. □

2.6 Generating Sets and Cayley Digraphs

Let's consider the dihedral group D_n ($n \ge 3$). Applying our new concept of cyclic groups, we can see that $\{\rho^0, \cdots, \rho^{n-1}\}$ is a cyclic subgroup of D_n generated by ρ, and $\{\mu^0, \mu^1\}$ is the cyclic subgroup generated by μ. But can we find more general structures that can have two generators?

The following formal definition makes it possible.

Definition 2.36. Let G be a group. We say that a subset of elements $S \subseteq G$ generates G if for every group element $g \in G$, there exists $a_1, \cdots, a_n \in S$ and $p_1, \cdots, p_n \in \mathbb{Z}$, where $n \in \mathbb{Z}_{\ge 0}$, such that the product $a_1^{p_1} \cdots a_n^{p_n} = g$.[1] The group G is said to

[1] The empty product when $n = 0$ is defined trivially as the identity $e \in G$.

be finitely generated if such a finite subset S exists. The subgroup generated by S is defined as the collection

$$\{a_1^{p_1} \cdots a_n^{p_n} \mid n \geq 0, \qquad a_1, \cdots, a_n \in S, \text{ and } \qquad p_1, \cdots, p_n \in \mathbb{Z}\}.$$

This is an intuitive definition: in D_n, we would consider all products like $\rho^{-2}\mu^3\rho^2\mu\rho$, which implies that $\{\rho, \mu\}$ generates D_n.

To establish the validity of this definition, observe that Proposition 2.26 implies the "subgroup" generated by S—a name unjustified so far—is indeed a subgroup of G. Also, S generates G precisely when the subgroup generated by S equals G. Further, when $n = 1$, we recover the definition for cyclic groups.

We provide another perspective on this definition. The phrasing "the smallest subgroup" as follows means the intersection of all subgroups under the given qualifications, which one can easily check is indeed a subgroup by Proposition 2.26.

Proposition 2.37. Suppose G is a group and $S \subseteq G$. Then, the subgroup generated by S is the smallest subgroup of G containing S.

Proof. Let K denote the subgroup generated by S and H the smallest subgroup of G containing S. Every product expression in K is from group axioms, so $K \subseteq H$. For the other direction, it is straightforward that K is a subgroup containing S, considering the expressions $a_1^{p_1}$ with $n = 1$, $a_1 \in G$, and $p_1 = 1$. Because H is the smallest such subgroup, we have $H \subseteq K$. □

This equivalent definition can be useful: the definition of the term "smallest" means we can leverage the definition of intersection and obtain useful "for all" statements for proofs.

Having established the definition, we introduce a tool to visualize the structure of how some finite elements generate a (finite) group: the Cayley digraph.

The following definition is technical; types of arrows/edges are introduced.

Definition 2.38. A digraph with $k \in \mathbb{Z}_{>0}$ types of edges is a pair (V, E) where V is a set of vertices and $E \subseteq \{1, \cdots, k\} \times V \times V$ is a set of edges.

Suppose a subset $S = \{g_1, \cdots, g_k\} \subseteq G$ where $k = |S|$ generates a finite group G. Its Cayley digraph is a digraph with G as its vertices and has $|S|$ different types of edges. We add an arrow of type $i \in \{1, \cdots, k\}$ from $a \in G$ to $b \in G$ if $ag_i = b$.

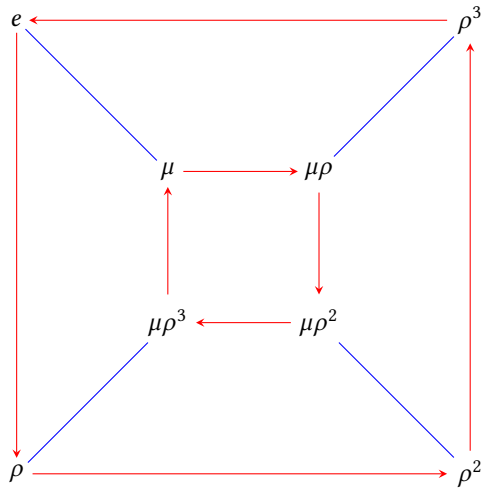

Figure 2: The Cayley digraph of D_4 generated by $\{\rho, \mu\}$, where red arrows stand for ρ and blue arrows stand for μ.

Above is the Cayley digraph for D_4. The outer loop indicates that

$$e \cdot \rho = \rho, \qquad \rho \cdot \rho = \rho^2, \qquad \rho^2 \cdot \rho = \rho^3, \qquad \rho^3 \cdot \rho = e.$$

The inner loop indicates that

$$\mu \cdot \rho = \mu\rho, \qquad \mu\rho \cdot \rho = \mu\rho^2, \qquad \mu\rho^2 \cdot \rho = \mu\rho^3, \qquad \mu\rho^3 \cdot \rho = \mu.$$

We use different colors to indicate the use of different generators along the way, and we omit arrows when there are actually arrows in both direction. So, the four blue arrow-less edges mean that

$$\rho^i \cdot \mu = \mu\rho^{4-i} \quad \text{and} \quad \mu\rho^i \cdot \mu = \rho^{4-i} \qquad \text{for } i = 1, 2, 3, 4.$$

Observe how the different rotational orientation of the inner and outer loop corresponds to μ flipping the orientation of the square.

3 Structure of Groups

3.1 Groups of Permutations

In this section, we'll dive deeper into the groups of permutations, and conclude Cayley's theorem, which states remarkably that every group is essentially a group of permutations.

We begin by extending the notion of isomorphism, removing the restriction that it be bijective. The resulting property concerns and preserves only the structure of the groups, which we call a homomorphism.

Definition 3.1. Let $(G_1, *_1)$ and $(G_2, *_2)$ be groups. A map $\phi\colon G_1 \to G_2$ is said to be homomorphism from G_1 to G_2 if

$$\phi(a *_1 b) = \phi(a) *_2 \phi(b)$$

for all $a, b \in G_1$.

A motivating example is the homomorphism ϕ from $(\mathbb{R}, +)$ to $(U, \times) \subset \mathbb{C}^*$ defined by

$$\phi(x) = \mathrm{e}^{ix}.$$

Indeed, ϕ maps the real line to the complex unit circle, and

$$\phi(a + b) = \mathrm{e}^{i(a+b)} = \mathrm{e}^{ia} \cdot \mathrm{e}^{ib} = \phi(a)\phi(b).$$

This is a homomorphism, but not an isomorphism—it is not injective. The real line wraps around the unit circle infinitely many times, each time taking up a segment of length 2π.

We'll now talk about some properties of a homomorphism that will set up our following discussion of Cayley's theorem.

Proposition 3.2. Suppose G_1 and G_2 are groups and ϕ is a homomorphism from G_1 to G_2. Then,

- If e_1 is the identity of G_1, then $\phi(e_1)$ is the identity of G_2;
- If $a \in G_1$, then $\phi(a^{-1}) = \phi(a)^{-1}$;
- If $H_1 \leq G_1$, then $\phi(H_1) \leq G_2$;
- If $H_2 \leq G_2$, then $\phi^{-1}(H_2) \leq G_1$.

Proof. This is quite straightforward. For the first item, note that $\phi(e_1) = \phi(e_1e_1) = \phi(e_1)\phi(e_1)$. Applying the cancellation law in G_2 (Proposition 2.6), we have $\phi(e_1) = e_2$, the identity of G_2. For the second item, note that $\phi(a)\phi(a^{-1}) = \phi(aa^{-1}) = \phi(e_1) = e_2$.

For the last two item, we leverage Proposition 2.26 by examining the group closure property and the inclusion of inverses. If $H_1 \leq G_1$, then for any $a, b \in G_1$, $\phi(a)\phi(b) = \phi(ab) \in \phi(H_1)$. We have shown in the second item that $\phi(a)^{-1} = \phi(a^{-1}) \in \phi(H_1)$ as well, so $\phi(H_1) \leq G_2$. Now suppose instead that $H_2 \leq G_2$. Let $a, b \in \phi^{-1}(H_2)$ be arbitrary; in other words, $\phi(a), \phi(b) \in H_2$. Then, $ab \in \phi^{-1}(\{\phi(ab)\}) = \phi^{-1}(\{\phi(a)\phi(b)\})$. The closure property of H_2 implies $\phi(a)\phi(b) \in H_2$, so $\phi^{-1}(\{\phi(a)\phi(b)\}) \subseteq \phi^{-1}(H_2)$. Therefore, $a \cdot b \in \phi^{-1}(H_2)$. For the inverse property, we have $a^{-1} \in \phi^{-1}(\{\phi(a^{-1})\}) = \phi^{-1}(\{\phi(a)^{-1}\}) \subseteq \phi^{-1}(H_2)$. The proof is now complete. □

We will now tackle Cayley's theorem. We aim to show that every group is a subgroup of S_n for some $n \geq 1$. The big idea is that we want to recover the notion of a permutation from no more than the group axioms alone. The tool we use is the cancellation law, which says that multiplication by some element is a bijection from the group to itself.

Then, we can associate a multiplication function to every element. We formalize this idea.

Definition 3.3. Suppose G is a group. To each element $a \in G$ is associated a map $\underline{\lambda_a}: G \to G$, defined as

$$\lambda_a(g) = a \cdot g \qquad \text{for all } g \in G.$$

This association which maps G to functions on G is called the left regular representation of G.

Now, the cancellation law tells us that each λ_a is an injection. Further, λ_a maps onto G since every $g \in G$ is equal to $\lambda_a(a^{-1}g)$. We can now assert that $\lambda_a \in S_G$, or that the left regular representation maps G to S_G. We're ready to prove Cayley's theorem.

Theorem 3.4 (Cayley's theorem)**.** *Every group is isomorphic to a subgroup of a group of permutations.*

Proof. Let G be a group. The left regular representation, which we call ϕ, maps G to S_G. We claim that ϕ is a homomorphism. Indeed, for all $a, b \in G$, the associated permutation $\phi(ab) = \lambda_{ab}$ is defined by

$$\lambda_{ab}(g) = ab \cdot g = \lambda_a(bg) = (\lambda_a \circ \lambda_b)(g) \qquad \text{for all } g \in G.$$

Because g is arbitrary, this implies that the functions $\phi(ab) = \phi(a) \circ \phi(b)$.

Then, $\phi(G) \leq S_G$ by Proposition 3.2, and $\phi|_G$ is a map onto this subgroup of permutations. It remains to show that ϕ is one-to-one. Indeed, if $\phi(a) = \phi(b)$ for some $a, b \in G$, then the functions $\lambda_a = \lambda_b$, or

$$ag = bg \qquad \text{for all } g \in G.$$

The cancellation law on the right implies then that $a = b$ (Proposition 2.6). Because the restriction preserves the homomorphism property by definition, $\phi|_G$ is an bijective homomorphism—an isomorphism—from G to $\phi(G) \leq S_G$. □

This is remarkable because if we can understand the structure of permutations, we should be able to apply them to any groups. We'll therefore continue our investigations of permutations.

Our first result is the formalization of the following: every reordering of the numbers $1, \cdots, n$ should be possible by interchanging pairs of numbers repeatedly.

Because we can represent any permutation in disjoint cycle notation (Proposition 2.18), we need only to study how to break any k-cycle up to a product of 2-cycles. For example, $(1, 3, 2, 4) = (1, 4)(1, 2)(1, 3)$. Note that from right to left, 1 will be mapped to 3, which occur in no other cycles on the left. Now, 3 will be mapped first to 1, and then to 2. Now, no cycles on the left of $(1, 2)$ contains 2, so we're done. Similarly, 2 is mapped to 1 and then 4. Finally, 4 is mapped to 1 in the leftmost cycle.

This construction can easily be extended to prove this result.

Proposition 3.5. Every permutation can be decomposed as a product of 2-cycles.

Proof. Due to the associativity of compositions, it suffices to show that any k-cycle in S_n is a product of 2-cycles, where $0 < k \leq n$. If $k = 1$, then $(i) \in S_n$ is equal to $(i, i +_n 1)(i +_n 1, i)$. If $k = 2$, then the cycle is already a product of 2-cycles.

Now suppose $k \geq 3$. Let $\sigma \coloneqq (i_1, \cdots, i_k)$ be an arbitrary k-cycle, where $i_\bullet \colon [k] \to [n]$ is injective. Consider the product

$$\pi \coloneqq (i_1, i_k) \cdot (i_1, i_{k-1}) \cdots (i_1, i_2).$$

Firstly, $\pi(i_1) = \Big((i_1, i_k) \cdot (i_1, i_{k-1})\Big)(i_2)$. Because none of the 2-cycles applied to i_2 contains i_2, $\pi(i_1) = i_2 = \sigma(i_2)$. Now consider $2 \leq j \leq k-1$, where i_j appears in exactly one of the factors in all but the first 2-cycles. Then similarly,

$$\pi(i_j) = \Big((i_1, i_k) \cdot (i_1, i_{k-1}) \cdots (i_1, i_2)\Big)(i_j) = \Big((i_1, i_k) \cdots (i_1, i_{j+1}) \cdot (i_1, i_j)\Big)(i_j) = \Big((i_1, i_k) \cdots (i_1, i_{j-1})\Big)(i_{j+1}) = i_{j+1},$$

because in the last equality, the permutation applied to i_{j+1} on the left hand side does not contain i_{j+1} and hence doesn't change its value. Lastly, i_k appears only in the leftmost 2-cycle, so $\pi(i_k) = i_1$. The proof is complete. □

This result is used in many cases, a notable example of which is the determinant. The rows of a square matrix are permuted, and we tack on the sign of each permutation:

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn} \sigma \cdot \left(\prod_{i \in [n]} A_{i,\sigma(i)} \right).$$

As you might recall, for a permutation written in a product of k 2-cycles, the sign is defined as $(-1)^k$. But—at least in my experience—I never knew why this definition is consistent. After all, this k is not unique:

$$(1, 2) = (1, 2)(1, 3)(1, 3) \in S_3,$$

so k could be either 1 or 3. It so happens in this case that they're both odd, but can we say this in full generality?

Group theory provides some insights. We detail a proof suggested by David M. Bloom, which relies on the concept of an *orbit*.

Definition 3.6. Suppose A is a non-empty set. Let $\sigma \in S_A$ and $a \in A$. The orbit of a is defined as the set

$$\{\sigma^k(a) \mid k \in \mathbb{Z}\} \subseteq A.$$

The orbits of σ partition A.

In the case of S_n, the orbit of a for $\sigma \in S_n$ is simply the length of the cycle containing a in the disjoint cycle notation. So for $(1, 3, 2)(4, 5) \in S_5$, the orbits of 1, 2, and 3 are all $\{1, 2, 3\}$, and the orbits of 4 and 5 are both $\{4, 5\}$. We will show this result for finite permutation groups, which are relevant for the determinant definition above.

Lemma 3.7. Suppose $\sigma \in S_n$ is a permutation of $1, \cdots, n$. If σ can be written as a product of k_1 2-cycles and as a product of k_2 2-cycles simultaneously, then k_1 and k_2 share the same parity; that is, they are either both odd or both even.

Proof. Let $\sigma \in S_n$ and $\tau = (i, j) \in S_n$ be arbitrary, where we assume $i > j$ without loss of generality. We claim first that the number of orbits of σ and $\tau\sigma$ differ by 1.

Case I. Suppose i and j are in different orbits of σ. $\tau\sigma$ will only differ from σ over elements in the two cycles containing i and j. Suppose that this restriction is equal to $(b, j, \times, \times, \times)(a, i, \times, \times)$ symbolically, where $\times$ represents other possible elements in the orbit/cycle and appears arbitrarily many times. We have

$$(i, j)(b, j, \times, \times, \times)(a, i, \times, \times) = (b, i, \times, \times, a, j, \times, \times, \times).$$

Therefore, in $\tau\sigma$, $b, i, \times, \times, a, j, \times, \times, \times$ are joined to one orbit. The orbits of other elements remain the same in σ and $\tau\sigma$. Therefore, σ has 1 more orbit than $\tau\sigma$.

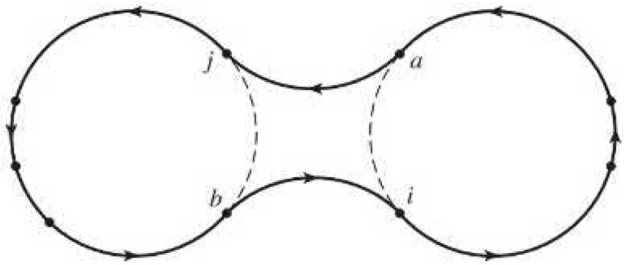

Figure 3: Case I in the proof of Lemma 3.7

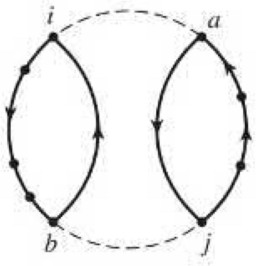

Figure 4: Case II in the proof of Lemma 3.7

Case II. Suppose i and j are in the same orbit. Suppose this orbit is $(a, i, \times, \times, b, j, \times, \times, \times)$. Then, restricted to these elements, $\tau\sigma$ is

$$(i, j)(a, i, \times, \times, \times, b, j, \times, \times) = (a, j, \times, \times)(b, i, \times, \times, \times).$$

In $\tau\sigma$, $a, i, \times, \times, \times, b, j, \times, \times$ are separated to two orbits. The orbits of other elements remain the same in σ and $\tau\sigma$. Therefore, σ has 1 fewer orbit than $\tau\sigma$.

Now, observe that the identity permutation $\iota \in S_n$ has n orbits, where each orbit is a singleton. Suppose σ is written as products

$$\tau_1 \cdots \tau_s \iota \qquad \text{and} \qquad \tau'_1 \cdots \tau'_t \iota.$$

Then, $\not{n} + i_1 + \cdots + i_s = \not{n} + j_1 + \cdots + j_t$, where $i_1, \cdots, i_s, j_1, \cdots, j_t \in \{-1, 1\}$. Because addition by 1 or -1 toggles the parity of the sum, $i_1 + \cdots + i_s$ shares the parity of s and $j_1 + \cdots + j_t$ shares the parity of t. This mean s and t share the same parity. □

This justifies the definition of the sign of permutations.

Definition 3.8. Let $\sigma \in S_n$ be a permutation written as a product of k factors of 2-cycles. Then, the sign of σ, denoted as $\operatorname{sgn}\sigma$, is defined as $(-1)^k$. We say that σ is even if k is even and that σ is odd if k is odd.

Even though there are many possible values of k for a given permutation, they all share the same parity, so the value of $(-1)^k$ remains unique. Therefore, $\operatorname{sgn}\sigma$ is well-defined.

If you add two even numbers, you get even numbers. But the sum of two odd numbers is always odd. In the same spirit, the closure property works for even permutations as well, but not odd permutations.

Proposition 3.9. Suppose $n \in \mathbb{Z}_{\geq 2}$. The collection of all even permutations $\sigma \in S_n$, denoted as $A_n < S_n$, forms a subgroup of S_n of order $n!/2$.

Proof. Consider an arbitrary 2-cycle $\tau \in S_n$, which exists because $n \geq 2$. Let $A_n \subseteq S_n$ be the collection of all even permutations and $B_n \subseteq S_n$ the set of all odd permutations. We claim that λ_τ, when restricted to A_n, maps one-to-one and onto B_n, which would imply that $|A_n| = |B_n| = |S_n|/2 = n!/2$. Indeed, for every odd permutation $\sigma' \in B_n$, $\tau^{-1}\sigma' \in A_n$ is even

and $\sigma = \lambda_\tau(\tau^{-1}\sigma) \in B_n$, so $\lambda_\tau|_{A_n}$ is onto. Whenever $\lambda_\tau(\sigma_1) = \lambda_\tau(\sigma_2)$ where $\sigma_1, \sigma_2 \in A_n$, we have $\tau\sigma_1 = \tau\sigma_2$, which implies $\sigma_1 = \sigma_2$ by the cancellation law.

We now show that A_n forms a subgroup by Proposition 2.27. Because A_n is finite, we need only to show closure. Indeed, if $\sigma_1, \sigma_2 \in A_n$, then σ_1 can be written as an even number k_1 of 2-cycles and so can σ_2 as an even number k_2 of 2-cycles. Then, $\sigma_1\sigma_2$ can be written as a product of $(k_1 + k_2)$ 2-cycles, where $k_1 + k_2$ is even. Then, Lemma 3.7 implies that $\sigma_1\sigma_2$ is again even. The proof is now complete. □

Alternatively, to gain additional insight, one can note that when restricted, sgn is a homomorphism from S_n to the group $(\{1, -1\}, \times)$ of order 2. Note that the second group is just $\mathbb{Z}_2$ (there's only one group of order 2).

Proposition 3.10. The sign of a permutation is a homomorphism from S_n to the group $(\{1, -1\}, \times)$ of order 2.

Proof. Suppose $\sigma \in S_n$ is a product of k_1 2-cycles and $\tau \in S_n$ a product of k_2 2-cycles. Then, $\sigma\tau$ is a product of $(k_1 + k_2)$ 2-cycles. There are four cases:

- If σ and τ are both odd, then $k_1 + k_2$ is even. Indeed, $\operatorname{sgn}(\sigma\tau) = (-1)^2 = 1 = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$;
- If σ and τ are both even, then $k_1 + k_2$ is even. Indeed, $\operatorname{sgn}(\sigma\tau) = 1^2 = 1 = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$;
- If σ is odd and τ is even, then $k_1 + k_2$ is odd. Indeed, $\operatorname{sgn}(\sigma\tau) = (-1) \cdot 1 = -1 = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$;
- If σ is odd and τ is even, then $k_1 + k_2$ is even. Indeed, $\operatorname{sgn}(\sigma\tau) = 1 \cdot (-1) = -1 = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$.

