Abstract Algebra

Joseph R. Mileti

January 22, 2012

Contents

Chapter 1

Introduction

1.1 What is Abstract Algebra?

Each of the sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ have naturally defined notions of addition and multiplication. Comparing these, we see many similarities. No matter which universe of numbers we are confining ourselves to live within, it does not matter which order you use to add two numbers. That is, we always have $a + b = b + a$. There are of course some differences as well. In each of the first three sets, there is no number which when mulitplied by itself gives 2, but in there latter two such a number does exist.

There is no reason to work only with objects you typically think of as numbers. You know how to add and multiply vectors, polynomials, matrices, functions, etc. In many cases, several of the "normal" properties of addition and multiplication are still true here. Sometimes the operations of addition and multiplication you define fail to have the same properties. For example, the "multiplication" of two vectors in $\mathbb{R}^3$ given by cross product is not commutative, nor is the multiplication of square matrices. The "multiplication" of two vectors in $\mathbb{R}^n$ given by the dot product takes two vectors and returns a number rather than another vector.

A primary goal of abstract algebra is to take the essential properties of these operations, codify them as axioms, and then study *all* occasions where they arise. Of course, we first need to ask the question: What is essential? Where do we draw the line for which properties we enforce by fiat? Our goal is to put enough in to force an interesting theory, but keep enough out to leave the theory as general and robust as possible. The delicate balancing act of "interesting" and "general" is no easy task.

1.2 Groups

We begin with the notion of a group. In this context, we deal with just one operation. We choose to start here in order to get practice with rigor and abstraction in as simple a setting as possible. Also, it turns out that groups appear across all areas of mathematics in many different guises.

Definition 1.2.1. *Let X be a set. A* binary operation *on X is a function $f\colon X^2 \to X$.*

In other words, a binary operation on a set X is a rule which tells you how to "put together" any two elements of X. For example, the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x^2 e^y$ is a binary operation on $\mathbb{R}$. Notice that a binary operation must be defined for *all* pairs of elements from X, and it must return an element of X. The function $f(x, y) = \frac{x}{y-1}$ is not a binary operation on $\mathbb{R}$ because it is not defined for any point of the form $(x, 1)$. The function $f\colon \mathbb{Z}^2 \to \mathbb{R}$ defined by $f(x, y) = \sin(xy)$ is defined on all of $\mathbb{Z}^2$, but it is not a binary operation on $\mathbb{Z}$ because although some values of f are integers (like $f(0, 0) = 1$), not all outputs are integers even when you provide integer inputs (for example, $f(1, 1) = \sin 1$ is not an integer).

Also, the dot product is not a binary operation on $\mathbb{R}^3$ because given two element of $\mathbb{R}^3$ it returns an element of $\mathbb{R}$.

Rather than using the standard function notation for binary operations, one typically uses the so-called infix notation. When we add two numbers, we write $x+y$ rather the far more cumbersome $+(x,y)$. For the binary operation involved with groups, we will follow this infix notation.

Definition 1.2.2. *A* group *is a set G equipped with a binary operation $\cdot$ and an element $e \in G$ such that*

1. *(Associativity) For all $a, b, c \in G$, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.*
2. *(Identity) For all $a \in G$, we have $a \cdot e = a$ and $e \cdot a = a$.*
3. *(Inverses) For all $a \in G$, there exists $b \in G$ with $a \cdot b = e$ and $b \cdot a = e$.*

In the abstract definition of a group, we have chosen to use the symbol $\cdot$ for the binary operation. This symbol may look like the "multiplication" symbol, but the operation need not be the usual multiplication in any sense. In fact, the $\cdot$ operation may be addition, exponentiation, or some crazy operation you never thought of before.

Any description of a group needs to provide three things. A set G, a binary operation $\cdot$ on G (i.e. function from G^2 to G), and a particular element $e \in G$. Absolutely any such choice of set, function, and element can comprise a group. To check whether it is indeed so, you need only check whether the above three properties are true of that fixed choice.

Here are a few examples of groups.