Therefore, $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$ for all $\sigma, \tau \in S_n$. □

This gives an alternative way of showing that A_n is a subgroup of S_n.

Corollary 3.11. $A_n \leq S_n$.

Proof. Note that A_n, the collection of all even permutations in S_n, is simply $\operatorname{sgn}^{-1}[\{1\}]$, the preimage of the trivial subgroup of the codomain. Then, A_n is a subgroup of S_n. □

3.2 Finitely Generated Abelian Groups

In this section, we'll take a look at the structure of finite abelian groups and, more generally, all finitely generated abelian groups. We begin by introducing the notion of the direct product of (finitely many) groups.

Definition 3.12. Suppose $(G_1, *_1), \cdots, (G_k, *_k)$ are groups. The direct product of $G_1, \cdots, G_k$, denoted as $G_1 \times \cdot \times G_k$, is the group of the Cartesian product equipped with elementwise group operations; that is,

$$(g_1, \cdots, g_k) \times (g_1', \cdots, g_k') \coloneqq (g_1 *_1 g_1', \cdots, g_k *_k g_k').$$

One can easily verify that the direct product of groups is indeed a group as claimed, and that the direct product of abelian groups is again abelian.

We've mentioned briefly before that there are two groups of order 4, namely $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$. The reason they're not isomorphic is that while $\mathbb{Z}_4$ is cyclic (1 has order 4), every element in $\mathbb{Z}_2 \times \mathbb{Z}_2$ has order at most 2: every element is an involution. If we look at the powers of $(1, 1)$, then we see that

$$(0,0), \quad (1,1), \quad \cancel{(0,0)}, \quad \cdots$$

But is this always the case? Let's take a look at $\mathbb{Z}_6$ vs. $\mathbb{Z}_2 \times \mathbb{Z}_3$. What would $(1, 1)$ generate? Well, just by repeatedly adding itself (note that there's no need to account for inverses additionally in finite groups), we have

$$e = (0,0), \quad (1,1), \quad (0,2), \quad (1,0), \quad (0,1), \quad (1,2), \quad \cancel{(0,0)}, \quad \cdots$$

So $(1,1)$ has order 6. This means $\mathbb{Z}_2 \times \mathbb{Z}_3$ is cyclic, and this group must be isomorphic to $\mathbb{Z}_6$. Is there a way we can quantify when these structures are isomorphic?

Proposition 3.13. The group $\mathbb{Z}_m \times \mathbb{Z}_n$ is isomorphic to $\mathbb{Z}_{mn}$ if and only if m and n are coprime.

Proof. First, suppose m and n are coprime. We claim that $(1,1)$ generates $\mathbb{Z}_m \times \mathbb{Z}_n$, for which it suffices to show that the map $\{0, \cdots, mn-1\} \to \mathbb{Z}_m \times \mathbb{Z}_n$ that takes i to $(1,1)^i = (i \bmod m, i \bmod n)$ is injective. Indeed, if $(i \bmod m, i \bmod n) = (j \bmod m, j \bmod n)$, we have

$$i \equiv j \pmod{m} \qquad \text{and} \qquad i \equiv j \pmod{n}.$$

Then, m and n both divide $(i-j)$. Because m and n are coprime, this implies that $mn \mid (i-j)$. Given that $i, j \in \{0, \cdots, mn-1\}$, this implies that $i = j$ necessarily.

For the other direction, assume that the gcd of m and n is $d > 1$. For any $r \in \mathbb{Z}_m$ and $s \in \mathbb{Z}_n$, we have $(r,s)^{mn/d} = (r^{mn/d}, r^{nm/d}) = (e_m, e_n) = e$, so the order of any element is at most $mn/d < mn$. Therefore, $\mathbb{Z}_m \times \mathbb{Z}_n$ cannot be cyclic. □

Of course, this result can easily be extended to more groups.

Corollary 3.14. The product group $\prod_{i=1}^n \mathbb{Z}_{m_i}$ is isomorphic to $\mathbb{Z}_{m_1 \cdots m_n}$ if and only if any two of the numbers $m_1, \cdots, m_n$ are relatively prime; that is, $\gcd(m_1, \cdots, m_n) = 1$.

Now, we take the same idea one step further to investigate the order of a particular element in such a product group. For example, what's the order of $(2,4)$ in $\mathbb{Z}_4 \times \mathbb{Z}_5$? We know that $|\langle 2 \rangle| = 2$ in $\mathbb{Z}_4$ and $|\langle 4 \rangle| = 5$ in $\mathbb{Z}_5$. Starting from $(0_{\mathbb{Z}_4}, 0_{\mathbb{Z}_5})$, repeatedly adding $(2,4)$ means that every 2 additions take the first entry to 0, and every 5 additions take the second entry to 0. Then, for both to be zero (so that the element in the product group is the identity), we need to add $\operatorname{lcm}(2,5) = 10$ times.

Proposition 3.15. Let $(a_1, \cdots, a_n) \in \prod_{i=1}^n G_i$, where each a_i is of finite order r_i. Then, the order of $(a_1, \cdots, a_n)$ is $\operatorname{lcm}(r_1, \cdots, r_n)$.

Proof. Suppose the groups $G_1, \cdots, G_n$ have identities $e_1, \cdots, e_n$ respectively. For each $i = 1, \cdots, n$, because a_i has order r_i, we have $a_i^k = e_i$ if and only if $r_i \mid k$, where $k \in \mathbb{Z}_{>0}$. Now, the order of $(a_1, \cdots, a_n)$ is the smallest positive integer k such that $(a_1, \cdots, a_n)^k = (a_1^k, \cdots, a_n^k) = (e_1, \cdots, e_n)$. This requires that $r_1 \mid k$, $\cdots$, $r_n \mid k$, and hence $k = \operatorname{lcm}(r_1, \cdots, r_n)$ by definition. □

These results are quite powerful: they tell us that $\mathbb{Z}_2 \times \mathbb{Z}_2 \ncong \mathbb{Z}_4$, but $\mathbb{Z}_2 \times \mathbb{Z}_3 \simeq \mathbb{Z}_6$. These result give us something strikingly similar to the prime factorization of integers:

Theorem 3.16 (Primary Factor Version of the Fundamental Theorem of Finitely Generated Abelian Groups)**.** *Every finitely generated abelian group G is isomorphic to a direct product of cyclic groups of the form*

$$\mathbb{Z}_{{p_1}^{r_1}} \times \mathbb{Z}_{{p_2}^{r_2}} \times \cdots \times \mathbb{Z}_{{p_n}^{r_n}} \qquad \times \qquad \mathbb{Z} \times \cdots \times \mathbb{Z},$$

where $p_1, \cdots, p_n$ are primes, not necessarily distinct, and $r_1, \cdots, r_n$ are positive integers. Further, the direct product is unique up to the rearrangement of the factors.

The proof is omitted here as this result is straightforward intuitively and logically; many technical details are necessary for a complete proof that do not add to our insights. A non-trivial result, however, is that $\prod_{i=1}^n \mathbb{Z} \ncong \prod_{i=1}^m \mathbb{Z}$ if $n \neq m$. A possible argument can be found here.

Theorem 3.17 (Invariant Factor Version of the Fundamental Theorem of Finitely Generated Abelian Groups)**.** *Every finitely generated abelian group G is isomorphic to a direct product of cyclic groups of the form*

$$\mathbb{Z}_{d_1} \times \mathbb{Z}_{d_n} \qquad \times \qquad \mathbb{Z} \times \cdots \times \mathbb{Z},$$

where $d_1, \cdots, d_n \in \mathbb{Z}_{\geq 2}$, not necessarily distinct, satisfy $d_1 \mid \cdots \mid d_n$. Further, the direct product is unique up to the rearrangement of the factors.

The idea is as follows. Suppose we have a primary factorization $\mathbb{Z}_2 \times \mathbb{Z}_4 \times \mathbb{Z}_9 \times \mathbb{Z}_5 \times \mathbb{Z}_{25} \times \mathbb{Z}_{125}$, where factors are ordered increasingly first by the base then by the power. We rearrange the factors as follows: the factors of the same base are written in one line flushed right.

$$\begin{array}{ccc} & 2 & 4 \\ & & 9 \\ 5 & 25 & 125 \end{array}$$

We then multiply each column to get the invariant factors: $\mathbb{Z}_5 \times \mathbb{Z}_{2\cdot 9\cdot 25} \times \mathbb{Z}_{4\cdot 125}$. The validity of this construction is obvious. Meanwhile, any different invariant factorization of the same order must have displaced some divisor of a factor, but $\mathbb{Z}_{p^{i+j}} \neq \mathbb{Z}_{p^i} \times \mathbb{Z}_{p^j}$.

3.3 Cosets and Lagrange's Theorem

In all our examples so far, the order of a subgroup always divides that of a finite group. This is a straightforward result for finite abelian groups: one can construct such a group explicitly leveraging the primary factor decomposition (Theorem 3.16). Is this true in greater generality for non-abelian groups as well?

The answer is given by Lagrange's theorem, which states precisely this divisibility relationship between a group and its subgroup. Given a subgroup $H \leq G$, the theorem is proven by demonstrating a partition of G, all sharing the cardinality of H. The uniform sizes, then, mean that the order of H divides that of G. The subsets that partition G are called the left cosets of H, constructed as equivalence classes of an equivalence relation $\sim_H$ on G.

Proposition 3.18. Let H be a subgroup of G. The relation $\sim_H$ on G defined through

$$a \sim_H b \qquad \Longleftrightarrow \qquad a^{-1}b \in H$$

is an equivalence relation.

Proof. We check each property of an equivalence relation. Suppose $a, b, c \in G$ are arbitrary.

Reflexivity. Because $a^{-1}a = e \in H$, $\sim_H$ is reflexive.

Symmetry. Suppose $a^{-1}b \in H$. Because $b^{-1}aa^{-1}b = a^{-1}bb^{-1}a = e$, $b^{-1}a$ is the inverse of $a^{-1}b$ and must be in H as well. Therefore, $\sim_H$ is symmetric.

Transitivity. Now suppose $a^{-1}b, b^{-1}c \in H$. Then, $a^{-1}c = (a^{-1}b) * (b^{-1}c) \in H$, so $\sim_H$ is transitive. □

The equivalence classes of an equivalence relation on G always partition G. For an arbitrary $a \in G \geq H$, the subset containing a is precisely the collection of all $x \in G$ such that $a \sim_H x$, or $a^{-1}x \in H$. This means that any $x \in G$ is in this subset precisely when $a^{-1}x = h$ for some $h \in H$, or $x = ah$. We therefore denote the equivalence class of a as aH.

Definition 3.19. Suppose G is a group. Let $a \in G$ and $H \leq G$. The subset $\{ah \mid h \in H\} \subseteq G$, denoted as $\underline{a * H}$ or simply $\underline{aH}$, is called the left coset of H containing a. Similarly, the subset $\{ha \mid h \in H\}$, denoted as $\underline{H * a}$ or $\underline{Ha}$, is defined as the right coset of H containing a.

Let's consider the additive group $(\mathbb{Z}, +)$ of integers and the subgroup $3\mathbb{Z}$ of integers divisible by 3. What are all the left cosets of $3\mathbb{Z}$?

First, the identity always works. The left coset of 0 is $0 + 3\mathbb{Z} = 3\mathbb{Z}$. Now, any integer not in $3\mathbb{Z}$ would be in a separate left coset, like 1. Then, $1 + 3\mathbb{Z} = \{\cdots, -5, -2, 1, 4, \cdots, \}$ is another left coset. Because 2 is not in either, we have yet another left coset $2 + 3\mathbb{Z} = \{\cdots, -4, -1, 2, 5, \cdots\}$. Now, these three left cosets partition $3\mathbb{Z}$, so we have found all left cosets of $3\mathbb{Z}$.

Given $H \leq G$ and $a \in G$, there is a natural map from H to aH by $h \mapsto ah$, which is onto by definition and one-to-one from the group cancellation law. The arbitrary choice for a means that every left coset of a subgroup has the same cardinality as that subgroup, even when G is infinite. We're now ready to prove Lagrange's theorem.

Theorem 3.20 (Lagrange)**.** *Let H be a subgroup of a finite group G. Then, the order of H divides the order of G.*

Proof. Suppose there are k cosets of H, $a_1H, \cdots, a_kH \subseteq G$ for some $a_1, \cdots, a_k \in G$. Consider the maps $f_i \colon H \to a_iH$ by $h \mapsto a_ih$ for $i \in \{1, \cdots, k\}$. Each map is surjective by definition and injective following the group cancellation law (Proposition 2.6). Hence, all cosets of H, which partition G, have the same cardinality of H. Then, $k \cdot |H| = |G|$, where $k \in \mathbb{Z}$. Thus, $|H|$ divides $|G|$. □

This theorem allows us to generalize many results that we can obtain from the prime factorization for finite abelian groups. An important example is the fact that every group of prime order is cyclic. This is obvious for abelian groups, since the prime factorization is unique. But it is not at all straightforward why this might be true for non-abelian groups (well, to be clearer, no non-abelian group have prime order).

Corollary 3.21. Every group of prime order is cyclic.

Proof. Suppose G is a group with prime order $p = |G| \geq 2$. Choose an arbitrary group element $g \in G \backslash \{e\}$. Because the only element of order 1 is the identity, the order of g is at least 2. Since it must divide p, it must equal p as p is a prime. Therefore, g generates G, and G is cyclic. □

In a similar perspective, we note the following:

Corollary 3.22. The order of an element of a finite group divides the order of the group.

Lastly, we introduce a new terminology.

Definition 3.23. The index of a subgroup H of a group G, denoted as $(G : H)$, is the cardinality of the set of equivalence classes of $\sim_H$.

Of course, when G is finite, the index is simply $(G : H) = |G|/|H|$. But this also works for infinite groups. For instance, $(\mathbb{Z} \colon 2\mathbb{Z}) = 2$.

4 Homomorphisms and Factor Groups

4.1 Factor Groups

Recall the concept of a quotient space: given a vector space V, say $\mathbb{R}^2$, and a subspace U like the diagonal line $y = x$, the quotient space V/U is the space of all lines parallel to the given $y = x$. In order to make these lines into another vector space, we define the operation of $v + U$, representing the line whose displaced by the vector v. So we define adding two lines $v + U$ and $w + U$ as $(v + U) + (w + U) = (v + w) + U$ and scaling a line $v + U$ as $c \cdot (v + U) = (c \cdot v) + U$.

Since addition forms an abelian group in every vector space, it's natural to think about the coarser group structure of addition on the quotient space. Generalizing this concept, we have factor groups, constructed as the left cosets of a subgroup. The notation aH is already the "$v + U$"—but there's a catch.

While vector spaces are abelian groups under addition, a general group may not be abelian. As a result, the left and right cosets of a subgroup may be different. While this doesn't bother us when defining cosets as simply subsets that partition a group (either left or right could work in the proof of Lagrange's theorem), this distinction means not all subgroups $H \leq G$ will lead to a well-defined group structure on G/H.

Definition 4.1. A subgroup H of a group G is said to be normal, denoted as $H \trianglelefteq G$, if $gH = Hg$ for all $g \in G$; that is, the left and right cosets of H coincide. In this case, we refer to either the left or the right cosets of H as simply cosets of H, and

define the factor group of G by H as the group of cosets of H endowed with the operation

$$(aH) * (bH) := (ab)H$$

for all $a, b \in G$.

It is non-trivial to show that this definition works. While every coset is captured by the notation aH for some $a \in G$, the choice for a is not unique. In fact, pick any other a from this left coset (which we'll call the *representative* of the left coset), and aH would give you the same left coset always. So, applying the definition for different representatives of the same groups, does their product in G/H coincide? In discrete math, during modular multiplication, we had to show explicitly that

$$a_1 \cdot b_1 \equiv a_2 \cdot b_2 \pmod{n}$$

if $a_1 \equiv a_2 \pmod{n}$ and $b_1 \equiv b_2 \pmod{n}$. It is precisely the same situation here.

We answer this question with the following lemma, which shows that the normality of a subgroup H of G is precisely the condition for a well-defined group structure on G/H.

Lemma 4.2. Let H be a subgroup of G. Then, G/H is made into a well-defined group with the definition above if and only if H is normal.

Before we begin the proof, note that the coset operations have a certain kind of associativity following that of the group:

$$(ab)H = a(bH) \qquad \text{and} \qquad (aH)b = a(Hb).$$

This is true for all left and right cosets, whether or not H is normal. Therefore, we remove the parentheses in these cases as they are not necessary.

Proof. For the $\Rightarrow$ direction, suppose G/H is a well-defined group. Let $g \in G$ and $x \in gH$. Then, $x \sim_H g$, so $xH = gH$ and $(xH) * (g^{-1}H) = xg^{-1}H$ must coincide with $(gH) * (g^{-1}H) = gg^{-1}H = eH$. Thus, $xg^{-1} \sim_H e$, so $xg^{-1} \in H$. In other words, $x = hg$ for some $h \in H$, so $x \in Hg$. The same argument applies to show that $x \in gH$ whenever $x \in Hg$, so $gH = Hg$ for all $g \in G$.

For the $\Leftarrow$ direction, suppose now that H is normal. Let $h, h' \in H$ be arbitrary, so that ahH and $bh'H$ range through all possible representations of the respective left cosets. Then,

$$(ahH) * (bh'H) = ahbh'H = ahbH = ahHb = aHb = abH.$$

Finally, we check that G/H is indeed a group when H is normal. Associativity follows from that of the group operation on G, the identity is H, and the inverse of a group element gH is $g^{-1}H$. □

Note that it is not necessary to check if the value of $g^{-1}H$ coincides for different g's representing the same gH. We need only to show that *an* inverse exists.

Here, we could have defined instead an operation on subsets $A, B \subseteq G$ by

$$A * B := \{ab \mid a \in A, b \in B\},$$

which yields $(aH)(bH) = (ab)H$ when H is normal.

The following are some results familiar to MATH 436 students who've taken linear algebra. Given a subspace (normal subgroup) U of V, there is a natural linear map (homomorphism) π from V to V/U by $\pi(v) = v + U$ whose null space (kernel) is U.

Shifting a line by a vector in a plane is linear in the shift, and a shift doesn't change the location of the line iff it's along the line. We phrase this result in terms of group theory.

Proposition 4.3. To each a normal subgroup H of G is associated a natural homomorphism $\gamma\colon G \to G/H$ by $g \mapsto gH$, whose kernel is H.

Proof. Suppose $a, b \in G$ are arbitrary. Then,

$$\gamma(ab) = abH = (aH) * (bH) = \gamma(a) * \gamma(b).$$

The kernel of γ is the collection of $g \in G$ such that $gH = H$, or $g \sim_H e$. Then, $\ker \gamma = H$. □

Now, every linear map (homomorphism) $T\colon V \to W$ induces an isomorphism $T/\text{null}\, T\colon V/\text{null}\, T \to W$ by $v + \text{null}\, T \mapsto T(v)$. To translate this to group theory terms, we need to make sure that the subgroup $\text{null}\, T$ is normal. We establish the following:

Proposition 4.4. The kernel of any group homomorphism is a normal subgroup.

Proof. Let $\phi\colon G \to G'$ be a group homomorphism with kernel $H = \ker \phi$. Note that for $g \in G$ and $h \in H$, $\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g^{-1}) = \phi(g)\phi(g^{-1}) = \phi(gg^{-1}) = \phi(e) = e$, so $ghg^{-1} \in H$. This means $gh = h'g$ for some $h' \in H$, so $gh \in Hg$. Because the choice of h was arbitrary, this implies that $gH = Hg$, so H is normal. □

The following result is known as the first isomorphism theorem, or the fundamental homomorphism theorem.

Theorem 4.5 (Fundamental Homomorphism Theorem for Groups)**.** *Let $\phi\colon G \to G'$ be a group homomorphism with kernel $H = \text{null}\, T$. Then, $\mu\colon G/H \to G$ by $gH \mapsto \phi(g)$ is a well-defined isomorphism from G/H to $\phi[G]$. In other words, $\phi = \mu \circ \gamma$.*

Proof. We first show that μ is well-defined. Let $g \in G$ and $h \in H$, so ghH ranges through all possible representations of gH. Then, $\mu(ghH) = \phi(gh) = \phi(g)\phi(h) = \phi(g) = \mu(gH)$.

To show that μ is a homomorphism, note that for all $g, g' \in G$, $\mu(gH)\mu(g'H) = \phi(g)\phi(g') = \phi(gg') = \mu(gg'H) = \mu((gH) * (g'H))$.

Now, μ is surjective by construction. It is injective because $\mu(gH) = \mu(g'H)$ implies $\phi(g) = \phi(g')$. Now, $\phi[gH] = \{\phi(gh) \mid \phi(h) = e\} = \{\phi(g)\}$, so $gH = g'H$. □

Finally, we provide some equivalent conditions of subgroup normality.

Proposition 4.6. Let H be a subgroup of G. The following are equivalent.

- $H \trianglelefteq G$;
- $gH = Hg$ for all $g \in G$;
- $ghg^{-1} \in H$ for all $g \in G$ and $h \in H$;
- $gHg^{-1} = H$ for all $g \in G$;
- There is a group G' and a homomorphism $\phi\colon G \to G'$ such that $\ker \phi = H$.

The proof is omitted as all ideas have been used in previous arguments.

Finally, we introduce a construct called the commutator subgroup, which arises from the inspection of the equation $ab = ba$ that assert a and b commute. Equivalently, $aba^{-1}b^{-1} = e$, which is the basis of the following definition.

Definition 4.7. Suppose G is a group. Denote with C the commutator subgroup of G, defined as the subgroup generated by all elements of the form $aba^{-1}b^{-1}$ where $a, b \in G$. Formally,

$$C = \langle \{aba^{-1}b^{-1} \mid a, b \in G\} \rangle.$$

As we'll see in a moment, C is normal. By taking the quotient of G/C, we are effectively identifying all $aba^{-1}b^{-1}$ with the identity. Indeed, the quotient group will be commutative. In fact, we have the following stronger result concerning the commutativity of a factor group.

Theorem 4.8. *Suppose G is a group. Then, $C \trianglelefteq G$. Further, for all $N \trianglelefteq G$, the factor group G/N is abelian if and only if $C \leq N$.*

Proof. First, observe that the inverse of an element $aba^{-1}b^{-1}$ from the generating set is $bab^{-1}a^{-1}$, also from the generating set. Thus, C contains precisely all finite products of elements from the generating set; that is,

$$C = \left\{ \prod_{i=1}^{k} (a_i b_i a_i^{-1} b_i^{-1}) \quad \middle| \quad k \geq 0 \text{ and } a_1, \cdots, a_k, b_1, \cdots, b_k \in G \right\}.$$

For the normality of C, we must show that $g(a_1 b_1 a_1^{-1} b_1^{-1}) \cdots (a_k b_k a_k^{-1} b_k^{-1}) g^{-1} \in C$. Note that between every pair of adjacent parentheses, we may insert $e = g^{-1}g$ without affecting the value of the expression, which separates the left hand side element as a product of elements $g a_i b_i a_i^{-1} b_i^{-1} g^{-1}$. By the closure of C, it suffices to show that all such elements are included in C. Indeed, given $g, a, b \in G$, we have

$$gaba^{-1}b^{-1}g^{-1} = gaba^{-1} \cdot (g^{-1}b^{-1}bg) \cdot b^{-1}g^{-1} = (ga \cdot b \cdot a^{-1}g^{-1} \cdot b^{-1}) \cdot (b \cdot g \cdot b^{-1} \cdot g^{-1}) \in C.$$

For the further claim, observe the following chain of equivalences:

$$\begin{aligned}
G/N \text{ is abelian} &\iff (aN) \cdot (bN) = (bN) \cdot (aN) \text{ for all } a, b \in G \\
&\iff ab \sim_N ba \text{ for all } a, b \in G \\
&\iff aba^{-1}b^{-1} \sim_N e \text{ for all } a, b \in G \\
&\iff aba^{-1}b^{-1} \in N \text{ for all } a, b \in G \\
&\iff C \leq N.
\end{aligned}$$

The proof is complete. □

4.2 Simple Groups

We expand upon our previous discussion on factor groups and cosets to gain some additional insights on the structure of groups.

Definition 4.9. A group is said to be simple if it is non-trivial and has no proper non-trivial normal subgroups.

Why does this concept matter, and what does it tell us? In general, we do not have results as strong as the primary factor decomposition for finitely generated groups: if G is a group of order n and k divides n, we need not have any subgroup H of order k. Rather, we ask if G is normal; for if not then we can keep quotienting the non-trivial normal subgroups until it is simple. Note that for the finite abelian cases, the only normal subgroups are $\mathbb{Z}_p$ (up to isomorphisms of groups) due to the factorization.

Definition 4.10. A subgroup M of G is said to be maximally normal if M is normal and no proper normal subgroup of G contains M properly.

Intuitively, quotienting by such a maximally normal subgroup M means we can't quotient any more afterwards. In other words:

Proposition 4.11. A normal subgroup M of G is maximally normal if and only if G/M is simple.

Proof. For the $\Rightarrow$ direction, suppose $M \trianglelefteq G$ is maximally normal and consider the surjective homomorphism $\gamma \colon G \to G/M$ by $g \mapsto gM$. Suppose for the sake of contradiction that G/M is not simple, where $\tilde{H}$ is a non-trivial normal proper subgroup of G/M. Consider $H = \phi^{-1}[\tilde{H}]$ with $|H| \geq |\tilde{H}| > 0$. Because $e_{G/M} = M \in \tilde{H}$ and any $m \in M$ has $mM = M$, $M \subseteq H$. Further,

since $\tilde{H}$ is non-trivial, we may assert about the pre-image that H properly contains M; similarly, since $\tilde{H} \subsetneq G/M$, $H \subsetneq G$. Therefore, M cannot be a maximally normal subgroup.

Conversely, let G/M be simple but assume for contradiction that M is not maximally normal, with a proper normal subgroup N of G strictly containing M. Then the normal subgroup $\phi[N]$ strictly contains the trivial subgroup $\phi[M] = \{e\}$ of G/M. Note that because N is a proper subgroup of G, $\phi[N]$ is also a proper subgroup of G/M. Then, G/M is not simple, a contradiction. □

4.3 Group Action on a Set

The concept of viewing a group element as something that acts on another object is not new. For instance, elements of S_n are functions that act on (are applied to) the numbers $1, \cdots, n$, and elements of D_n act on regular n-gons. In most abstract ways, the heart of Cayley's theorem is that a fixed group element can be thought of as an action on the group via left-multiplication by this element.

We will generalize this concept with the concept of group actions. As is the case for groups, the entire structure of groups actions are encapsulated by a "binary operation" but on different sets.

Definition 4.12. Suppose X is a set and G a group. An group action, or simply an action, of G on X is a binary map $*\colon G \times X \to X$ such that

- $e * x = x$ for all $x \in X$;
- $(ab) * x = a * (b * x)$ for all $a, b \in G$ and $x \in X$.

X is said to be a G-set where the context of the associated group action $*$ is understood. G is said to act on X by $*$. When no ambiguity can arise, one simply writes gx for $g * x$ where $g \in G$ and $x \in X$.

The two additional stipulations ensure that the structure of the group action is compatible with that of the group. We interpret the $*$ operation as right-associative, so that $a * b * x$ means $a * (b * x)$ naturally.

The following are some examples:

- S_n acts on $\{1, \cdots, n\}$ by $\sigma * i = \sigma(i)$;
- D_n acts on regular n-gons by rotation;
- $\mathrm{GL}_n(\mathbb{R})$ acts on $\mathbb{R}^n$ by $A * v = Av$;
- $\mathrm{GL}_n(\mathbb{R})$ acts on $M_n(\mathbb{R})$ by $A * B = ABA^{-1}$;

We state some natural definitions following the concept of group actions.

Proposition 4.13. Suppose G acts on X by $*$. Then, to each $g \in G$ is associated a canonical bijection, or a canonical permutation, $\lambda_g\colon X \to X$ via $x \mapsto gx$.

Proof. Let $y \in X$ and consider $x = g^{-1} * y$. Then, $g * x = g * g^{-1} * y = (gg^{-1}) * y = e * y = y$, so λ_g is surjective. Now, suppose $\lambda_g(x) = \lambda_g(y)$, or $g * x = g * y$, where $x, y \in X$. Then, applying g^{-1} to both sides yields $g^{-1} * g * x = g^{-1} * g * y$, or $e * x = e * y$. Thus, $x = y$ and λ_g is injective. □

Note how the above proof hinges upon the two stipulations from the definition of group actions, so that we may obtain the left cancellation law of group actions.

A bijection from X to X is a permutation on X, so this association is a map from G to S_X.

Lemma 4.14. Suppose G acts on X by $*$. Then, the map $\phi\colon G \to S_X$ by $g \mapsto \lambda_g$ is a canonical homomorphism.

Proof. Let $a, b \in G$ and $x \in X$. Then, $\phi(a) \circ \phi(b) = \lambda_a \circ \lambda_b = x \mapsto a * b * x = \lambda_{ab} = \phi(ab)$. □

Given that G acts on X by $*$, we can naturally ask, what group elements from G would leave the entirety of X unchanged? Such elements form a normal subgroup of G.

Definition 4.15. Suppose G acts on X by $*$. The group action $*$ is said to be faithful if the only element $g \in G$ such that $g * x = x$ for all $x \in X$ is $g = e$; or equivalently, the canonical homomorphism $\phi : G \to S_X$ is injective.

Proposition 4.16. Suppose G acts on X by $*$. Then, $N \coloneqq \{g \in G \mid g * x = x \text{ for all } x \in X\}$ forms a normal subgroup of G. Further, the action of G/N on X by $gN * x = g * x$ is faithful.

Proof. We first show $N \leq G$. Suppose $a, b \in N$ and $x \in X$. Then, $(ab) * x = a * b * x = a * x = x$, so $ab \in N$. Similarly, applying the left cancellation law (Proposition 4.13), $a * x = x$ implies $a^{-1} * x = x$, so $a^{-1} \in N$.

Suppose further that $g \in G$, so $g^{-1} * x \in X$. Note that $gag^{-1} * x = g * a * (g^{-1} * x) = g * (g^{-1} * x) = (gg^{-1}) * x = e * x = x$, so $gag^{-1} \in N$ and normality follows from Proposition 4.6.

We now show the action is well-defined. Suppose $gN = hN$; that is, the action of $g^{-1}h \in N$ is the identity on X. Then, $g * x = g * (g^{-1}h * x) = h * x$ indeed. Finally, suppose $gN * x = hN * x$ for all $x \in X$, where $g, h \in G$ and $x \in X$; that is, $g * x = h * x$. Then, $g^{-1}h * x = x$, so $g^{-1}h \in N$. In other words, $g \sim_N h$, or $gN = hN$. □

In retrospect, one can note that N is the kernel of the canonical homomorphism to S_X and is hence normal by Proposition 4.4.

Restricting our attention to group elements that leave a particular $x \in X$ fixed, we obtain another subgroup which we call the stabilizer. Note that with this restriction, such a subgroup need not be normal anymore.

Definition 4.17. Suppose G acts on X by $*$ and let $x \in X$. The stabilizer of x is defined as

$$G_x = \{g \in G \mid g * x = x\}.$$

In other words, the stabilizer of a set element x is the collection of all group elements that do not change, or stabilize, x.

Proposition 4.18. Suppose G acts on X by $*$ and let $x \in X$. Then, the stabilizer of x is a subgroup of G.

Proof. Suppose $g, h \in G_x$; that is, $g * x = h * x = x$. Then, $gh * x = g * h * x = g * x = x$, so $gh \in G_x$. Further, acting on both sides of $g * x = x$ by g^{-1} yields $g^{-1} * x = x$, so $g^{-1} \in G_x$. Thus, $G_x \leq G$ by Proposition 2.26. □

Next, we introduce the concept of orbits, which we have already seen, e.g., in the proof that $\operatorname{sgn} : S_n \to \{-1, +1\}$ is well-defined. One can note that the abstract definition below is strikingly similar to that of right cosets, which rightfully derive from the algebraic compatibility of the group action and the group operation.

Definition 4.19. Suppose G acts on X by $*$ and let $x \in X$. The orbit of x, denoted as $G * x$ or simply Gx, is defined as

$$G * x \coloneqq \{g * x \mid g \in G\}.$$

Proposition 4.20. Suppose G acts on X by $*$. Then, the orbits of elements of X partition X as the equivalence classes of the equivalence relation $\sim_*$ on X where $x \sim_* y$ if and only if $x \in G * y$.

Proof. Because $x = e * x$, we have $x \sim_* x$ for all $x \in X$, so $\sim_*$ is reflexive.

For all $x, y \in X$, if $x \sim_* y$, then $x \in G * y$. Fix $g \in G$ such that $x = g * y$. Then, applying g^{-1} to both sides yields $g^{-1} * x = g^{-1} * g * y = e * y = y$, so $y \sim_* x$. Therefore, $\sim_*$ is symmetric.

Lastly, suppose $x, y, z \in X$ are such that $x \sim_* y$ and $y \sim_* z$. Fix $g, h \in G$ such that $x = g * y$ and $y = h * z$. Then, $x = g * y = g * h * z = (gh) * z$, so $x \sim_* z$. Hence, $\sim_*$ is transitive. □

Equipped with more machinery, we can explore a few more abstract examples to gain greater insight into the concept.

- G acts faithfully on G by $a * b = ab$, which has a unique orbit. The canonical *injective* homomorphism from G to S_G gives rise to Cayley's theorem;
- $H \leq G$ acts on G by $h * g = hg$. The orbit of $g \in G$ is $H * g = Hg$, the right coset of H containing g. Lagrange's theorem is recovered (how?) by noting the stabilizer of any element is the trivial subgroup;
- G acts on the left cosets of $H \leq G$ by $a * bH = abH$.
- G acts on $H \trianglelefteq G$ by $g * h = ghg^{-1}$. The action has a unique orbit. Because H is normal, $g * h \in H$. Compatibility can be seen from $ehe^{-1} = h$ and $(ab)h(ab)^{-1} = abhb^{-1}a^{-1} = a * b * h$.

Note how this definition is strikingly similar to that of a right coset. But there's a catch: while all right cosets share the same cardinality (Theorem 3.20), this is not true for even the simplest orbits! This is because the group action $*$ has the left but not right cancellation law. If the group G is really big, then $g_1 * x = g_2 * x$ can be possible with $g_1 \neq g_2$; that is, there could be so many elements in G that the left cancellation law of $*$ can be guaranteed not to hold.

So can we quantify the cardinality of a orbit easily? It turns out that the orbit of a set element is deeply tied to its stabilizer.

Theorem 4.21. *Suppose G acts on X by $*$ and let $x \in X$. Then, $|G * x| = (G : G_x)$.*

Proof. We establish a bijective correspondence between elements of $G * x$ and left cosets of G_x. Suppose $\pi : G * x \to \{gG_x \mid g \in G\}$ is defined through $\pi(g * x) = gG_x$. This map is well-defined: if $g * x = h * x$ for $g, h \in G$, then $h^{-1}g \in G_x$. Hence, $g \sim_{G_x} h$ by definition, and $gG_x = hG_x$.

Surjectivity is by construction. To show injectivity, suppose $gG_x = hG_x$, which by the same logic above implies $g * x = h * x$. □

We now introduce the concept of the *center* of a group, the collection of the group elements that commute with everything.

Definition 4.22. Suppose G is a group. The center of G, denoted as $Z(G)$ or simply Z, is defined as the collection of group elements which commute with all group elements; that is, $Z(G) \coloneqq \{z \in Z \mid zg = gz \text{ for all } g \in G\}$.

Such a natural construction is, not surprisingly, a subgroup. More remarkably, it is normal. This can be seen by leveraging conjugation again. To justify the machinery, we make a quick definition of the automorphism group.

Definition 4.23. Suppose G is a group. An <u>automorphism</u> on G is a isomorphism from G to G. The <u>automorphism group</u> of G, denoted as $\mathrm{Aut}(G)$, is defined as the collection of all automorphisms on G endowed with the operation of composition.

Because isomorphisms are closed under composition and inverses, the automorphism group is clearly a group.

Proposition 4.24. Suppose G is a group. Then, the center of G is a normal subgroup of G.

Proof. Consider the homomorphism $\phi : G \to \mathrm{Aut}(G)$ by $\phi(g)(a) = gag^{-1}$. One can verify that $\phi(gh)(a) = ghah^{-1}g^{-1} = g\phi(h)(a)g^{-1} = \phi(g)(\phi(h)(a)) = (\phi(g) \circ \phi(h))(a)$. The kernel of this homomorphism is

$$\ker \phi = \{z \in G \mid zgz^{-1} = g \text{ for all } g \in G\} = \{z \in G \mid zg = gz \text{ for all } g \in G\} = Z(G).$$

Hence, by Proposition 4.4, the center $Z(G)$ is normal. □

Turning to a particular construction of a group action, we gain some further insights into the fundamental structure of groups.

Proposition 4.25. Suppose p is prime and G is a group of order p^n for some integer n. Then, the center $Z(G)$ has order divisible by p.

Proof. Consider the action of G on G by $g * h = ghg^{-1}$. The set elements of G whose orbits have cardinality 1 are precisely those $z \in G$ such that $g * z = gzg^{-1} = z$ for all $g \in G$; in other words, $gz = zg$ for all $g \in G$, so such elements are precisely those of the center $Z(G)$.

Suppose all other r orbits, which have cardinality at least 2, are represented as $G*g_1, \cdots, G*g_r$. Because the orbits partition G, we have $p^n = |Z(G)| + \sum_{i=1}^{r} |G * g_i|$. Rearranging the terms and applying Proposition 4.21, we have

$$|Z(G)| = p^n - \sum_{i=1}^{n} (G : G_{g_i}).$$

Because p divides p^n and each $(G : G_{g_i})$, we conclude that p must divide $|Z(G)|$ as well. □

Corollary 4.26. Suppose p is prime and G is a group of order p^n for some integer n. Then, the center $Z(G)$ is not trivial.

The result above is quite strong, and we illustrate its power with the following statement.

Proposition 4.27. Suppose p is prime and G is a group of order p^2. Then, G is abelian.

Proof. By Corollary 4.26 and Lagrange's theorem (Theorem 3.20), $|Z|$ is either p or p^2. In the latter case, G is immediately abelian.

Suppose now that Z has order p. The same argument as above implies the center must be cyclic, generated by some element $z \in G$. Because $Z \trianglelefteq G$, we may consider the factor group G/Z which also has order p and is cyclic. Suppose $G/Z = \langle gZ \rangle$ where $g \in G$.

Now, for arbitrary $a, b \in G$, suppose $a \in g^i Z$ and $b \in g^j Z$ where $a = g^i z_1$ and $b = g^j z_2$. Then,

$$ab = g^i(z_1 g^j)z_2 = g^{i+j}(z_1 z_2) = g^j(g_i z_2)z_1 = g^j z_2 g_i z_2 = ba,$$

so G is also abelian in this case. □

Lastly, we present Cauchy's theorem, which serve to strengthen Lagrange's theorem and provide additional insights as to classifying group in a similar spirit to the primary factor decomposition of finite abelian groups. This theorem is a weaker converse of Lagrange's theorem: the order of every subgroup divides the group containing it, but there need not exist a subgroup whose order is an arbitrary divisor of the group order. This is, however, the case when the divisor is prime.

Theorem 4.28 (Cauchy)**.** *Suppose p is prime and G is a group whose order is divisible by p. Then, G has an element, and hence a subgroup, of order p.*

Proof. Let $\mathbb{Z}_p$ act on

$$\tilde{G} \coloneqq \{(g_0, \cdots, g_{p-1}) \mid g_0, \cdots, g_{p-1} \in G \text{ and } g_0 \cdots g_{p-1} = e\}$$

by $n * (g_0, \cdots, g_{p-1}) = (g_n, \cdots, g_{p-1}, g_0, \cdots, g_{n-1})$, shifting the tuple as a cycle by n to the right. One can verify easily that the action is valid. Note that one must verify that the image of the action is in $\tilde{G}$ as well, which follows from $ab = e \implies ba = e$.

By similar argument as in the proof of Proposition 4.25, there are at least p orbits of cardinality 1. The unique element in such an orbit must admit the form $(g, \cdots, g)$ for some $g \in G$. Indeed, if any two entries differ, then shifting one to the other yields a different set element, leading to a contradiction. One of such possibilities is trivially $(e, \cdots, e)$. Since $p \geq 2$, there must exist some $g \in G \backslash \{e\}$ such that $(g, \cdots, g) \in \tilde{G}$, or $g^p = e$. Hence, the order of g, which is known not to be 1, must divide p, and therefore must equal p. □

5 Rings and Fields

5.1 Basic Definitions

Some prototypical groups we studied in depth were $\mathbb{Z}_n$ and $\mathbb{Z}$. Both have the operation of addition, a commutative group operation, and their structures have been well studied as finitely generated abelian groups. But there is also the operation of multiplication which work very nicely with addition. This section aims to generalize these operations further and gain insights into such structures.

Definition 5.1. A ring $(R, +, \cdot)$ is a set R endowed with two binary operations $+\colon R \times R \to R$ and $\cdot\colon R \times R \to R$ such that

- $(R, +)$ is an abelian group;
- $\cdot$ is associative;
- $(a + b) \cdot c = a \cdot c + b \cdot c$ and $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in R$.

We sometimes omit $\cdot$ and simply juxtapose ring elements ab to denote the multiplication $a \cdot b$. Similar to groups, we may write R rather than the 3-tuple to represent the group. 0 denotes the identity of the group operation of addition.

Some commonplace examples are $\mathbb{Z}_n$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. Importantly, note that $(R\backslash\{0\}, \cdot)$ need not be a group. For example, in $\mathbb{Z}_4$, $2 \cdot 2 = 0$, so $\cdot$ restricted to $R\backslash\{0\}$ is not even a well-defined binary operation.

We introduce some (i.e., many) terminologies regarding rings.

Definition 5.2. A ring R is said to be commutative if $\cdot$ is commutative.

Definition 5.3. A ring R is said to be unitary or a ring of unity if $\cdot$ admits an identity not equal to 0. 1 denotes this identity when it exists.

We make sure the identity is 0 so that the trivial ring $\{0\}$ counts as a ring but not a unitary one.

Note that if a binary operation has an identity, it must be unique. Hence, 1 is well-defined when it exists.

Definition 5.4. A unitary ring R is said to be a division ring if every nonzero element admits a multiplicative inverse; that is, $\forall a \in R\backslash\{0\}, \exists b \in R, ab = ba = 1$. In other words, a ring R is a division rings if $(R\backslash\{0\}, \cdot)$ forms a group.

Definition 5.5. A field is a commutative division ring.

Let's dive into some more examples in greater specificity:

- $\mathbb{Z}$ is a commutative ring but not unitary. It is hence neither a division ring or a field;
- $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all fields;
- $\mathbb{H}$, the set of 2×2 complex matrices of the form

$$\begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix}$$

  where $a, b, c, d \in \mathbb{R}$, is a non-commutative division ring.