1. $(\mathbb{Z}, +, 0)$ is a group, as are $(\mathbb{Q}, +, 0)$, $(\mathbb{R}, +, 0)$, and $(\mathbb{C}, +, 0)$.
2. $(\mathbb{Q}\backslash\{0\}, \cdot, 1)$ is a group. We need to omit 0 because it has no multiplicative inverse. Notice that the product of two nonzero elements of $\mathbb{Q}$ is nonzero, so $\cdot$ really is a binary operation on $\mathbb{Q}\backslash\{0\}$.
3. The set of invertible 2×2 matrices over $\mathbb{R}$ with matrix multiplication and identity $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. It is important to note that the product of two invertible matrices is itself invertible, and that the inverse of an invertible matrix is itself invertible.
4. $(\{1, -1, i, -i\}, \cdot, 1)$ is a group where $\cdot$ is multiplication of complex numbers.
5. $(\{T, F\}, \oplus, F)$ where $\oplus$ is "exclusive or" on T (interpreted as "true") and F (interpreted as "false").

In contrast, here are some examples of non-groups.

1. $(\mathbb{Z}, -, 0)$ is not a group. Notice that $-$ is not associative because $(3-2)-1 = 1-1 = 0$ but $3-(2-1) = 3-1 = 2$. Also, 0 is not an identity because although $a - 0 = a$ for all $a \in \mathbb{Z}$, we have $0 - 1 = -1 \neq 1$.
2. Let S be the set of all odd elements of $\mathbb{Z}$, with the additional inclusion of 0, i.e. $S = \{0\} \cup \{n \in \mathbb{Z} : n$ is odd$\}$. Then $(S, +, 0)$ is not a group. Notice that $+$ is associative, that 0 is an identity, and that inverses exist. However, $+$ is not a binary operation on S because $1 + 3 = 4 \notin S$. In other words, S is not *closed* under $+$.
3. $(\mathbb{R}, \cdot, 1)$ is not a group because 0 has no inverse.

Returning to the examples of groups above, notice that the group operation is commutative in all the examples above except for 3. To see that the operation in 3 is not commutative, simply notice that each of the matrices

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

are invertible (they both have determinant 1), but

$$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix} = \begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}$$

while

$$\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix} = \begin{pmatrix}1 & 1\\ 1 & 2\end{pmatrix}$$

Thus, there are groups in which the group operation $\cdot$ is not commutative. The special groups that satisfy this additional fundamental property are named after Niels Abel.

Definition 1.2.3. *A group $(G, \cdot, e)$ is* abelian *if $\cdot$ is commutative, i.e. if $a \cdot b = b \cdot a$ for all $a, b \in G$. A group which is not abelian is called* nonabelian*.*

We will see many examples of nonabelian groups in time.

Notice that for a group G, there is no requirement at all that the set G or the operation $\cdot$ are in any way "natural" or "reasonable". For example, suppose that you work with the set $G = \{3, \aleph, @\}$ with operation $\cdot$ defined by the following table.

$\cdot$	3	$\aleph$	@
3	@	3	$\aleph$
$\aleph$	3	$\aleph$	@
@	$\aleph$	@	3

Interpret the above table as follows. To determine $a \cdot b$, go to the row corresponding to a and the column corresponding to b, and $a \cdot b$ will be the corresponding entry. You can indeed check that this does form a group with identity element $\aleph$. I will warn you that this is a painful check because there are 27 choices of triples for which you need to verify associativity. This table view of a group G described above is called the *Cayley table* of G.

Here is another example of a group using the set $G = \{1, 2, 3, 4, 5, 6\}$ with operation $*$ defined by the following Cayley table:

$*$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	1	6	5	4	3
3	3	5	1	6	2	4
4	4	6	5	1	3	2
5	5	3	4	2	6	1
6	6	4	2	3	1	5

It is straightforward to check that 1 is an identity for G and that every element has an inverse. However, as above, I very strongly advise against checking associativity directly. We will see later how to build many new groups and verify associativity without directly trying all possible triples. Notice that this last group is an example of a finite nonabelian group because $2 * 3 = 6$ while $3 * 2 = 5$.

1.3 Basic Properties of Groups

First, we notice that in the definition of a group it was only stated that there exists an identity element. There was no comment on how many such elements there are. We first prove that there is only one such element in every group.

Proposition 1.3.1. *Let $(G, \cdot, e)$ be a group. There exists a unique identity element in G, i.e. if $d \in G$ has the property that $a \cdot d = a$ and $d \cdot a = a$ for all $a \in G$, then $d = e$.*

Proof. The key element to consider is $d \cdot e$. On the one hand, we know that $d \cdot e = d$ because e is an identity element. On the other hand, we have $d \cdot e = e$ because d is an identity element. Therefore, $d = d \cdot e = e$, so $d = e$. □

We now move on to a similar question about inverses. The axioms only stated the existence of an inverse for every given element. We now prove uniqueness.