Definition 5.6. A zero divisor of a ring R is an non-zero element a such that $ab = 0$ for some non-zero $b \in R\backslash\{0\}$. A commutative ring R with unity is said to be an integral domain if it admits no zero divisors.

Every finite integral domain is a field. But in fact, we can strengthen this result by taking away the guaranteed existence of unity.

Proposition 5.7. A finite commutative ring with no zero divisors is a field.

Proof. Let $a \in R\backslash\{0\}$ be non-zero and suppose $R = \{a_1, \cdots, a_n\}$ has cardinality n.

We first show that the ring is unitary. Let $1 \leq i, j \leq n$ be arbitrary. Then, $a \cdot a_i = a \cdot a_j$ implies $a \cdot (a_i - a_j) = 0$. Because R is an integral domain and $a \neq 0$, we have $a_i - a_j = 0$, or $i = j$. Hence, $a \cdot a_1, \cdots, a \cdot a_n \in R$ are pairwise distinct and must enumerate all elements of R. In particular $a \cdot a_{i'} = a$ for some $1 \leq i' \leq n$. Similarly, for any $b \in R\backslash\{0\}$, $b \cdot a_{j'} = b$ for some $1 \leq j' \leq n$. Now, because R is commutative,

$$0 = ab - ba = (aa_{i'})b - (ba_{j'})a = ab(a_{i'} - a_{j'}).$$

Because $a \neq 0$, $b(a_{i'} - a_{j'}) = 0$; because $b \neq 0$, $a_{i'} = a_{j'}$, or $i' = j'$. Therefore, $a_{i'} = a_{j'}$ is a multiplicative identity which we henceforth call 1.

Because $a \cdot a_1, \cdots, a \cdot a_n$ contains all elements of R, $a \cdot a_k = a_k \cdot a = 1$ for some $1 \leq k \leq n$. □

Corollary 5.8. A finite integral domain is a field.

This result is immediate from the definition of an integral domain.

Corollary 5.9. If p is prime, then $\mathbb{Z}_p$ is a field.

Proof. We appeal to the fact from discrete math that if $p \mid ab$, then $p \mid a$ or $p \mid b$. If $a, b \in \mathbb{Z}_p$ are such that $ab \equiv 0 \pmod{p}$, then $p \mid ab$. This implies either $a = 0$ or $b = 0$. □

Definition 5.10. Suppose R is a unitary ring. An element $a \in R$ is said to be a unit if a admits a multiplicative inverse. The collection of all units of a ring R is denoted as

$$\mathcal{U}(R) = \{a \in R \mid \exists b \in R, ab = ba = 1\}.$$

Proposition 5.11. The units of a unitary ring form a group under multiplication.

Proof. Suppose $a, b \in R$ are units with multiplicative inverses a^{-1} and b^{-1} respectively. Because $(ab)(b^{-1}a^{-1}) = 1$, the product ab is invertible as well, which establishes closure. The associativity of multiplication over units follows from that of the ring. A unitary ring by definition admits a multiplicative identity. The units by definition admit multiplicative inverses. □

We finish this section by introducing the notion of the characteristic of a ring.

Definition 5.12. Let R be a ring. The characteristic of R, denoted as $\operatorname{char} R$, is defined as the least $n \in \mathbb{Z}_{>0}$ such that for all $a \in R$,

$$\underbrace{a + \cdots + a}_{n \text{ copies}} = 0$$

when such an n exists. Otherwise, the characteristic of R is said to be 0.

One can easily factor out the arbitrary a; that is, we can replace the a in the definition above with simply 1.

Considering the finite fields, we have the following:

Proposition 5.13. The characteristic of a field is either a prime or 0. In the latter case, further, there is an injective homomorphism from $\mathbb{Q}$ to the field.

5.2 Fermat's Little Theorem and Euler's Generalization

With the tools that rings provide, we can prove some remarkable results about natural numbers.

Theorem 5.14 (Fermat's Little Theorem)**.** *Suppose a is an integer which is not divisible by a prime number p. Then, p divides $a^{p-1}-1$; that is, $a^{p-1} \equiv 1 \pmod{p}$.*

Proof. We consider the field $\mathbb{Z}_p$ (Corollary 5.9), in which multiplication is the group $(\mathbb{Z}_p^*, \times_p)$, where $\mathbb{Z}_p^* = \mathbb{Z}_p \backslash \{0\}$. Because a is not divisible by p, $a \not\equiv 0 \pmod{p}$; therefore, $\tilde{a} \coloneqq a \bmod p \in \mathbb{Z}_p^*$. Then, the order of $\tilde{a}$ divides $|\mathbb{Z}_p^*| = p-1$ by Corollary 3.22 of Lagrange's theorem. Note that the power of a group element coincides with the power in integer multiplication modulo p. Hence, $\tilde{a}^{p-1} = 1$ implies $\tilde{a}^{p-1} \equiv a^{p-1} \equiv 1 \pmod{p}$. □

Corollary 5.15. Let a be an integer and p a prime number. Then, $a^p \equiv a \pmod{p}$.

Proof. If a is not divisible by p, then $a^{p-1} \equiv 1 \pmod{p}$, and multiplying both sides by a gives the intended result. Otherwise, $a \equiv 0 \pmod{p}$, and the result is trivially satisfied. □

This result is practically useful because it lends us a effective tool for calculating modular exponentiation. For example, 15^{19} modulo 7 is $15^{19} = 15^{3\cdot 6} \cdot 15^1 \equiv 15 \equiv 1 \pmod{7}$.

Can we generalize this result further? For a general n replacing p, the units form a multiplicative group, but they need to cover all $n-1$ elements. We therefore introduce the following function.

Definition 5.16. Euler's totient function $\phi \colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ is defined by $\phi(n) = |\mathcal{U}(\mathbb{Z}_n)|$ for $n \geq 2$ and $\phi(1) = 1$ otherwise.

One may recall a more elementary definition from discrete math, which we state as follows as a Proposition.

Proposition 5.17. For $n \in \mathbb{Z}_{\geq 2}$, $\phi(n)$ equals the number of integers from 1 to n that are coprime to n.

Proof. If $n = 1$, then $\gcd(1,1) = 1$, which coincides with $\phi(1) = 1 = |\{1\}|$. Now suppose $n \geq 2$. Clearly neither 0 or n is coprime to n; it therefore suffices to show that the integers from 1 to $n-1$ coprime to n are precisely the units of the ring $\mathbb{Z}_n$.

Let $a \in \mathbb{Z}_n^*$. Suppose $\gcd(a,n) = 1$. Then, by Bezout's identity, $ax + ny = 1$ for some $x, y \in \mathbb{Z}$, which implies that $ax \equiv 1 \pmod{n}$ and that a admits a multiplicative inverse $x \bmod n$. If instead $\gcd(a,n) = d > 1$, then $a \cdot (n/d)$ is an integer multiple of n, so n/d and a are both 0 divisors. □

Noting that $\phi(p) = p-1$ from the fact that $\mathbb{Z}_p$ is a field, we can generalize Fermat's little theorem by replacing the power $p-1$ with $\phi(p)$.

Theorem 5.18 (Euler)**.** *Suppose $n \in \mathbb{Z}_n$ and $a \in \mathbb{Z}$ is coprime to n. Then, $a^{\phi(n)} \equiv 1 \pmod{n}$.*

Proof. In $\mathbb{Z}_1$, the congruence $a^{\phi(1)} \equiv 1 \equiv 0 \pmod{1}$ is trivial, Now suppose $n \geq 2$. Because a is coprime to n, $a \not\equiv 0 \pmod{n}$; therefore, $\tilde{a} \coloneqq a \bmod n \in \mathcal{U}(\mathbb{Z}_n)$. Then, the order of $\tilde{a}$ divides $|\mathcal{U}(\mathbb{Z}_n)| = \phi(n)$ by Corollary 3.22 of Lagrange's theorem. Note that the power of a group element coincides with the power in integer multiplication modulo n. Hence, $\tilde{a}^{\phi(n)} = 1$ implies $\tilde{a}^{\phi(n)} \equiv a^{\phi(n)} \equiv 1 \pmod{n}$. □

We end this section with an attempt for a formula for $\phi(n)$ in general. Along the way, we happen to recover the Chinese remainder theorem!

Proposition 5.19. Suppose $R_1, \cdots, R_n$ are rings. Then, $\mathcal{U}(R_1 \times \cdots \times R_n) = \mathcal{U}(R_1) \times \cdots \times \mathcal{U}(R_n)$.

Proof. Let $(a_1, \cdots, a_n) \in R_1 \times \cdots \times R_n$ be arbitrary. Then,

$$\begin{aligned}(a_1, \cdots, a_n) \in \mathcal{U}(R_1 \times \cdots \times R_n) &\iff \exists (b_1, \cdots, b_n) \in R_1 \times \cdots \times R_n, a_1 b_1 = b_1 a_1 = \cdots = a_n b_n = b_n a_n = 1 \\ &\iff (\exists b_1 \in R_1, a_1 b_1 = b_1 a_1 = 1) \wedge \cdots \wedge (\exists b_n \in R_n, a_n b_n = b_n a_n = 1)\end{aligned}$$

$$\iff a_1 \in \mathcal{U}(R_1) \wedge \cdots \wedge a_n \in \mathcal{U}(R_n)$$
$$\iff (a_1, \cdots, a_n) \in \mathcal{U}(R_1) \times \cdots \times \mathcal{U}(R_n).$$

The proof is finished. □

Theorem 5.20 (Chinese Remainder Theorem)**.** *Suppose $m_1, \cdots, m_s \in \mathbb{Z}_{\geq 2}$ are pairwise coprime. Then, $\mathbb{Z}_{m_1 \cdots m_s} \simeq \mathbb{Z}_{m_1} \times \cdots \times \mathbb{Z}_{m_s}$ as rings.*

Proof. Let $\phi : \mathbb{Z}_{m_1 \cdots m_s} \to \mathbb{Z}_{m_1} \times \cdots \times \mathbb{Z}_{m_s}$ by $a \mapsto (a \bmod m_1, \cdots, a \bmod m_s)$. Suppose $a, b \in \mathbb{Z}_{m_1 \cdots m_s}$ are arbitrary. Then, $\phi(a+b) = ((a+b) \bmod m_1, \cdots, (a+b) \bmod m_s) = (a \bmod m_1, \cdots, a \bmod m_s) + (b \bmod m_1, \cdots, b \bmod m_s) = \phi(a) + \phi(b)$. Similarly, $\phi(a \cdot b) = ((a \cdot b) \bmod m_1, \cdots, (a \cdot b) \bmod m_s) = (a \bmod m_1, \cdots, a \bmod m_s) \cdot (b \bmod m_1, \cdots, b \bmod m_s) = \phi(a) \cdot \phi(b)$. Therefore, ϕ is a homomorphism of rings.

For injectivity, suppose $\phi(a) = \phi(b)$; that is, $\phi(a - b) = 0$. Then, $\phi(a) = \phi(b) \Rightarrow a - b = 0$; it therefore suffices to show that the kernel of ϕ is trivial. Indeed, if $\phi(a) = 0$, then $a \bmod m_1 = \cdots = a \bmod m_s = 0$; that is, each of $m_1, \cdots, m_s$ divides a. Because $m_1, \cdots, m_s$ are pairwise coprime, their greatest common divisor is 1, so their least common multiple is $m_1 \cdots m_s / 1 = m_1 \cdots m_s$. Hence, $m_1 \cdots m_s \mid a$, and $a = 0$.

For surjectivity, let $(r_1, \cdots, r_s) \in \mathbb{Z}_{m_1} \times \cdots \times \mathbb{Z}_{m_s}$ be arbitrary. Define $N \coloneqq m_1 \cdots m_s$ and $N_i \coloneqq N/m_i$ for each $i = 1, \cdots, m$. Because $m_1, \cdots, m_s$ are pairwise coprime, they comprise disjoint prime factors; hence, N_i and m_i also comprise disjoint prime factors, and $\gcd(N_i, m_i) = 1$. This implies that $\tilde{N}_i \coloneqq N_i \bmod m_i \in \mathcal{U}(\mathbb{Z}_{m_i})$. Let $b_i \in \mathbb{Z}_{m_i}$ be the multiplicative inverse of N_i, so that $c_i \coloneqq r_i b_i$ satisfies $c_i \cdot N_i \equiv r_i \pmod{m_i}$. Let

$$a \coloneqq \sum_{j=1}^{s} c_j N_j.$$

Note that $m_i \mid N_j$ whenever $i \neq j$ by construction. Then,

$$a \bmod m_i = \left(\sum_{j=1}^{s} c_j N_j \bmod m_i \right) \bmod m_i = c_i N_i \bmod m_i = r_i.$$

Hence, $\phi(a) = (r_1, \cdots, r_s)$. The proof is finished. □

Lemma 5.21. Suppose $m_1, \cdots, m_s \in \mathbb{Z}_{\geq 2}$ are pairwise coprime. Then, $\phi(m_1 \cdots m_s) = \phi(m_1) \cdots \phi(m_s)$; that is, ϕ is *weakly multiplicative*.

Proof. Observe that

$$\begin{aligned}
\phi(m_1 \cdots m_s) &= |\mathcal{U}(\mathbb{Z}_{m_1 \cdots m_s})| \\
&= |\mathcal{U}(\mathbb{Z}_{m_1} \times \cdots \times \mathbb{Z}_{m_s})| && \text{(Theorem 5.20)} \\
&= |\mathcal{U}(\mathbb{Z}_{m_1}) \times \cdots \times \mathcal{U}(\mathbb{Z}_{m_s})| && \text{(Proposition 5.19)} \\
&= |\mathcal{U}(\mathbb{Z}_{m_1})| \times \cdots \times |\mathcal{U}(\mathbb{Z}_{m_s})| && \text{(Counting a Cartesian product)} \\
&= \phi(m_1) \cdots \phi(m_s).
\end{aligned}$$

The proof is complete. □

Proposition 5.22. Let p be a prime and suppose $s \in \mathbb{Z}_{>0}$. Then, $\phi(p^s) = p^s - p^{s-1}$.

Proof. Let $a \in \mathbb{Z}_{p^s}$ be arbitrary. Then,

$$\begin{aligned}
a \notin \mathcal{U}(\mathbb{Z}_{p^s}) &\iff \gcd(a, p^s) > 1 \\
&\iff \exists d \in \mathbb{Z}_{\geq 2} \text{ such that } d \mid a \text{ and } d \mid p^s
\end{aligned}$$

$$\iff \exists i \in \mathbb{Z} \cap [1, s] \text{ such that } p^i \mid a$$
$$\iff \exists q \in \mathbb{Z} \cap [0, p^{s-1}) \text{ such that } a = pq.$$

There are exactly p^{s-1} such $q := a/p \in \mathbb{Z} \cap [0, p^{s-1})$. Hence, $|\mathbb{Z}_{p^s} \backslash \mathcal{U}(\mathbb{Z}_{p^s})| = |\{a \in \mathbb{Z}_{p^s} \mid \exists q \in \mathbb{Z} \cap [0, p^{s-1}) \text{ such that } a = pq\}|$ $= |\mathbb{Z} \cap [0, p^{s-1})| = p^{s-1}$. Hence, $|\mathcal{U}(\mathbb{Z}_{p^s})| = |\mathbb{Z}_{p^s}| - |\mathcal{U}(\mathbb{Z}_{p^s})| = p^s - p^{s-1}$. □

Corollary 5.23. Suppose $n \in \mathbb{Z}_{\geq 2}$ has prime factorization $n = p_1^{r_1} \cdots p_s^{r_s}$, where $p_1, \cdots, p_s$ are distinct primes and $r_1, \cdots, r_s \in \mathbb{Z}_{>0}$. Then, $\phi(n) = \prod_{i=1}^{n} (p_i^{s_i} - p_i^{s_i - 1}) = n \cdot \prod_{i=1}^{n} (1 - 1/p_i)$.

6 Constructing Rings and Fields

6.1 The Field of Quotients of an Integral Domain

We will redo the construction of $\mathbb{Q}$ from $\mathbb{Z}$ in real analysis in the context of a more general integral domain D. That is, we set out to construct, in a sense, the "smallest field containing D."

Throughout this section, D shall refer to a given integral domain. We consider the format fractions

$$S = \{(a, b) \in D \times D \mid b \neq 0\},$$

on which we define the equivalence relation $\sim$ via

$$(a, b) \sim (c, d) \qquad \iff \qquad ad = bc.$$

Proposition 6.1. $\sim$ is an equivalence relation.

Proof. Let $a, c, e \in D$ and $b, d, f \in D^*$. Reflexivity is shown by $(a, b) \sim (a, b)$ because $ab = ba$ in a commutative ring D. Symmetry is justified by noting $(a, b) \sim (c, d)$ implies $ad = bc$, so $cb = da$ and $(c, d) \sim (a, b)$. Lastly, $\sim$ is transitive because $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$ imply $ad = bc$ and $cf = de$. Multiplying both the two equation gives $adcf = bcde$, or $cd(af - be) = 0$. Because $c, d \neq 0$, $cd \neq 0$ and $af - be = 0$, so $af = be$ and $(a, b) \sim (e, f)$. □

We will denote the equivalence classes of an element (a, b) of S as $[(a, b)]$, and the collection of all such equivalence classes as $F = \{[(a, b)] \mid (a, b) \in S\}$. We will now define our usual addition and multiplication.

Proposition 6.2. There exist unique binary operations + and $\cdot$ over F such that

$$[(a, b)] + [(c, d)] = (ad + bc, bd) \qquad \text{and} \qquad [(a, b)] \cdot [(c, d)] = [(ac, bd)]$$

for all $a, b, c, d \in D$.

Proof. Suppose $a, a', c, c' \in D$ and $b, b', d, d' \in D^*$ are such that $(a, b) \sim (a', b')$ and $(c, d) \sim (c', d')$. Then, $ab' = a'b$ and $c'd = cd'$. Hence, $(ab' - a'b)dd' = bb'(c'd - cd')$, or $ab'dd' + bb'cd' = a'bdd' + bb'c'd$. Hence, $(ad + bc)b'd' = (a'd' + b'c')bd$, so $(ad + bc, bd) \sim (a'd' + b'c', b'd')$. This shows + is well-defined.

Now, $ab' = a'b$ and $c'd = cd'$ also imply $ab'cd' = a'bc'd$, or $(ac)(b'd') = (a'c')(bd)$, so $(ac, bd) \sim (a'c', b'd')$. Therefore, $\cdot$ is well-defined. □

Now, equipped with the set F with binary operations + and $\cdot$, we are ready to show that it is a field.

Theorem 6.3. *$(F, +, \cdot)$ is a field. Further, there exists an injective homomorphism of rings from D to F.*

Proof. That F is a field follows routine verification. One needs to verify that **(i)** + is associative and commutative, **(ii)** $\cdot$ is associative, and **(iii)** the distributive laws hold. In addition, one needs to show that $[(0, 1)]$ is the identity in $(F, +)$, $[(-a, b)]$ is the additive inverse of $[(a, b)]$, $[(1, 1)]$ is the identity in $(F^*, \cdot)$, and $[(b, a)]$ is the multiplicative identity of $[(a, b)] \in F^*$.

Obviously, the map $\iota : D \to F$ by $k \mapsto [(k, 1)]$ is an injective homomorphism of rings. □

Definition 6.4. Let D be an integral domain. The field of quotients of D is the field F as constructed above.

If we apply this construction on what is already a field, we recover the "same" field back:

Corollary 6.5. A field K is isomorphic to its field of quotients F via the isomorphism $\phi : K \to F, x \to [(x, 1)]$.

Proof. Let $a \in K$ and $b \in K^*$ be arbitrary. Then, $\phi(ab^{-1}) = [(a, b)]$. □

We now make precise the notion that F is the "smallest" field "containing" D.

Theorem 6.6. *Suppose D is an integral domain and K an arbitrary field. Let F be the field of quotients of D and $\iota : D \to F$ the injective homomorphic embedding. If $\phi : D \to K$ is an injective homomorphism, then there exists a unique injective homomorphism $\tilde{\phi} : F \to K$ such that $\tilde{\phi} \circ \iota = \phi$.*

Proof. Let $(a, b) \in S$. Note that $[(a, b)] = [(a, 1)] \cdot [(1, b)] = \iota(a) \cdot \iota(b)^{-1}$. Now,

$$\tilde{\phi}([(a, b)]) = \tilde{\phi}(\iota(a) \cdot \iota(b)^{-1}) = \tilde{\phi}(\iota(a)) \cdot \tilde{\phi}(\iota(b))^{-1} = \phi(a) \cdot \phi(b)^{-1}.$$

We claim the equation above is well-defined and uniquely defines $\tilde{\phi}$. First, note that $b \neq 0$ implies $\phi(b) \neq 0$, so $\phi(b)^{-1}$ exists. Now suppose $(a, b) \sim (a', b')$, or $ab' = ba'$. Then, $\phi(a)\phi(b') = \phi(a')\phi(b)$, or $\phi(a) \cdot \phi(b)^{-1} = \phi(a') \cdot \phi(b')^{-1}$ by the commutativity of $(K^*, \cdot)$. Indeed, for an arbitrary a, $\tilde{\phi}(\iota(a)) = \tilde{\phi}([(a, 1)]) = \phi(a) \cdot \phi(1)^{-1} = \phi(a) \cdot 1 = \phi(a)$. □

6.2 Rings of Polynomials and Their Factorization

One may recall from linear algebra that polynomials with coefficients in a given field form a vector space of infinite dimensions. One may relax the restriction of *field* coefficients to *ring* coefficients. The resultant structure of the polynomials is no longer a vector space, but it is a ring in its own right.

We could view univariate polynomials with coefficients in a ring R as the direct sum

$$\bigoplus_{i=0}^{\infty} R,$$

but this takes away the algebraic structure inherent in polynomials. Rather, we resort to the following formal sum:

Definition 6.7. Let R be a ring and x an indeterminate variable. The polynomials in x with coefficients in R are defined as the following collection of formal series:

$$R[x] \coloneqq \left\{ \sum_{i=0}^{\infty} a_i x^i \mid \text{all but finitely many } a_i\text{'s are non-zero} \right\},$$

where x is assumed to follow the axioms of the ring R.

While we may freely use the algebraic structure of R for x, one should bear in mind the following nuance: it becomes difficult to ask what a statement like "$f(x) = 0$" means: is it that $f(x)$ is the identically 0 polynomial, or is it a question about the zeros of the polynomial f? We avoid qualifiers because x is not a variable. Therefore, to avoid ambiguity, we only ever refer to the latter situation as "x is a zero of f."

One notes the following obvious statement:

Proposition 6.8. Let R be a ring and suppose $\{a_i\}_{i=0}^{\infty}$ is a sequence in R. Then, all but finitely many a_i's are non-zero if and only if some tail of $\{a_i\}_{i=0}^{\infty}$ is identically zero.

It is therefore natural to ask how one may write or denote such a polynomial. It suffices to omit the longest identically zero tail. We therefore write $f(x) = x$ instead of

$$f(x) = \sum_{i=0}^{\infty} a_i x^i, \qquad \text{where } a_i = \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The same train of thought gives rise to the *degree* of a polynomial.

Definition 6.9. Let R be a ring and suppose $f(x) \in R[x]$ where $f(x) = \sum_{i=0}^{\infty} a_i x^i$. The degree of $f(x)$, denoted as $\deg f$, is defined as $\max\{n \in \mathbb{Z}_{\geq 0} \mid a_n \neq 0\}$. Note that $\deg 0 = -\infty$ trivially.

The choice of $\deg 0 \coloneqq -\infty$ is made so that we have the following:

Proposition 6.10. Let R be a ring and suppose $f, g \in R[x]$. Then,

$$\deg(f+g) \leq \max\{\deg f, \deg g\}, \qquad \text{where "=" holds if } \deg f \neq \deg g$$
$$\deg(f \cdot g) = \deg f + \deg g.$$

This is quite an intuitive statement: x plus $3x^2$ has degree $\max\{1, 2\} = 2$ and $(x+1) \cdot (2x)$ has degree $1 + 1 = 2$. One nuance is that the "=" condition is not an "only if" condition: the equality could hold even if $\deg f = \deg g$, but a more detailed discussion is necessary.

We now turn from the definition to the inherent structure of such polynomials: the pointwise addition and multiplication on polynomials are nicely defined in such a way that gives rise to the structure of a ring.

Proposition 6.11. Let R be a ring. Then, $(R[x], +, \cdot)$ is a ring, where $+$ and $\cdot$ denote pointwise operations of addition and multiplication on polynomials by

$$(f+g)(x) \coloneqq f(x) + g(x),$$
$$(f \cdot g)(x) \coloneqq f(x) \cdot g(x)$$

for all $f, g \in R[x]$.

The routine proof of checking the axioms is needed though by no means challenging, omitted here for brevity. Note that one can think directly in terms of the coefficients. If $f(x) = \sum a_i x^i$ and $g(x) = \sum b_i x^i$, then

$$f(x) + g(x) = \sum (a_i + b_i) x^i \qquad \text{and} \qquad f(x) \cdot g(x) = \sum c_i x^i,$$

where $c_i = a_0 b_i + a_1 b_{i-1} + \cdots + a_{i-1} b_1 + a_i b_0$.

The idea of division is so powerful that it spans our elementary education to college. The long division algorithm is another such example that has wide applications in our discussions.

We will henceforth restrict our attentions to fields rather than general rings. The main reason is that without a field, we could have

$$(x-1)(x+1) = (x-3)(x+3),$$

if we are working in a ring where $1 = 9$ like $\mathbb{Z}_8$. As we will later see, this is guaranteed not to occur in a field. At the same time, we can't divide by anything unless it's a unit.

Proposition 6.12. Let F be a field. Then, $F[x]$ is an integral domain.

Proof. Suppose $f(x) \cdot g(x) = 0$, where $f(x), g(x) \in F[x]$ and $f(x) \neq 0$. Then, $\deg f + \deg g = \deg 0 = -\infty$, where $\deg f \geq 0$. The only possibility is $\deg g = -\infty$; that is, $g(x) = 0$. □

Theorem 6.13 (Long Division)**.** *Let F be a field and suppose $f(x) \in F[x]$ and $g(x) \in F[x]\backslash\{0\}$. Then, there exist unique polynomials $q(x), r(x) \in R[x]$, where $\deg r < \deg g$, such that $f(x) = g(x) \cdot q(x) + r(x)$.*

Proof. Suppose $f(x) = a_0 + \cdots + a_n x^n$ and $g(x) = b_0 + \cdots + b_m x^m$, where $a_n \neq 0$ and $b_m \neq 0$.

Existence. If $\deg f < \deg g$, then we may simply take $q(x) = 0$ and $r(x) = f(x)$. Otherwise, we consider the following recursive procedure:

- While $\deg f \geq \deg g$, repeat:
  - Let $\tilde{q}(x) \coloneqq a_n b_m^{-1} \cdot x^{n-m}$ and $\tilde{r}(x) = f(x) - g(x) \cdot \tilde{q}(x)$;
  - Observe that
    $$g(x) \cdot \tilde{q}(x) = (b_0 + \cdots + b_m x^m) \cdot \frac{a_n}{b_m} x^{n-m} = \frac{b_0 a_n}{b_m} x^{n-m} + \cdots + a_n x^n;$$
    that is, $\deg(g \cdot q) = n = \deg f$ and $g \cdot q$ have the same leading coefficient as f. Then, $\tilde{r}(x)$ must have degree strictly less than n;
  - Update $f(x)$ to be $\tilde{r}(x)$.
- Output $q(x)$ as the sum of all such $\tilde{q}(x)$'s above and $r(x)$ as the final $\tilde{r}(x)$.

Indeed, given

$$\begin{aligned} f &= g \cdot \tilde{q}_1 + \tilde{r}_1, \\ \tilde{r}_1 &= g \cdot \tilde{q}_2 + \tilde{r}_2, \\ &\vdots \\ \tilde{r}_{s-1} &= g \cdot \tilde{q}_s + \tilde{r}_s, \end{aligned}$$

one plugs back to obtain

$$\begin{aligned} \tilde{r}_{s-1} &= g \cdot \tilde{q}_s + \tilde{r}_s, \\ \tilde{r}_{s-2} &= g \cdot \tilde{q}_{s-1} + g \cdot \tilde{q}_s + \tilde{r}_s = g \cdot (\tilde{q}_{s-1} + \tilde{q}_s) + \tilde{r}_s, \\ &\vdots \\ \tilde{r}_1 &= g \cdot (\tilde{q}_2 + \cdots + \tilde{q}_s) + \tilde{r}_s, \\ f &= g \cdot \underbrace{(\tilde{q}_1 + \tilde{q}_2 + \cdots + \tilde{q}_s)}_{q(x)} + \underbrace{\tilde{r}_s}_{r(x)}. \end{aligned}$$

This validates the existence of such $q(x)$ and $r(x)$.

Uniqueness. Suppose now that $f = g \cdot q + r = g \cdot q' + r'$; that is, $g \cdot (q - q') = r' - r$. Note that $\deg(g \cdot (q - q')) = \deg g + \deg(q - q')$, while $\deg(r' - r) \leq \max\{\deg r', \deg r\} < \deg g$. Unless $\deg(q - q') = -\infty$, we cannot possibly obtain the supposed equality. Hence, $q = q'$, which forces $r' - r = g \cdot 0 = 0$, and this implies $r = r'$. The proof is now complete. □

Corollary 6.14. Let F be a field, $f(x) \in F[x]\backslash\{0\}$, and $\alpha \in F$. Then, α is a zero of $f(x)$ if and only if $f(x) = (x - \alpha) \cdot g(x)$ for some non-zero $g(x) \in F[x]\backslash\{0\}$.

Proof. Suppose α is a zero of $f(x)$. Let $f(x) = (x-\alpha) \cdot q(x) + r(x)$, where $q(x), r(x) \in F[x]$ are the quotient and the remainder from the division algorithm. Then, $r(x)$ has degree at most 0, which must be a constant c. Then, $f(\alpha) = 0 \cdot q(\alpha) + c = c = 0$, so $r(x) = c = 0$. Hence, $f(x) = (x - \alpha) \cdot g(x)$. If $g(x)$ were 0, then $f(x) = 0$ also, a contradiction. We may therefore take $q(x)$ as $g(x)$.

Conversely, let $f(x) = (x-\alpha)\cdot g(x)$ for some $g(x) \in F[x]$. Then, $f(\alpha) = 0 \cdot g(\alpha) = 0$. The proof is complete. □

This allows us to recover the following upper-bound on the number of zeros of a polynomial over a field.

Corollary 6.15. Suppose $f(x) \in F[x]\backslash\{0\}$ is a non-zero polynomial over a field F. Then, $f(x)$ has no more zeros than $\deg f$.

Proof. Suppose $\alpha_1, \cdots, \alpha_n \in F$ are distinct zeros of $f(x)$. We will show inductively that for all $i \in \{1, \cdots, n\}$, there exists non-zero $g_i(x) \in F[x]\backslash\{0\}$ such that $f(x) = (x-\alpha_1)\cdots(x-\alpha_i)\cdot g_i(x)$.

Base case. Because α_1 is a zero of $f(x)$, $f(x) = (x-\alpha_1)\cdot g_1(x)$ for some $g_1(x) \in F[x]\backslash\{0\}$ by the preceding Corollary.

Inductive case. Let $i \in \{2, \cdots, n\}$ and suppose inductively that $f(x) = (x-\alpha_1)\cdots(x-\alpha_{i-1})\cdot g_{i-1}(x)$ for some $g_{i-1}(x) \in F[x]\backslash\{0\}$. Note that evaluating the polynomial $(x-\alpha_1)\cdots(x-\alpha_{i-1})$ at $x = \alpha_i$ does not yield 0 because non-zero elements of F are multiplied. Then, because F is an integral domain, $f(\alpha_i) = 0$ implies $g_{i-1}(\alpha_i) = 0$; that is, α_i is a zero of $g_{i-1}(x)$. By the preceding corollary, $g_{i-1}(x) = (x-\alpha_i)\cdot g_i(x)$ for some $g_i(x) \in F[x]\backslash\{0\}$. Therefore, $f(x) = (x-\alpha_1)\cdots(x-\alpha_i)\cdot g_i(x)$.

Hence, in particular, $f(x) = (x-\alpha_1)\cdots(x-\alpha_n)\cdot g_n(x)$ for some $g_n(x) \in F[x]\backslash\{0\}$. Because $g_n(x) \neq 0$, we see that $f(x)$ has degree at least n. □

The last corollary of the division algorithm is extremely powerful. It allows us to assert that the units of a field always form a *cyclic* multiplicative group. This statement would otherwise be extremely hard to prove.

Corollary 6.16. Every finite subgroup of the multiplicative group $(F^*, \cdot)$ of a field F is cyclic.

Proof. Observe that $(F^*, \cdot)$ is a finitely generated abelian group, which admits an invariant factorization $\mathbb{Z}_{d_1} \times \cdots \times \mathbb{Z}_{d_n}$ via $\phi\colon (G, \cdot) \to (\mathbb{Z}_{d_1} \times \cdots \times \mathbb{Z}_{d_n}, +)$. Suppose for the sake of contradiction that $n \geq 2$. Then, $|\langle \phi^{-1}(1, 0, \cdots, 0)\rangle|$ has degree d_1. Further, for each $i \in \mathbb{Z}_{d_1}$,

$$\begin{aligned} |\langle \phi^{-1}(i, d_2/d_1, 0, \cdots, 0)\rangle| &= |\langle (i, d_2/d_1, 0, \cdots, 0)\rangle| \\ &= \operatorname{lcm}(d_1/\gcd(i, d_1), d_1, 1, \cdots, 1) \\ &= d_1. && (d_1/\gcd(i, d_1) \text{ divides } d_1) \end{aligned}$$

We have identified $1 + d_1 > d_1$ zeros of $g^n - 1 \in F[g]$, which is impossible. Hence, $n = 1$ and $(F^*, \cdot) \simeq (\mathbb{Z}_{d_1}, +)$ is cyclic. □

In addition to these powerful corollaries, the division algorithm also lends us additional machinery to talk about factoring polynomials.

Definition 6.17. A polynomial $f(x) \in F[x]$ is said to be reducible if there exist $g(x), h(x) \in F[x]$ with $\max\{\deg g, \deg h\} < \deg f$ such that $f(x) = g(x)\cdot h(x)$. Otherwise, it is said to be irreducible.

The versatility of polynomials, when extended to general fields, is manifold. For one, it is obvious that $\sqrt{2} \in \mathbb{Q}$ if and only if $x^2 - 2 \in \mathbb{Q}[x]$ is reducible. Hence, questions about rationality are really just questions about the reducibility of polynomials in $\mathbb{Q}[x]$.

Proposition 6.18. Let $f(x) \in F[x]$ have degree $\deg f \in \{2, 3\}$. Then, $f(x)$ is reducible if and only if it has a zero.

Proof. Suppose first that $f(x)$ is reducible as $g(x)\cdot h(x)$, with $\max\{\deg g, \deg h\} < \deg f$. Without loss of generality, suppose $\deg g \leq \deg h$. Because $\deg g + \deg h = \deg f$, it follows that $\deg g = 1$. Hence, $g(x) = a\cdot x + b$ for some $a \in F^*$ and $b \in F$, and $f(-b/a) = (a\cdot(-b/a) + b)\cdot g(-b/a) = 0 \cdot g(-b/a) = 0$.

Conversely, suppose $f(\alpha) = 0$ for some $\alpha \in F$. Then, $f(x) = (x-\alpha)\cdot g(x)$ for some $g(x) \in F[x]$. Because $\deg g = \deg f - 1 < 1$, the factorization above exemplifies the reducibility of $f(x)$. □

Below, we provide a powerful criterion for factoring polynomials in $\mathbb{Z}[x]$, due to Gauss.

Lemma 6.19 (Gauss)**.** Let $f(x) \in \mathbb{Z}[x]$. Then, $f(x)$ reduces to a product $g(x) \cdot h(x)$ of polynomials $g(x), h(x) \in \mathbb{Q}[x]$ of degree r and s if and only if $f(x)$ reduces to a product $g'(x) \cdot h'(x)$ of polynomials $g'(x), h'(x) \in \mathbb{Z}[x]$ of degree r and s.

Some machinery is necessary for a detailed proof.

Definition 6.20. The content of a polynomial $f(x) \in \mathbb{Z}[x]$, denoted as $\text{cont}(f(x))$, is defined as the greatest common divisor of the coefficients of $f(x)$. Such $f(x)$ is said to be primitive if $\text{cont}(f(x)) = 1$.

Proposition 6.21. Let p be a prime and $f(x), g(x) \in \mathbb{Z}[x]$. If p divides all coefficients of the product $f(x) \cdot g(x)$, then p divides all coefficients of $f(x)$ or p divides all coefficients of $g(x)$.

Proof. Suppose $f(x) = a_0 + \cdots + a_r x^r$, $g(x) = b_0 + \cdots + b_s x^s$, and $P(x) = f(x) \cdot g(x) = c_0 + \cdots + c_n x^n$. Assume for contradiction that p neither divides all $a_0, \cdots, a_r$ nor divides all $b_0, \cdots, b_s$. Let $i \in \{1, \cdots, r\}$ and $j \in \{1, \cdots, s\}$ be the minimal value such that $p \nmid a_i$ and $p \nmid b_j$ respectively.

Note that

$$p \mid c_{i+j} \qquad \Rightarrow \qquad p \mid (\underbrace{a_0 b_{i+j} + \cdots}_{\text{part (i)}} + a_i b_j + \underbrace{\cdots + a_{i+j} b_0}_{\text{part (ii)}}),$$

where each $a_0, \cdots, a_{i-1}$ in part (i) is divisible by p and each $b_{j-1}, \cdots, b_0$ in part (ii) is divisible by p as well. Because p divides the entire sum, this compels $p \mid a_i b_j$; that is, $p \mid a_i$ or $p \mid a_j$, a contradiction. □

Corollary 6.22. Primitive polynomials are closed under multiplication.

Proof. Let $f(x), g(x) \in \mathbb{Z}[x]$ be primitive. Let d be the greatest common divisor of the coefficients of the product $f(x) \cdot g(x)$. If $d > 1$, then choose an arbitrary prime p from the prime factorization of d. Note that $p \mid d$ divides all coefficients of $f(x) \cdot g(x)$. Then, p divides all coefficients of $f(x)$ or p divides all coefficients of $g(x)$. But either scenario is absurd, since $f(x)$ and $g(x)$ are both assumed to be primitive. □

Corollary 6.23. The content is multiplicative over $\mathbb{Z}[x]$; that is, $\text{cont}(f(x) \cdot g(x)) = \text{cont}(f(x)) \cdot \text{cont}(g(x))$ for all $f(x), g(x) \in \mathbb{Z}[x]$.

Proof. Let $d_1 = \text{cont}(f(x))$ and $d_2 = \text{cont}(g(x))$. Divide all coefficients of f and g by d_1 and d_2 respectively to obtain primitive polynomials $\tilde{f}(x), \tilde{g}(x)$. Then, $\tilde{f}(x) \cdot \tilde{g}(x)$ is primitive as well. Hence, the content of $f(x) \cdot g(x) = d_1 d_2 \cdot \tilde{f}(x) \cdot \tilde{g}(x)$ is $d_1 d_2 = \text{cont}(f(x)) \cdot \text{cont}(g(x))$. □

We finally have the machinery to prove Gauss' lemma.

Proof. The "if" direction is trivial. For the "only if" direction, suppose $f(x) \in \mathbb{Z}[x]$ is reducible in $\mathbb{Q}[x]$ as $(a_0 + \cdots + a_n x^n) \cdot (b_0 + \cdots + b_m x^m)$. Let $A, B \in \mathbb{Z}_{>0}$ be such that $A \cdot a_0, \cdots, A \cdot a_n, B \cdot b_0, \cdots, B \cdot b_m \in \mathbb{Z}$. For $i \in \{0, \cdots, n\}$ and $j \in \{0, \cdots, m\}$, let $\alpha_i = A \cdot a_i$ and $\beta_j = B \cdot b_j$, set $d_\alpha = \gcd(\alpha_0, \cdots, \alpha_n)$ and $d_\beta = \gcd(\beta_0, \cdots, \beta_m)$, and define $\tilde{\alpha}_i = \alpha_i / d_\alpha$ and $\tilde{\beta}_i = \beta_i / d_\beta$. Then,

$$AB \cdot f(x) = d_\alpha d_\beta \cdot (\tilde{\alpha}_0 + \cdots + \tilde{\alpha}_n x^n) \cdot (\tilde{\beta}_0 + \cdots + \tilde{\beta}_m x^m).$$

Now,

$$AB \cdot \text{cont}(f(x)) = \text{cont}(AB \cdot f(x)) = \text{cont}(\alpha_0 + \cdots + \alpha_n x^n) \cdot \text{cont}(\beta_0 + \cdots + \beta_m x^m) = d_\alpha \cdot d_\beta,$$

so $d_\alpha \cdot d_\beta / AB = \text{cont}(f(x)) \in \mathbb{Z}_{>0}$ is an integer. Therefore,

$$f(x) = \frac{d_\alpha \cdot d_\beta}{AB} (\tilde{\alpha}_0 + \cdots + \tilde{\alpha}_n x^n) \cdot (\tilde{\beta}_0 + \cdots + \tilde{\beta}_m x^m) = \left(\frac{d_\alpha \cdot d_\beta \cdot \tilde{\alpha}_0}{AB} + \cdots + \frac{d_\alpha \cdot d_\beta \cdot \tilde{\alpha}_n}{AB} x^n \right) \cdot (\tilde{\beta}_0 + \cdots + \tilde{\beta}_m x^m).$$

□

Corollary 6.24. Suppose $f(x) = a_0 + \cdots + a_{n-1}x^{n-1} + x^n \in \mathbb{Z}[x]$ is a monic polynomial with $a_0 \neq 0$ and $f(x)$ has a zero in $\mathbb{Q}$. Then, $f(x)$ has a zero in $\mathbb{Z}$ which divides a_0.

Proof. Suppose $\alpha \in \mathbb{Q}$ is a zero of $f(x)$. Then, by the division algorithm we may write $f(x) = (x - \alpha) \cdot g(x)$ for some $g(x) \in \mathbb{Q}[x]$. Because $f(x)$ is monic, the factorization of $f(x)$ in $\mathbb{Z}[x]$ guaranteed by Gauss' lemma must admit the form

$$f(x) = (x - m) \cdot (x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_1 x - a_0/m).$$

Now, m is immediately an integral zero. Further, $m \mid a_0$ because $-a_0/m \in \mathbb{Z}$. □

We finish this section with another useful criterion for irreducibility.

Proposition 6.25 (Eisenstein)**.** Let p be a prime. Suppose $f(x) = a_0 + \cdots + a_n x^n \in \mathbb{Z}[x]$ is such that

- $a_n \not\equiv 0 \pmod{p}$,
- $a_0, \cdots, a_{n-1} \equiv 0 \pmod{p}$, and
- $a_0 \not\equiv 0 \pmod{p^2}$,

then $f(x)$ is irreducible in $\mathbb{Q}[x]$.

Proof. By Gauss' lemma (Lemma 6.19), it is sufficient to show that $f(x)$ is irreducible in $\mathbb{Z}[x]$. Suppose on the contrary that $f(x) = (b_r x^r + \cdots + b_0) \cdot (c_s x^s + \cdots + c_0)$, where b_r and c_s are non-zero and $\max\{r, s\} < n$. Because $p \mid b_0 c_0$ but $p^2 \mid b_0 c_0$, exactly one of b_0 and c_0 is divisible by p. Without loss of generality, let $p \mid c_0$ but $p \nmid b_0$. Similarly, $b_r c_s \not\equiv 0 \pmod{p}$ implies neither b_r nor c_s is divisible by p.

Let $m = \min\{k \in \{0, \cdots, s\} \mid p \nmid c_k\}$. If $r \geq m$, then

$$a_m = b_0 c_m + \cancel{b_1 c_{m-1} + \cdots + b_m c_0} = b_0 c_m \not\equiv 0 \pmod{p};$$

if $r < m$, then

$$a_m = b_0 c_m + \cancel{b_1 c_{m-1} + \cdots + b_r c_{m-r}} = b_0 c_m \not\equiv 0 \pmod{p}.$$

The only possibility for m, then, is $m = n$. But this means $n = m \leq s \leq n$, so $s = n$, a contradiction. □

6.3 Homomorphisms and Factor Rings

We will attempt to generalize the concept of factor groups to rings. Clearly, every additive subgroup of a ring is normal and we can construct the additive factor group, but there's no reason to expect it to respect the multiplicative structure of the ring. It turns out a new criterion is required, which unsurprisingly demands some sort of closure under multiplication.

Definition 6.26. Let R be a ring and $(N, +) \leq (R, +)$ an additive subgroup. Then, N is said to be an ideal, denoted as $(N, +, \cdot) \trianglelefteq (R, +, \cdot)$,[2] if

$$rN \coloneqq \{rn \mid n \in N\} \subseteq N \qquad \text{and} \qquad Nr \coloneqq \{nr \mid n \in N\} \subseteq N$$

for all $r \in R$.

We already know that addition would be well-defined; we'll now show that the ideal condition above is precisely what makes the (additive) cosets compatible under multiplication.

Proposition 6.27. Let R be a ring and $N \trianglelefteq R$. Then, there is a unique binary operation $\cdot$ on the cosets R/N such that

$$(a + N) \cdot (b + N) = ab + N$$

for all $a, b \in R$. Further, this operation agrees with element-wise multiplication of the sets.

[2] When unambiguous, we sometimes write simply $N \trianglelefteq R$.

Proof. Let $a, a', b, b' \in R$ with $a - a' \in N$ and $b - b' \in N$. Fix $n_1, n_2 \in N$ such that $a' = a + n_1$ and $b' = b + n_2$. Then,

$$a'b' - ab = (a + n_1)(b + n_2) - ab = an_2 + n_b + n_1n_2 \in N$$

by definition.

Finally, observe that the element-wise product is

$$\{(a + n_1)(b + n_2) \mid n_1, n_2 \in N\} = \{ab + \underbrace{an_2 + n_1b + n_1n_2}_{\text{surjective from } N \times N \text{ to } N} \mid n_1, n_2 \in N\} = ab + N.$$

□

Corollary 6.28. Let R be a ring and $N \trianglelefteq R$. Then, $(R/N, +, \cdot)$ is a ring which is called the factor ring of R by N.

The proof is omitted; the ring axioms for R/N follow immediately from those of R.

One may recall the heavy use of homomorphisms earlier to get a better understanding of the structures of (factor) groups. Similarly, we will make use of this machinery which we now develop.

Definition 6.29. Let R, R' be rings. A map $\phi \colon R \to R'$ is said to be a homomorphism of rings if

$$\phi(a + b) = \phi(a) + \phi(b) \qquad \text{and} \qquad \phi(a \cdot b) = \phi(a) \cdot \phi(b)$$

for all $a, b \in R$.

Every homomorphism of rings is obviously also a homomorphism of the underlying additive groups, so there are no more ring homomorphisms than there are underlying additive group homomorphisms. It is in general much more restrictive though. For example, while there are infinitely many homomorphisms from $(\mathbb{Z}, +)$ to $(\mathbb{Z}, +)$ by mapping $a \mapsto na$ for any fixed $n \in \mathbb{Z}_{\geq 0}$, there are only 2 from $(\mathbb{Z}, +, \cdot)$ to $(\mathbb{Z}, +, \cdot)$. This is because $\phi(1) = \phi(1)^2$, or $n = n^2$, compels $n \in \{0, 1\}$.

Homomorphisms of rings share almost identical properties with those of groups. After all, all "homomorphisms" are is a map that preserves structures.

Proposition 6.30. Let $\phi \colon R \to R'$ be a homomorphism of rings. Then,

- $\phi(0) = 0$;
- $\phi(-a) = -\phi(a)$ for all $a \in R$;
- If $S \leq R$, then $\phi[S] \leq R'$;
- If $S' \leq R'$, then $\phi^{-1}[S'] \leq R$;
- If R is unitary, then $\phi[R]$ is unitary and $\phi(1_R) = 1_{\phi[R]}$;
- If $N \trianglelefteq R$, then $\phi[N] \trianglelefteq \phi[R]$;
- If $N' \trianglelefteq \phi[R]$ or $N' \trianglelefteq R'$, then $\phi^{-1}[N'] \trianglelefteq R$.

We are now ready to state and prove the fundamental homomorphism theorem for rings.

Theorem 6.31 (Fundamental Homomorphism Theorem for Rings)**.** *Let $\phi \colon R \to R'$ be a homomorphism of rings, $N \coloneqq \ker \phi$, and $\gamma \colon R \to R/N$ via $a \mapsto a + N$. Then, there exists an injective homomorphism $\mu \colon R/N \to R'$ such that $\phi = \mu \circ \gamma$.*

Proof. We claim that $\mu(a + N) \coloneqq \phi(a)$ is well-defined. If $a + N = a' + N$ where $a, a' \in R$, then $a - a' \in N$ and $\phi(a - a') = 0$, which implies $\phi(a) = \phi(a')$.

Let $b \in R$. Then, $\mu(a + N) + \mu(b + N) = \phi(a) + \phi(b) = \phi(a + b) = \mu((a + b) + N)$ and $\mu(a + N) \cdot \mu(b + N) = \phi(a) \cdot \phi(b) = \phi(a \cdot b) = \mu(a \cdot b + N)$. □

6.4 Prime and Maximal Ideals

This section is an extension of the discussion of simple groups and maximally normal subgroups. We begin by showing that fields are "simple."

Proposition 6.32. Let R be a ring with unity and $N \trianglelefteq R$. If N contains a unit in R, then $N = R$.

Proof. Suppose $a \in N$ is a unit. Then, because $a^{-1} \in R$, $a^{-1}a = 1 \in N$. Hence, for all $r \in R$, $r \cdot 1 = r \in N$. Because $N \subseteq R$ and $N \supseteq R$, we have $N = R$. □

Corollary 6.33. The only ideals of a field are the trivial and improper ideals.

Hence, a field is analogous to a simply group in the sense that it is non-trivial and contains no non-trivial, proper ideal. We know that G/N is simple iff N is maximally normal. In the same spirit, we define a maximal ideal of a ring to obtain a similar result.

Definition 6.34. An ideal M of a ring R is said to be a maximal ideal if M is proper and no proper ideal of R properly contains M.

Proposition 6.35. Let R be a commutative ring with unity and $M \trianglelefteq R$. Then, M is a maximal ideal of R if and only if R/M is a field.

Proof. Let $M \triangleleft R$ be maximal. Then, for all $a \in R \backslash M$, $aR + M$ is an ideal of R properly containing M, which must be the entire ring; that is, $aR + M = R \ni 1$. Hence, $ar + m = 1$ for some $r \in R$ and $m \in M$, which implies $ar + M = 1 + M$, or $(a + M) \cdot (r + M) = 1 + M$. Hence, any non-zero coset $a + M \neq 0 + M$ is invertible.

Conversely, let R/M be a field with $M \subsetneq I \triangleleft R$. Then, I/M is a non-trivial ideal of R/M under the canonical homomorphism $\gamma : R \to R/M$, and hence I/M is the entire field R/M. Now, for each $r \in R$, there must exist some $i \in I$ and $m \in M$ such that $r - i = m \in M \subset I$, so $r \in I + I = I$. Therefore, $R \subseteq I$ and $I \subseteq R$, which implies $R = I$. □

We also introduce the notion of a prime ideal. The elementary fact that $p \mid ab \implies p \mid a \vee p \mid b$ can be rewritten as $ab \in p\mathbb{Z} \implies a \in p\mathbb{Z} \vee b \in p\mathbb{Z}$. Such a $p\mathbb{Z}$ is said to be a prime ideal of $\mathbb{Z}$.

Definition 6.36. A proper ideal I of a ring R is said to be a prime ideal if $ab \in I$ implies $a \in I$ or $b \in I$ for all $a, b \in I$.

Clearly, I is zero in R/I, so we have the following equivalence immediately.

Corollary 6.37. Let R be a commutative ring with unity and $I \trianglelefteq R$ an ideal. Then, I is prime if and only if R/I is an integral domain.

Meanwhile, a field is always an integral domain. Therefore,

Corollary 6.38. Let R be a commutative ring with unity and $M \triangleleft R$ a maximal ideal. Then, M is a prime ideal.

Now, we turn to a way of formalizing the following idea. A ring with unity has 1, so it has "2" = 1 + 1, "3" = 1 + 1 + 1, etc. Meanwhile, it also has -1, -2, etc.

Proposition 6.39. Let R be a ring with unity. Then, $\phi : \mathbb{Z} \to R$ via

$$n \mapsto \text{“}n \cdot 1\text{”} = \begin{cases} \overbrace{1 + \cdots + 1}^{n \text{ copies}}, & \text{if } n \geq 0, \\ -\phi(-n), & \text{otherwise} \end{cases}$$

is a homomorphism of rings.

Corollary 6.40. A field F either has prime characteristic p and contains a subfield isomorphic to $\mathbb{Z}_p$ or has characteristic 0 and contains a subfield isomorphic to $\mathbb{Q}$.

We finish this section with some discussions on the ideal structure of $F[x]$. We first introduce the notion of principal ideals, which correspond with cyclic subgroups.

Definition 6.41. Let R be a commutative ring with identity and $a \in R$. The principal ideal generated by a, denoted as $\langle a \rangle$, is defined as $aR = \{ar \mid r \in R\}$. An ideal $I \trianglelefteq R$ is said to be a principal ideal if $I = \langle a \rangle$ for some $a \in R$.

We have seen before that all subgroups of $\mathbb{Z}$ are of the form $n\mathbb{Z}$ for some $n \in \mathbb{Z}$. Because a subring is an additive subgroup, this implies that all ideals of $\mathbb{Z}$ are principal. This is true for $F[x]$ as well for a general field F.

Definition 6.42. An integral domain D is said to be a principal ideal domain (PID) if every ideal of D is principal.

Proposition 6.43. The polynomials over a field form a principal ideal domain.

Proof. Suppose I is an ideal of $F[x]$. If $I = \{0\}$, then $I = \langle 0 \rangle$ trivially. Now suppose I is non-trivial, and fix a non-zero polynomial $g(x) \in I \backslash \{0\}$ with minimal degree $\min \deg[I \backslash \{0\}]$. Then, for any $f(x) \in I$, we may apply the long division algorithm (Theorem 6.13) to obtain $f(x) = g(x) \cdot q(x) + r(x)$, where $q(x), r(x) \in F[x]$ with $\deg r < \deg g$. Because $g(x) \in I$, we have $g(x) \cdot q(x) \in I$, and thus $r(x) = f(x) - g(x) \cdot q(x) \in I$. Because $\deg r < \deg g = \min \deg[I \backslash \{0\}]$, we conclude that $r(x) = 0$. Therefore, for any $f(x) \in I$, $f(x) = g(x) \cdot q(x)$ for some $q(x) \in I$. Hence, $I = g(x)R = \langle g(x) \rangle$ is a principal ideal. □

In general, the generator of an ideal in an integral domain, when it exists, is unique up to multiplication by a unit. We provide a detailed proof for the specific case concerning the domain of polynomials over a field.

Corollary 6.44. Let F be a field. Then, for every ideal I there exists $f(x) \in F[x]$, unique up to multiplication by a non-zero constant in F, such that $I = \langle f(x) \rangle$

Proof. If $I = \{0\}$, then $\langle f(x) \rangle = 0$ implies that $f(x) \cdot g(x) = 0$ for all $g(x) \in F[x]$. Fixing a particular non-zero $g(x) \in F[x] \backslash \{0\}$, this implies that $f(x) = 0$ by Proposition 6.12.

Suppose now that $I = \langle f(x) \rangle = \langle g(x) \rangle$, where $f(x), g(x) \in F[x] \backslash \{0\}$. In particular, $f(x) \in \langle g(x) \rangle = g(x)R$, so $f(x) = g(x) \cdot h(x)$ for some $h(x) \in F[x]$. By symmetry, we can conclude as well that $g(x) = f(x) \cdot h'(x)$ for some $h'(x) \in F[x]$. Thus, $f(x) = f(x) \cdot h'(x) \cdot h(x)$, which implies that $f(x) \cdot (1 - h'(x) \cdot h(x)) = 0$. Because $f(x) \neq 0$, this implies $1 - h'(x) \cdot h(x) = 0$. Therefore, $\deg h + \deg h' = \deg 1 = 0$, where $\deg h, \deg h' \in \mathbb{Z}_{\geq 0} \cup \{-\infty\}$. The only possibility is when $\deg h = \deg h' = 0$, which implies $h(x) \cdot h'(x) = 1$. Consequently, both $h(x)$ and $h'(x)$ are non-zero constants. □

Next, we consider the maximal ideals of $F[x]$, which allows us to conclude $F[x]/\langle f(x) \rangle$ is a field for certain $f(x)$'s. It turns out the condition is precisely when $f(x)$ is irreducible.

Proposition 6.45. Let F be a field and $f(x) \in F[x]$. Then, $\langle f(x) \rangle$ is a maximal ideal of $F[x]$ if and only if $f(x)$ is irreducible.

Proof. Suppose $\langle f(x) \rangle$ is a maximal ideal and let $f(x) = g(x) \cdot h(x)$, where $g(x), h(x) \in F[x]$. Then, $\langle f(x) \rangle = f(x)R = g(x)h(x)R \subseteq g(x)R = \langle g(x) \rangle$; that is, $\langle f(x) \rangle \subseteq \langle g(x) \rangle \subseteq F[x]$. Then, $\langle g(x) \rangle$ is either $\langle f(x) \rangle$, in which case $g(x) = f(x)$ and $h(x) = 1$, or $F[x] = \langle 1 \rangle$, in which case $g(x) = 1$ and $h(x) = f(x)$ (by the uniqueness of the long division algorithm). In either cases, the condition that $\max\{\deg g, \deg h\} < \deg f$ is violated. Therefore, $f(x)$ is irreducible.

Conversely, suppose $f(x)$ is irreducible. Let $g(x) \in F[x]$ be arbitrary, where $\langle f(x) \rangle \subseteq \langle g(x) \rangle \subseteq F[x]$. In particular, $f(x) \in \langle g(x) \rangle$, so $f(x) = g(x) \cdot h(x)$ for some $h(x) \in F[x]$. Because $f(x)$ is irreducible, either $g(x) = 1$ or $h(x) = 1$. If $g(x) = 1$, then $\langle g(x) \rangle = 1F[x] = F[x]$. If $h(x) = 1$, then $g(x) = f(x)$ by the uniqueness of long division, so $\langle g(x) \rangle = \langle f(x) \rangle$. In either cases, we fail to obtain an ideal $\langle g(x) \rangle$ strictly between $\langle f(x) \rangle$ and $F[x]$. The arbitrary choice of $g(x)$ therefore implies that $\langle f(x) \rangle$ is a maximal ideal. □

Corollary 6.46. Let F be a field and $f(x) \in F[x]$. Then, $F[x]/\langle f(x) \rangle$ is a field if and only if $f(x)$ is irreducible.

7 Extension Fields

7.1 Introduction to Extension Fields

We will work towards a big goal: every polynomial can be forced to have a zero. Why is this any significant? Let's take a look at the construction of $\mathbb{C}$. We imagined a "number" i out of nowhere, assuming that $i^2 = -1$. This assumption is equivalent to saying that i is a zero of $x^2 + 1 = 0$, or simply that $x^2 + 1$ has a zero, where we can only ever reasonably assume x is over $\mathbb{R}$. After all, $\mathbb{C}$ doesn't exist yet! I've always felt a bit iffy that we made it out of nowhere, but we have the consolation that $\mathbb{C}[x]$ is complete. The machinery we develop here will help justify the notion of $\mathbb{C}$.

Definition 7.1. A field E is an extension field of a field F is $F \leq E$.

It is typically sufficient to use the following definition.

Proposition 7.2. A field E is isomorphic to an extension field of a field F if and only if there exists an injective homomorphism from F to E.

To show this, we can simply rename elements of the image of F in E to their corresponding counterpart in F in a one-to-one manner. An isomorphism really does mean that two fields (rings) are exactly the same. A complete proof is omitted. The idea is that an extension field E of F need not contain F as sets; it is sufficient to have an injective homomorphism (of rings/fields) from F to E. Given such an injection, we use it to rename elements of F to their images in E, thus creating $E \geq \tilde{F}$ where $\tilde{F} \simeq F$.

Theorem 7.3 (Kronecker)**.** *Let F be a field and suppose $f(x) \in F[x]$ is a non-constant polynomial. Then, there exists an extension field E of F such that $f(x)$ as a polynomial in $E[x]$ has a zero in E.*

Proof. Let $\phi : F \to F[x]/\langle f(x)\rangle$ be the composition of the canonical homomorphism $F[x] \to F[x]/\langle f(x)\rangle$ and the inclusion homomorphism $F \to F[x]$. We claim that ϕ is injective. Indeed, let $a, b \in F$ be such that $\phi(a) = \phi(b)$, or $a - b \in \langle f(x)\rangle$. Then, $a - b = f(x) \cdot g(x)$ for some $g(x) \in F[x]$. Then, $0 \geq \deg(a - b) = \deg f + \deg g$, where $\deg f \geq 1$. The only possibility is $\deg g = -\infty$, which implies $\deg(a - b) = \infty$ and $a = b$. Therefore, there exists an extension field E of F.

Let $f(x) = a_n x^n + \cdots + a_0$, where $n = \deg f \geq 1$. To show $f(x)$ as a polynomial in $E[x]$ has a zero in E, it is sufficient to prove that $\tilde{f}$ has a zero, where $\tilde{f} \in (F[x]/\langle f(x)\rangle)[\tilde{x}]$ is the polynomial whose coefficients are the coefficients of $f(x)$ mapped under ϕ; that is, $\tilde{f}(\tilde{x}) = \phi(a_n)\tilde{x}^n + \cdots + \phi(a_0)$. We claim that $\tilde{f}(\alpha) = 0$, where $\alpha = x + \langle f(x)\rangle \in F[x]/\langle f(x)\rangle$. Indeed,

$$\begin{aligned}
\tilde{f}(\alpha) = \sum_{i=0}^{n} \phi(a_i) \cdot \alpha^i &= \sum_{i=0}^{n} (a_i + \langle f(x)\rangle) \cdot (x + \langle f(x)\rangle)^i \\
&= \sum_{i=0}^{n} (a_i + \langle f(x)\rangle) \cdot (x^i + \langle f(x)\rangle) \\
&= \sum_{i=0}^{n} (a_i x^i + \langle f(x)\rangle) \\
&= (\sum_{i=0}^{n} a_i x^i) + \langle f(x)\rangle \\
&= f(x) + \langle f(x)\rangle \\
&= 0_{F[x]/\langle f(x)\rangle}.
\end{aligned}$$

The proof is complete. □

We also take a look at a way of representing field elements of $F[x]/\langle f(x)\rangle$. The division algorithm allows us to view this factor ring the same as we view $\mathbb{Z}/n\mathbb{Z}$, where the division algorithm allows us to represent each coset uniquely with the remainder.

To draw this parallel, we denote the equivalence relation $\alpha(x) \sim_{F[x]/\langle f(x)\rangle} \beta(x)$ as $\alpha \equiv \beta \pmod{f(x)}$. This relation is really a congruence: addition and multiplication produce the same result when both inputs are replaced by equivalent elements by the congruence.

Definition 7.4. Let F be a field and $f(x) \in F[x]$ irreducible. Two polynomials $g(x), h(x) \in F[x]$ are said to be congruent modulo $f(x)$ if $g(x) \sim_{F[x]/\langle f(x)\rangle} h(x)$.

Proposition 7.5. Let F be a field and $f(x) \in F[x]$ irreducible. Then, any element of $F[x]/\langle f(x)\rangle$ can be written uniquely as $r(x) + \langle f(x)\rangle$, where $r(x) \in F[x]$ has degree strictly less than $\deg f$.

Proof. Let $h(x) \in F[x]$ be arbitrary and suppose $h(x) + \langle f(x)\rangle = r(x) + \langle f(x)\rangle$, where $r(x) \in F[x]$ and $\deg r < \deg f$. Then, $h(x) - r(x) \in \langle f(x)\rangle$, so $h(x) = f(x) \cdot q(x) + r(x)$ for some $q(x) \in F[x]$. The uniqueness of the long division algorithm (Theorem 6.13) implies that the choice of $r(x)$ is unique, and one such choice is given explicitly by the algorithm. □

So, $\mathbb{C}$ is really just $\mathbb{R}[x]/\langle x^2 + 1\rangle$, where elements of $\mathbb{C}$, represented uniquely as $(ax + b) + \langle x^2 + 1\rangle$ for $a, b \in \mathbb{R}$, are really $ai + b$. Here, $i =$ "$1i + b$" $= x + \langle x^2 + 1\rangle$. This machinery equips us to dig deeper into number theory.

Definition 7.6. An element α of an extension field E of a field F is said to be algebraic over F if it is a zero of a non-zero polynomial in $F[x]$. Otherwise, α is said to be transcendental over F.

Definition 7.7. A real or complex number $x \in \mathbb{R} \cup \mathbb{C}$ is said to be an algebraic number if it is algebraic over $\mathbb{Q}$; otherwise, x is said to be a transcendental number.

A prototypical example is that $\sqrt{2}$, which is not in $\mathbb{Q}$ but is algebraic over $\mathbb{Q}$ because $x^2 - 2 = 0$. It is well-known, but not at all obvious, that both e and π are transcendental over $\mathbb{Q}$. But neither is transcendental over $\mathbb{R}$: e is a zero of the polynomial $(x - \mathrm{e}) \in \mathbb{R}[x]$ and π is a zero of the polynomial $(x - \pi) \in \mathbb{R}[x]$.

Corollary 7.8. A complex number is an algebraic number if and only if it is a zero of a polynomial over $\mathbb{Z}$.

Proof. The "if" direction is obvious. Conversely, given a zero z in $\mathbb{C}$ of a non-constant polynomial $a_0 + \cdots + a_n x^n$ $(n \geq 1)$ over $\mathbb{Q}$, multiply all coefficients by the least common multiple of the denominators of the coefficients to obtain a polynomial in $\mathbb{Z}$ with the same zeros to obtain a polynomial $ca_0 + \cdots + ca_n x^n \in \mathbb{Z}[x]$ with the same zeros. In particular, z remains a zero of the constructed polynomial. □

We will leverage the tool of homomorphisms heavily throughout our discussion of extension fields. The following is a prototypical example that highlights its utility.

Lemma 7.9. Suppose $\alpha \in E \geq F$ and let $\phi_\alpha \colon F[x] \to E$ be the evaluation homomorphism of $F[x]$ at α. Then, α is transcendental over F if and only if ϕ_α is injective.

Proof. Suppose for the "if" direction that ϕ_α is injective. Then, $\ker \phi_\alpha$ is trivial, so no polynomial in $F[x]$, other than the zero polynomial, evaluates to 0 at α. In particular, no non-constant polynomial in $F[x]$ evaluates to 0 at α. By definition, therefore, α is transcendental over F.

Suppose for the other direction that α is transcendental. Consider the kernel of ϕ_α, which necessarily contains 0. For any other $f(x) \in F[x] \backslash \{0\}$, there are two possibilities: **(a)** if $\phi_\alpha(f(x)) = f(\alpha)$ is a (non-zero) constant, then it never evaluates to 0; **(b)** if $f(\alpha)$ is non-constant, then because α is transcendental, $f(\alpha) \neq 0$. Therefore, $\ker \phi_\alpha$ is trivial, which implies that ϕ_α is injective. □

Consider $\alpha \in E \geq F$ algebraic over F. Of the many polynomials in $F[x]$, any of which has α as a zero, can we find a particular representative for α? The following theorem replies affirmatively by choosing a monic, irreducible representative.

Theorem 7.10. *Let $\alpha \in E \geq F$ be algebraic over F. Then, there exists a unique monic, irreducible polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$.*

Proof. Let $I \coloneqq \{f(x) \in F[x] \mid f(\alpha) = 0\}$, which we claim is an ideal of $F[x]$. One can verify that it is a subgroup under addition by checking closure under addition and additive inverses, and that it absorbs multiplication by ring elements on both sides.

Then, there exists a generator $\tilde{p}(x) \in F[x]$ of I. We now argue that $\tilde{p}(x)$ is non-constant. Suppose the contrary, so either $I = 0$ when $\tilde{p}(x) = 0$, or $I = R$ when $\tilde{p}(x) \in F\backslash\{0\}$. If $I = 0$, then α is by definition transcendental, a contradiction. The other case that $I = R$ is absurd by inspecting its definition, since polynomials of degree 1 (i.e., non-zero constants) cannot have any zeros, including α, another contradiction. Because $\tilde{p}(x)$ is non-constant, all possible $\tilde{p}(x)$ are equivalent up to multiplication by a non-zero constant. Therefore, denoting $\tilde{p}(x) = \tilde{a}_0 + \cdots + \tilde{a}_n x^n$ where $\tilde{a}_n \neq 0$ and $n \geq 1$, the monic generator $p(x) \coloneqq \tilde{a}_n^{-1} \cdot \tilde{p}(x)$ of I must be unique.

It remains to show that $p(x)$ is irreducible. Indeed, let $p(x) = f(x) \cdot g(x)$ where $f(x), g(x) \in F[x]$. Then, $p(\alpha) = 0$ implies that either $f(\alpha) = 0$ or $g(\alpha) = 0$. Suppose $f(\alpha) = 0$, or $f(x) \in I = \langle p(x) \rangle$, without loss of generality. By similar logic as above, f is a non-constant polynomial as well. Then, $f(x) = p(x) \cdot g'(x)$ for some $g'(x) \in F[x]$. Combining the two equations, we have $f(x) = f(x) \cdot g(x) \cdot g'(x)$, or $f(x) \cdot (1 - g(x) \cdot g'(x)) = 0$. Because $f(x) \neq 0$, $g(x) \cdot g'(x) = 1$, so $g(x)$ must be a (non-zero) constant. The proof is now complete. □

Definition 7.11. Let $\alpha \in E \geq F$ be algebraic over F. The unique monic, irreducible polynomial $p(x) \in F[x]$ which generates the ideal of polynomials in $F[x]$ having α as a zero, according to Theorem 7.10, is denoted as $\underline{\text{irr}(\alpha, F)}$ and called the minimal polynomial for α over F. The degree of $\text{irr}(\alpha, F)$, called the degree of α over F, is denoted as $\underline{\deg(\alpha, F)}$.

It's noteworthy to remark that there is no apparent algorithm that identifies the minimal polynomial. Consider the following example:

Let $\alpha \coloneqq \sqrt{1 + \sqrt{3}} \in \mathbb{R}$, which is an algebraic number. Then, $\alpha^2 = 1 + \sqrt{3}$, so $(\alpha^2 - 1)^2 = 3$. In other words, $\alpha^4 - 2\alpha^2 - 2 = 0$. By the Eisenstein criterion (Proposition 6.25) with $p = 2$, we see that $f(x) \coloneqq x^4 - 2x^2 - 2 \in \mathbb{Q}[x]$ is a monic, irreducible polynomial. The uniqueness of the minimal polynomial therefore implies $\text{irr}(\alpha, \mathbb{Q}) = x^4 - 2x^2 - 2$.

All previous discussions on extension fields concern a given extension field E of F. We now build towards to other direction: given a field F and some "element" with some desired properties, can we *create* an extension field E of F that's just large enough to include the given element and, of course, the entirety of F?

Let's first narrow down the elements of consideration. We can certainly describe α as a (non-existent) zero of a polynomial in $F[x]$, where we can manually construct $F[\alpha]$ already. Otherwise, we should assume complementarily that α is transcendental over F (although we do not know in general which field α is in).

To this extent, we consider what we shall call simple extensions of a field F, where the desired "element" comes from some larger given field E which must exist.

Case I. Suppose $\alpha \in E \geq F$ is algebraic over F. Then, $F[x]/\langle \text{irr}(\alpha, F) \rangle$ is isomorphic to an extension field of F. We define $F(\alpha)$ as the image of the injective homomorphism $F[x]/\langle \text{irr}(\alpha, F) \rangle \hookrightarrow E$ which takes a coset to the evaluation of its minimal degree representative at α, satisfying $F \leq F(\alpha) \leq E$. Note that any member of such a coset evaluates to the same value in E at $\alpha \in E$. $F(\alpha)$ is the result of an effort to create an extension field of F just large enough to include α as well.

Case II. Suppose instead that $\alpha \in E \geq F$ is transcendental over F. By Lemma 7.9, the evaluation homomorphism $\phi_\alpha \colon F[x] \to E$ is injective. In other words, every polynomial in $F[x]$ evaluates to a different number in E at α. These evaluations must be included (after mapping by some injective homomorphism) in the extension field by field axioms, and including them gives us a domain $F[\alpha]$. This notation is understood as the ring of polynomials over the indeterminate α, which is a different mathematical object from the field element $\alpha \in E$; however, every element of the domain is identified

naturally with a number in E via evaluation at $\alpha \in E$. In this case, we would have to expand the extension field to the field of fractions of $F[x]$, which we denote as $F(\alpha)$, whose notation is understood in the same way as $F[\alpha]$. Note that $F \hookrightarrow F(\alpha) \hookrightarrow E$ in a natural way,[3] so we may redefine $F(\alpha)$ as its image in E.

Note that $F(\alpha)$ is by definition dependent on $E \ni \alpha$. By renaming, there are infinitely many other extension fields $\tilde{F}(\alpha)$ of F isomorphic to $F(\alpha)$. This observation highlights why we often say "F is a subfield of E" when we really mean "F is isomorphic to a subfield of E," or equivalently "there exists an injective homomorphism from F to E." However, we will refrain from indulging in these shorthands for the sake of rigorous logic. Regardless, given a specific $E \ni \alpha$, $F(\alpha)$ has a unique definition algorithmically.

Definition 7.12. Let $\alpha \in E \geq F$. The simple extension of F by α, denoted as $F(\alpha)$, is defined as follows:

- When α is algebraic over F, $F(\alpha)$ is an extension field of F in E isomorphic to $F[x]/\langle \mathrm{irr}(\alpha, F)\rangle$;
- When α is transcendental over F, $F(\alpha)$ is an extension field of F in E isomorphic to $F(x)$.

That such extension fields are unique in E is not immediately straightforward, since two isomorphic subfields need not equal each other as sets (e.g., $\mathbb{Q}(\pi)$ is countable, so there exists some transcendental $z \notin \mathbb{Q}(\pi)$ so that $\mathbb{Q}(\pi) \simeq \mathbb{Q}(z)$ but clearly differ as sets. We provide another perspective of defining simple extensions which will achieve this purpose.

Definition 7.13. An extension field E of F is said to be a simple extension if $E = F(\alpha)$ for some $\alpha \in E$.

Note that we could equivalently weaken the definition above to $E \simeq F(\alpha)$ instead of $E = F(\alpha)$.

We first state a corollary of Proposition 7.5 in the context of simple extensions.

Corollary 7.14. Let $E = F(\alpha)$ be a simple extension of F, where $\alpha \in E$ is algebraic over F. Then, every element x of E can be written uniquely as a tuple $(a_0, \cdots, a_{n-1}) \in F^n$ such that $x = a_0 + \cdots + a_{n-1}\alpha^{n-1}$, where $n = \deg(\alpha, F)$.

We also include a useful and straightforward lemma.

Lemma 7.15. Let $\alpha \in E \geq F$ and $\beta \in F(\alpha)$. Then, $F(\beta) \leq F(\alpha)$.

Proof. Let $z \in F(\beta)$ and $\beta = c_0 + \cdots + c_k\alpha^k$, where $c_0, \cdots, c_k \in F$. By definition, there exist polynomials $f(x), g(x) \in F[x]$, where $g(x) \neq 0$, such that $z = f(\beta)/g(\beta)$. Let $f(x) = a_0 + \cdots + a_m x^m$ and $g(x) = b_0 + \cdots + b_n x^n$. Then,

$$z = \frac{a_0 + \cdots + a_m \cdot (c_0 + \cdots + c_k\alpha^k)^m}{b_0 + \cdots + b_n \cdot (c_0 + \cdots + c_k\alpha^k)^n},$$

in which $c_0, \cdots, c_k, b_0, \cdots, b_n, c_0, \cdots, c_k \in F \leq F(\alpha)$ and $\alpha \in F(\alpha)$. Therefore, applying all operations in $F(\alpha)$ naturally, we conclude that $z \in F(\alpha)$ by field axioms. □

Now, we leverage our knowledge about vector spaces here, demonstrating an elegant application of linear algebra.

Proposition 7.16. Let $\alpha \in E \geq F$ be algebraic over F. Then, $F(\alpha)$ is an n-dimensional vector space over F that admits the basis $\{1, \cdots, \alpha^{n-1}\}$, where $n = \deg(\alpha, F)$. Further, every element $\beta \in F(\alpha)$ is algebraic over F with $\deg(\beta, F) \leq \deg(\alpha, F)$.

Proof. First, $F(\alpha)$ is by definition a vector space over F, which inherits addition and scalar multiplication naturally from field operations. We now show that $\{1, \cdots, \alpha^{n-1}\}$ is a basis. Indeed, by Corollary 7.14 above, every $v \in F(\alpha)$ can be written uniquely as $v = c_0 \cdot 1 + \cdots + c_{n-1} \cdot \alpha^{n-1}$, where $c_0, \cdots, c_{n-1} \in F$.

For every $\beta \in F(\alpha)$, $F \leq F(\beta) \leq F(\alpha)$ by Lemma 7.15. Because $F(\beta)$ is a vector space over F with the same operations, and $F(\beta) \subseteq F(\alpha)$, $F(\beta)$ is a subspace of $F(\alpha)$, which then has dimension at most n. Applying the argument from the previous paragraph to $F(\beta)$, we see that $\deg(\beta, F) \leq \deg(\alpha, F)$. □

[3] The hooked right arrow $\hookrightarrow$ denotes an injective homomorphism here.

7.2 Algebraic Extensions

The second half of Proposition 7.16 tells us that if $\alpha \in E \geq F$ is algebraic over F and $\beta \in F(\alpha)$, then β is algebraic over F. In other words, $F(\alpha)$ is such an extension of F that every element of $F(\alpha)$ is algebraic over F. We therefore generalize simple extensions in the following way, by considering the larger family of algebraic extensions of a field.

Definition 7.17. An extension field $E \geq F$ is said to be an algebraic extension of F if every element of E is algebraic over F.

We also leverage the notion of dimensionality to define *finite extensions*.

Definition 7.18. Suppose $E \geq F$ is a finite-dimensional vector space over F. Then, E is said to be a finite extension of degree n over F, where n is the dimension of E over F, or simply a finite extension of F.

Definition 7.19. Given a field extension $E \geq F$, the degree of E over F, denoted as $[E : F]$, is defined as the degree of E over F.

We begin our analysis by considering the special case of $[E : F] = 1$.

Proposition 7.20. Suppose $E \geq F$. Then, $[E : F] = 1$ if and only if $E = F$.

Proof. Suppose $[E : F] = 1$ and fix a basis $\{v\}$ of E over F, where $v \in E$. Then, $E = \{c \cdot v \mid c \in F\}$. In particular, $1 = c_1 \cdot v$ for some $c_1 \in F$, so $v = c_1^{-1} \in F$ as well; further, v is a unit in F. Therefore, $E = vF = \langle v \rangle_F = \langle 1 \rangle_F = F$. □

It turns out that not only simple extensions, but all finite extensions, are algebraic extensions.

Theorem 7.21. *A finite extension $E \geq F$ is an algebraic extension of F.*

Proof. Denote n as the dimension of E over F and suppose $\alpha \in E$. Then, the list of $n+1$ vectors $1, \cdots, a^n$ cannot be linearly independent. Therefore, there exist scalars $c_0, \cdots, c_n \in F$, not all zeros, such that

$$c_0 + \cdots + c_n \alpha^n = 0;$$

in other words, there exists a non-zero polynomial $f(x) \in F[x] \backslash \{0\}$ such that $f(\alpha) = 0$. Therefore, α is algebraic over F. □

We have as a corollary a useful criterion for algebraicity.

Corollary 7.22. Let $\alpha \in E \geq F$. Then, α is algebraic over F if and only if $[F(\alpha) : F]$ is finite; that is, $F(\alpha)$ is a finite extension of F.

Proof. The $\Rightarrow$ direction is the first part of Proposition 7.16. For the $\Leftarrow$ direction, $F(\alpha)$ is an algebraic extension of F by Theorem 7.21 above, so $\alpha \in F(\alpha)$ in particular is algebraic over F. □

The notation $[E : F]$ may be reminiscent of the index $[G : H]$ of a subgroup $H \leq G$, which comes with a multiplication equation. Similarly, we have the following result concerning the degree of finite extensions.

Proposition 7.23. Suppose $F \leq E \leq K$, where E is a finite extension over F and K a finite extension over E. Then, K is a finite extension over F, and $[K : F] = [K : E][E : F]$.

Proof. Suppose $\{e_1, \cdots, e_n\} \in E$ is a basis of E over F and $\{k_1, \cdots, k_m\} \in K$ a basis of K over E. Then, for every $k \in K$, there exist unique scalars $b_1, \cdots, b_m \in E$ such that $k = \sum_{j=1}^{m} b_j k_j$. Subsequently, each b_j is a linear combination $\sum_{i=1}^{n} a_{ij} e_i$, where $a_{ij} \in F$, so

$$k = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} a_{ij} e_i \right) k_j = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \cdot e_i k_j.$$

Therefore, $\{e_i \cdot k_j \mid i = 1, \cdots, n, \quad j = 1, \cdots, m\}$ spans K over F.

It remains to show linear independence. Suppose constants $\{c_{ij} \mid i = 1, \cdots, n \quad j = 1, \cdots, m\}$ are such that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} e_i k_j = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} c_{ij} e_i \right) \cdot k_j = 0.$$

Because $\{k_j\}$ is linearly independent, $\sum_i c_{ij} e_i = 0$ for all j. Because $\{e_i\}$ is linearly independent, $c_{ij} = 0$ for all i and all j. The proof is now complete. □

Corollary 7.24. Suppose $F_0 \leq \cdots \leq F_n$, where F_i is a finite extension of F_{i-1} for each $i = 1, \cdots, n$. Then, F_n is a finite extension of F_0, and

$$[F_n : F_0] = [F_n : F_{n-1}] \cdots [F_1 : F_0].$$

The corollary above is a trivial extension of Proposition 7.23 by induction.

Corollary 7.25. Let $\alpha \in E \geq F$ be algebraic over F and suppose $\beta \in F(\alpha)$. Then, $\deg(\beta, F)$ divides $\deg(\alpha, F)$.

Proof. Let $m = [F(\alpha) : F] < +\infty$, $k = [F(\alpha) : F(\beta)]$ by Lemma 7.15, and $n = [F(\beta) : F] < +\infty$. We first show that k is finite. Let $\alpha_0, \cdots, \alpha_m \in F(\alpha)$ be linearly independent over $F(\beta)$. In particular, they are a list of $m+1$ linearly independent vectors over $F \leq F(\beta)$. But this is impossible, since the dimension of $F(\alpha)$ over F is m. Therefore, $k \leq m < +\infty$.

Now, because $F \leq F(\beta) \leq F(\alpha)$ is a chain of consecutive finite extensions, we have $[F(\alpha) : F] = [F(\alpha)$ □

Corollary 7.25 is extremely powerful in computational problems. By the Corollary, $x^3 - 2$ has no zeros in $\mathbb{Q}(\sqrt{2})$: $x^3 - 2$ is irreducible by Eisenstein with $p = 2$, so a zero would have degree 3 which does not divide $2 = \deg(\sqrt{2}, \mathbb{Q})$.

To characterize finite extension fields, we'll first look back at simple extensions. We show that a simple extension $F(\alpha) \geq F$ for some $\alpha \in E \geq F$ really is the smallest extension field of F in E containing α.

Definition 7.26. Let $(S, \leq)$ be a partially ordered set. An element $s \in S$ is said to be the smallest element of S if $s \leq a$ for any $a \in S$. An element $s \in S$ is said to be a minimal element if no element $a \in S$ is strictly smaller than s: for all $a \in S$, $a = s$ whenever $a \leq s$.

Corollary 7.27. There is a unique minimal element of a partially ordered set if and only if there exists smallest element of the set. If this is true, then the minimal element and the smallest element coincide.

A partially ordered set (poset) $(S, \leq)$ is a directed, acyclic, transitive graph with loops on all vertices, whose edges are precisely $\leq = \{(a, b) \mid a, b \in S \quad \text{and} \quad a \leq b\}$. A minimal element is a vertex with in-degree 0. In a partially ordered set, neither a minimal element nor a smallest element is guaranteed to exist (e.g., $(\mathbb{Z}, \leq)$). A minimal element, when it exists, need not be unique (e.g., all primes are minimal in $(\mathbb{Z}_{\geq 2}, |)$ under the divisibility relation). However, when the smallest element exists, it is necessarily unique by a simple symmetry argument. When S is a collection of subsets of a common set, there is a constructive way to find the smallest element.

Corollary 7.28. Let $\mathcal{S}$ be a family of subsets of a set X, partially ordered by inclusion. Then, the smallest element of $\mathcal{S}$ exists if and only if the intersection $\bigcap_{T \in \mathcal{S}} T$ is an element of $\mathcal{S}$. Further, if this is true, then the smallest element of $\mathcal{S}$ is $\bigcap_{T \in \mathcal{S}} T$.

Proof. Let $T^* := \bigcap_{T \in \mathcal{S}} T \subseteq X$. If $T^* \in \mathcal{S}$, for every $T \in \mathcal{S}$, T^* is the intersection of some sets including T. Therefore, $T^* \subseteq T$, and T^* is the smallest element of $\mathcal{S}$. If instead $T^* \notin \mathcal{S}$, but assuming for contradiction that the smallest element T_0 of $\mathcal{S}$ exists, then $T^* \subseteq T_0$ and $T_0 \subseteq T^* = \cap_{T \in \mathcal{S}} T$, both by definition. Therefore, $T_0 = T^* \notin \mathcal{S}$, a contradiction.

It remains to note that by definition, the smallest element is the intersection of all subsets of X from S when it exists. □

Theorem 7.29. *Let $\alpha \in E \geq F$. Then, $F(\alpha)$ is the smallest extension field of F in E that contains α; that is, $F(\alpha)$ is the smallest element in $(\{F' \leq E \mid F' \geq F \quad \text{and} \quad \alpha \in F'\}, \leq)$.*

Proof. If α is algebraic over F, then $F(\alpha) \simeq F[x]/\langle \text{irr}(\alpha, F)\rangle$. Every element $z \in F(\alpha)$ is uniquely represented as a polynomial in α over F of degree strictly less than $\text{irr}(\alpha, F)$. Every such polynomial must evaluate to an element that lies in any extension field of F in E containing α by field axioms.

If instead α is transcendental over F, then every element $z \in F(\alpha)$ is uniquely represented as a rational function in α over F. Similarly by field axioms, z is an element that lies in any extension field of F in E containing α. □

We can now generalize simple extensions naturally to multiple numbers from E.

Definition 7.30. Let $\alpha_1, \cdots, \alpha_n \in E \geq F$. The field $F(\alpha_1) \cdots (\alpha_n)$, denoted as $F(\alpha_1, \cdots, \alpha_n)$, is defined as the field adjoining to F the elements $\alpha_1, \cdots, \alpha_n$.

The parentheses above are by definition left-associative. In other words, we first extend F to $F(\alpha_1)$. Then, we extend $F(\alpha_1)$ by α_2, and so on, until we finally extend $F(\alpha_1) \cdots (\alpha_{n-1})$ by α_n. Now, extending the equivalent definition for simple extensions, we have the following equivalence:

Theorem 7.31. *Let $\alpha_1, \cdots, \alpha_n \in E \geq F$. Then, $F(\alpha_1) \cdots (\alpha_n)$ is the smallest extension field of F in E containing $\alpha_1, \cdots, \alpha_n$; that is, $F(\alpha_1) \cdots (\alpha_n)$ is the smallest element of $(\{F' \leq E \mid F' \geq F \mid \alpha_1, \cdots, \alpha_n \in F'\}, \leq)$.*

Proof. We first show that $\min\{F' \leq E \mid F' \geq \min\{F'' \leq E \mid F'' \geq F \wedge \alpha_1 \in F''\} \wedge \alpha_2 \in F'\} = \min\{F' \leq E \mid F' \in F \wedge \alpha_1 \in F' \wedge \alpha_2 \in F'\}$. It is sufficient to show that $F' \geq \min\{F'' \leq E \mid F'' \geq F \wedge \alpha_1 \in F''\} = F(\alpha_1)$ is equivalent to $F' \geq F \wedge \alpha_1 \in F'$. For the $\Rightarrow$ direction, note that $F' \geq F(\alpha_1) \supseteq F \cup \{\alpha_1\}$, so $F' \geq F$ and $\alpha_1 \in F'$ obviously. Conversely for the $\Leftarrow$ direction, simply note that F' is a valid choice of F'', so $F' \geq F(\alpha_1)$ by definition.

In other words, we have shown that $F(\alpha_1)(\alpha_2) = F(\alpha_1, \alpha_2)$ is the smallest extension field of F in E that contains α_1, α_n. By induction, $F(\alpha_1) \cdots (\alpha_n)$ is the smallest extension field of F in E that contains $\alpha_1, \cdots, \alpha_n$. □

Because the logical AND $\wedge$ is commutative, we have the following corollary:

Corollary 7.32. Let $\alpha_1, \cdots, \alpha_n \in E \geq F$. Then, $F(\alpha_1) \cdots (\alpha_n) = F(\alpha_{\sigma(1)}) \cdots (\alpha_{\sigma(n)})$ for any permutation $\sigma \in S_n$.

Proof. Observe that $F(\alpha_1) \cdots (\alpha_n)$ and $F(\alpha_{\sigma(1)}) \cdots (\alpha_{\sigma(n)})$ are both the smallest extension field of F in E that contains $\alpha_1, \cdots, \alpha_n$. □

Similarly, the rational functions over a field in n indeterminates satisfy $F(x_1, \cdots, x_n) \simeq F(x_1) \cdots (x_n)$, although a complete treatment is omitted for brevity.

We can now show the equivalence of finite extensions and extensions of adjoining n variables.

Proposition 7.33. Let $E \geq F$. Then, E is a finite extension of F if and only $E = F(\alpha_1, \cdots, \alpha_n)$ for some $\alpha_1, \cdots, \alpha_n \in E$ algebraic over F and some $n \in \mathbb{Z}_{\geq 0}$.

Proof. Suppose $E = F(\alpha_1, \cdots, \alpha_n)$. Each $\alpha_1, \cdots, \alpha_n \in E$ is algebraic over F, so it is algebraic over any extension field of F in E. In particular, for the chain of extensions

$$F \leq F(\alpha_1) \leq \cdots \leq F(\alpha_1, \cdots, \alpha_n) = E, \tag{1}$$

α_{i+1} is algebraic over $F(\alpha_1, \cdots, \alpha_i)$ for all $i \in \{1, \cdots, n-1\}$, so every extension from Chain 1 is a simple extension by an algebraic element and hence a finite extension. Applying Proposition 7.24, $E = F(\alpha_1, \cdots, \alpha_n)$ is a finite extension of F.

Suppose instead that E is a finite extension of F, which is then an algebraic extension by Theorem 7.21. Let $\alpha_1, \cdots, \alpha_n$ be a basis of E in F, all of which are algebraic over F. Then, $F(\alpha_1, \cdots, \alpha_n) \leq E$. To show inclusion in the other direction, note that every $e \in E$ is expressible uniquely as $e = c_1\alpha_1 + \cdots + c_n\alpha_n$, where $c_1, \cdots, c_n \in F$. By field axioms, every e is a member of $F(\alpha_1, \cdots, \alpha_n)$. The proof is complete. □

The proposition above allows us to conclude that the algebraic extension is a transitive relation.

Corollary 7.34. Suppose $F' \geq F$ and $F'' \geq F'$ are algebraic extensions. Then, F'' is an algebraic extension of F.

Proof. Let $\alpha \in F''$, which is algebraic over F'. Thus, $f(\alpha) = 0$ for some non-constant $f(x) = a_0 + \cdots + a_n x^n \in F'[x] \backslash F'$. Because $a_0, \cdots, a_n \in F'$ are algebraic over F, $F(a_0, \cdots, a_n)$ is a finite extension of F by Proposition 7.33. Further, α is a zero of $f(x) \in F(a_0, \cdots, a_n)[x] \leq F'[x]$, so α is algebraic over $F(a_0, \cdots, a_n)$ and $[F(a_0, \cdots, a_n, \alpha) : F(a_0, \cdots, a_n)]$ is finite by the same Proposition. Applying 7.23, we note that $F(a_0, \cdots, a_n, \alpha)$ is a finite extension, and hence an algebraic extension, of F. Therefore, $\alpha \in F(a_0, \cdots, a_n, \alpha)$ is algebraic over F. □

Lastly, we remark that the algebraic elements of a field forms a field.

Proposition 7.35. Let $E \geq F$. Then, the set $\{\alpha \in E \mid \alpha \text{ is algebraic over } F\}$ is an extension field of F in E.

Proof. It is sufficient to show that $\tilde{F}_E := \{\alpha \in E \mid \alpha \text{ is algebraic over } F\}$ satisfies field closure axioms. Given $\alpha, \beta \in \tilde{F}_E$ which are both algebraic over F, $F(\alpha, \beta)$ is a finite (Proposition 7.33) and hence algebraic (Theorem 7.21) extension of F. Therefore, every member of $F(\alpha, \beta)$, which includes $\alpha + \beta, -\alpha, \alpha \cdot \beta$ always and includes α^{-1} when $\alpha \neq 0$, is algebraic over F. □

We now proceed to discuss algebraically closed field, which establishes the importance of $\mathbb{C}$ relative to $\mathbb{Q}$.

Definition 7.36. A field F is said to be algebraically closed if every non-constant polynomial over F has a zero in F.

Similar to the case in $\mathbb{C}$, we can split any such polynomial into linear factors.

Proposition 7.37. A field F is algebraically closed if and only if every non-constant polynomial over F factors into linear factors.

Proof. Suppose F is algebraically closed and let $f(x) \in F[x]$ be a non-constant polynomial. Denote $n = \deg f$. Let z_1 be a zero of $f_0(x) := f(x)$. Then, $f_0(x) =: (x - z_1) \cdot f_1(x)$, where $\deg f_1(x) = n - 1$. Continuing until $f_0(x) = (x - z_1) \cdots (x - z_n) \cdot f_n(x)$, where $\deg f_n = n - n = 0$ means f_n is a constant, which we absorb into, say, $(x - z_1)$.

Suppose conversely that every non-constant polynomial $f(x)$ over F factors into linear factors. Suppose $ax + b$ is one such factor. Then, $-a/b$ is a zero of $f(x)$. □

Corollary 7.38. An algebraically closed field has no proper algebraic extension. That is, no proper extension field $E \gneq F$ of F is an algebraic extension.

Proof. Suppose that E is a algebraic extension of F, which is algebraically closed. Let $e \in E$, which is algebraic over F. Therefore, e the zero of some non-constant polynomial over F, which must be in F. Therefore, $F \leq E \leq F$, and $E = F$. □

Notably, the complex numbers form an algebraically closed field. In other words, $\mathbb{C}$ already contains all possible zeros of polynomials; there are no more to add.

Lastly, we show that every field F has an algebraic closure, so that the extension $\mathbb{Q} \leq \mathbb{C}$ can be generalized:

Definition 7.39. An algebraic closure of a field F is an algebraic extension of F that is algebraically closed.

That every field has an algebraic closure depends on the Axiom of Choice, which we state as follows.

Axiom 7.1 (Axiom of Choice)**.** *For every set S, there exists a choice function $f\colon 2^S\backslash\{\varnothing\} \to S$ such that $f(X) \in X$ for all non-empty $X \subseteq S$.*

In other words, f chooses an element from X for every non-empty subset X from some given set. This formulation implies (and is in fact equivalent to) Zorn's lemma. We first define some basic terminology.

Definition 7.40. Let $(S, \leq)$ be a partially ordered set. A chain C in S is a totally ordered subset of S; that is, $a \leq b$ or $b \leq a$ for all $a, b \in C$.

Recall that a maximal element of a poset is one such that no element is larger. This is the statement of Zorn's lemma:

Lemma 7.41 (Zorn)**.** Let $(S, \leq)$ be a partially ordered set. If every chain C in S has an upper bound u in S; that is, $u \geq c$ for all $c \in C$, then S has a maximal element.

It turns out Zorn's lemma is equivalent to the Axiom of Choice. Another such statement is known as Tarski's theorem:

Theorem 7.42 (Tarski)**.** *Under ZF, let A be an infinite set. Then, $|A \times A| = |A|$ if and only if the Axiom of Choice holds.*

Lastly, we shall require some machinery to deal with set theory, independent of the Axiom of Choice.

Theorem 7.43. *For every set S, the power set is strictly larger than S; that is, $|2^S| > |S|$.*

Proof. Suppose on the contrary that $f\colon S \to 2^S$ is a surjection, and consider the set $X = \{x \in S \mid x \notin f(x)\} \in 2^S$. Because f is surjective, there exists $x_0 \in S$ such that $f(x_0) = X$. If $x \in f(x) = X$, then $x \notin f(x)$; if $x \notin f(x)$, then $\neg(x \notin f(x))$, so $x \in f(x)$. This is a contradiction, so f cannot be surjective. □

The following statement allows us to show that the power set of a set A is "sufficiently larger" than A.

Corollary 7.44. Let A be an infinite set. Then, $|A| \leq |\{S \subseteq A : |S| = 2\}|$.

Proof. Let $a_0 \in A$. We first remark that $|A\backslash\{a_0\}| = |A|$. Let $\tilde{S} \subseteq A$ be countably infinite, so that $S \coloneqq \tilde{S} \cup \{a_0\}$ remains countably infinite. Fix a bijection $f\colon \mathbb{Z}_{\geq 1} \to S\backslash\{a_0\}$ and define a bijection $g\colon \mathbb{Z}_{\geq 1} \to S$ via $g(1) \coloneqq a_0$ and $g(i-1) \coloneqq f(i)$ for all $i \in \mathbb{Z}_{\geq 2}$. Therefore, $|S| = |\mathbb{Z}_{\geq 1}| = |S\backslash\{a_0\}|$. Let $h\colon S \to S\backslash\{a_0\}$ be a bijection. Then, define a bijection $\phi\colon A \to A\backslash\{a_0\}$ via

$$\phi(x) \coloneqq \begin{cases} h(x), & \text{if } x \in S, \\ x, & \text{otherwise,} \end{cases}$$

which proves our remark. It only remains to note that

$$|A| = |A\backslash\{a_0\}| = |\{\{a_0, a\} \mid a \in A\backslash\{a_0\}\}| \leq |\{S \subseteq A : |S| = 2\}|.$$

The proof is complete. □

We now state and prove the theorem of interest:

Theorem 7.45. *Every field has an algebraic closure.*

Proof. Let F be a field and construct a set $A = \bigcup_{f \in F[x]} \{\omega_{f,i} \mid i = 1, \cdots, \deg f\}$.[4] Let $\Omega = 2^A \cup F$, which has $|\Omega| > |A|$.

Let $S \coloneqq \{E \subseteq \Omega \mid E \text{ is an algebraic extension of } F\}$. In more precise terms,

$$S = \{E \subseteq \Omega \mid \text{there exist binary operations } +, \cdot\colon E \times E \to E \text{ such that } (E, +, \cdot) \text{ is an algebraic extension of } F\}.$$

[4] The specific construction of $\omega_{f,i}$ does not matter; it only matters that A has more elements than all potential zeros $\omega_{f,i}$ for every polynomial. One can take $\omega_{f,i}$ as the ordered pair (f, i) for completeness.

We shall index $S = \{E_j \mid j \in J\}$ via an index set J.[5] Suppose $C = \{E_{j_k} \mid k \in K\}$ is a chain in S and let $U \coloneqq \bigcup_{k\in K} E_{j_k}$, which we make into a field by the following operations. Let $\alpha, \beta \in U$ with $\alpha \in E_{j_k}$ and $\beta \in E_{j_{k'}}$, where $E_{j_k} \leq E_{j_{k'}}$ without loss of generality. Define $\alpha +_U \beta$ as the sum of $\alpha, \beta \in E_{j_{k'}}$ and $\alpha \cdot_U \beta$ as the product of $\alpha, \beta \in E_{j_{k'}}$. Note that the choice of j_k and $j_{k'}$ is irrelevant here: if $E_{j_k} \leq E_{j_{k'}} \leq E_{j_{k''}}$, then the addition $a + b$ (resp. the multiplication $a \cdot b$) produces the same result whether taking place in $E_{j_{k'}}$ or $E_{j_{k''}}$. It remains to show that U is an algebraic extension of F. Indeed, for any $a \in U$, by definition $a \in E_{j_k}$ for some $j_k \in J$. Then, a is algebraic over F because $E_{j_k} \in S$ is an algebraic extension of F.

Therefore, by Zorn's Lemma (Lemma 7.41), there exists a maximal element $\overline{F}$ of S; that is, no element of S properly contains $\overline{F}$. Suppose for contradiction that $\overline{F}$ is not algebraically closed, so there exists a non-constant polynomial $f(x) \in \overline{F}[x]\backslash\overline{F}$ which admits no zeros in $\overline{F}$. We assume $f(x)$ to be irreducible without loss of generality.[6] Then $\overline{F}' \coloneqq \overline{F}[x]/\langle f(x)\rangle$ is an algebraic extension of $\overline{F}$ and $\overline{F}$ is an algebraic extension of F. Hence, by Corollary 7.34, $\overline{F}'$ is algebraic extension of F.

Observe that $\overline{F}' \leq \bigcup_{f(x)\in F[x]/\langle F\rangle}\{\alpha \in E \mid f(\alpha) = 0\}$ because $\overline{F}'$ is an algebraic extension of F. Because for every such $f(x)$, $|\{\alpha \in E \mid f(\alpha) = 0\}| \leq \deg f = |\{\omega_{f,i} \in A \mid i \in \{1, \cdots, \deg f\}\}|$, and such an association always maps different $f(x)$ to different ω_{f_i}, we conclude that $\overline{F}' \leq A$. Therefore, $|\overline{F}'\backslash\overline{F}| \leq |\overline{F}'| \leq |A| \leq |\{S \subseteq A : |S| = 2\}| \leq |2^A\backslash\tilde{A}| \leq |2^A\backslash\overline{F}| \leq |(2^A \cup F)\backslash\overline{F}| = |\Omega\backslash\overline{F}|$, where $\tilde{A} = \{\{a\} \mid a \in A\}$. The penultimate inequality follows from $|\overline{F}| \leq |\tilde{A}| = |A|$ and the last inequality follows from $F \leq \overline{F}$. We are hence given an injection $\psi\colon \overline{F}'\backslash\overline{F} \hookrightarrow \Omega\backslash\overline{F}$.

Finally, we construct an injection $\phi\colon \overline{F}' \hookrightarrow \Omega$ which fixes $\overline{F}$; that is, $\phi|_{\overline{F}} = \iota$ is an inclusion map:

$$\phi(x) \coloneqq \begin{cases} x, & \text{if } x \in \overline{F}, \\ \psi(x), & \text{otherwise.} \end{cases}$$

By ϕ we may therefore rename $\overline{F}'$ to $\phi[\overline{F}'] \leq \Omega$, an algebraic extension of F in Ω which properly contains $\overline{F}$. This contradicts the maximality of $\overline{F}$, which shows that $\overline{F}$ must be algebraically closed. □

8 Geometric Constructions

We shall first try to "define" constructible numbers informally. Then, we deduce some analytic properties that characterize them to proceed further algebraically.

Definition 8.1. A real number α is said to be constructible if a line segment of length $|\alpha|$ can be constructed in finitely many steps from a given line segment of length 1 from a straightedge and a compass.

Importantly, the constructible numbers form a field. The following proof should be taken with a grain of salt given the imprecise nature of the definition above.

Proposition 8.2. The set of constructible numbers is a subfield of $\mathbb{R}$.

Proof. Let A and B be the endpoints of the given unit interval and suppose α and β are constructible. The construction of $\alpha \pm \beta$ is straightforward. For $\alpha \cdot \beta$, fix a point O and fix A, P such that $|OA| = |\alpha|$ and $|OP| = 1$. Draw B on ray $\overrightarrow{OP}$ ($\overrightarrow{OP} \cdot \overrightarrow{OB} \geq 0$) such that $|OB| = |\beta|$. Draw draw line BQ parallel to PA such that Q lies on line $\overline{OA}$. Then, $|OQ| = |\alpha \cdot \beta|$.

Now suppose further $\beta \neq 0$. Then, α/β is constructed as follows. Fix a point O and fix A, P such that $|OA| = |\alpha|$ and $|OP| = 1$. Draw B on ray $\overrightarrow{OP}$ ($\overrightarrow{OP} \cdot \overrightarrow{OB} \geq 0$) such that $|OB| = |\beta|$. Draw line PQ parallel to AB such that Q lies on line OA. Then, $|OQ| = |\alpha/\beta|$. □

We may now proceed more analytically.

Proposition 8.3. The constructible numbers C form the smallest $\mathbb{Q} \leq C \leq \mathbb{R}$ such that $\sqrt{\cdot}$ maps $C \cap [0, +\infty)$ into C.

[5] This is always possible by, e.g., letting $E_j \coloneqq j$ and $J = S$.

[6] A reducible (non-constant) polynomial always factors into irreducible factors, and it has a zero iff any factor has a zero. To show no non-constant polynomial has a zero, it suffices to show no irreducible polynomial has a zero. In other words, assume $f(x)$ is "the minimal polynomial" of the zero.

TODO. □

9 Finite Fields

Definition 9.1. A field is said to be Galois if it is a finite field.

Theorem 9.2. *Let E be a finite extension of degree n over a finite field F. Then, $|E| = |F|^n$.*

Proof. Let $\alpha_1, \cdots, \alpha_n$ be a basis of E over F. Then, every $\beta \in E$ is uniquely written as $\beta = c_1 \cdot \alpha_1 + \cdots + c_n \cdot \alpha_n$, with $c_1, \cdots, c_n \in F$. Thus, $|E| = |F|^n$. □

Proposition 9.3. Let F be a finite field. Then, a subfield $H \leq F$ is isomorphic to $\mathbb{Z}_p$ for some prime p.

Proof. Let $p = \operatorname{char} F$ and map $\phi : \mathbb{Z}_p \to F$ by $n \mapsto n \cdot 1_F$, where ϕ is a field homomorphism and $n \in \mathbb{Z}$. By construction, ϕ is injective, and $\phi(\mathbb{Z}_p) \leq F$ is isomorphic to $\mathbb{Z}_p$. □

Corollary 9.4. Let E be a finite field. Then, $|E| = |\operatorname{char} E|^n$ for some $n \in \mathbb{Z}_{\geq 1}$.