Proposition 1.3.2. *Let G be a group. For each $a \in G$, there exists a unique $b \in G$ such that $a \cdot b = e$ and $b \cdot a = e$.*

Proof. Fix $a \in G$. By the group axioms, we know that there exists an inverse of a. Suppose that b and c both work as inverses, i.e. that $a \cdot b = e = b \cdot a$ and $a \cdot c = e = c \cdot a$. The crucial element to think about is $(b \cdot a) \cdot c = b \cdot (a \cdot c)$. We have

$$\begin{aligned} b &= b \cdot e \\ &= b \cdot (a \cdot c) \\ &= (b \cdot a) \cdot c \\ &= e \cdot c \\ &= c \end{aligned}$$

hence $b = c$. □

Definition 1.3.3. *Given an element a in a group G, we let a^{-1} denote the unique element such that $a \cdot a^{-1} = e$ and $a^{-1} \cdot a = e$.*

Proposition 1.3.4. *Let G be a group and let $a, b \in G$.*

1. *There exists a unique element $c \in G$ such that $a \cdot c = b$, namely $c = a^{-1} \cdot b$.*
2. *There exists a unique element $c \in G$ such that $c \cdot a = b$, namely $c = b \cdot a^{-1}$.*

Proof. We prove 1 and leave 2 as an exercise (it is completely analogous). Notice that $c = a^{-1} \cdot b$ works because

$$\begin{aligned} a \cdot (a^{-1} \cdot b) &= (a \cdot a^{-1}) \cdot b \\ &= e \cdot b \\ &= b \end{aligned}$$

Suppose now that $d \in G$ satisfies $a \cdot d = b$. We then have that

$$\begin{aligned} d &= e \cdot d \\ &= (a^{-1} \cdot a) \cdot d \\ &= a^{-1} \cdot (a \cdot d) \\ &= a^{-1} \cdot b \end{aligned}$$

Here's an alternate presentation of the latter part (it is the exact same proof, just with more words and a little more motivation). Suppose that $d \in G$ satisfies $a \cdot d = b$. We then have that $a^{-1} \cdot (a \cdot d) = a^{-1} \cdot b$. By associatively, the left-hand side is $(a^{-1} \cdot a) \cdot d$, which is $e \cdot d$, with equals d. Therefore, $d = a^{-1} \cdot b$. □

Corollary 1.3.5 (Cancellation Laws)**.** *Let G be a group and let $a, b, c \in G$.*

1. *If $a \cdot c = a \cdot d$, then $c = d$.*
2. *If $c \cdot a = d \cdot a$, then $c = d$.*

Proof. Suppose that $a \cdot c = a \cdot d$. Letting b equal this common value, it follows the uniqueness part of Proposition 1.3.4 that $c = d$. Alternatively, multiply both sides on the left by a^{-1} and use associativity. Part 2 is completely analogous. □

In terms of the Cayley table of a group G, Proposition 1.7 says that every element of a group appears exactly once in each row of G and exactly once in each column of G. In fancy terminology, the Cayley table of a group is a Latin square.

Proposition 1.3.6. *Let G be a group and let $a \in G$. We have $(a^{-1})^{-1} = a$.*

Proof. By definition we have $a \cdot a^{-1} = e = a^{-1} \cdot a$. Thus, a satisfies the requirement to be the inverse of a^{-1}. In other words, $(a^{-1})^{-1} = a$. □

Proposition 1.3.7. *Let G be a group and let $a, b \in G$. We have $(a \cdot b)^{-1} = b^{-1} \cdot a^{-1}$.*

Proof. We have

$$\begin{aligned}(a \cdot b) \cdot (b^{-1} \cdot a^{-1}) &= ((a \cdot b) \cdot b^{-1}) \cdot a^{-1} \\ &= (a \cdot (b \cdot b^{-1})) \cdot a^{-1} \\ &= (a \cdot e) \cdot a^{-1} \\ &= a \cdot a^{-1} \\ &= e\end{aligned}$$

and similarly $(b^{-1} \cdot a^{-1}) \cdot (a \cdot b) = e$. Therefore, $(a \cdot b)^{-1} = b^{-1} \cdot a^{-1}$. □

Chapter 2

The Integers

2.1 Induction and Well-Ordering

We first recall three fundamental and intertwined facts about the natural numbers, namely Induction, Strong Induction, and Well-Ordering on $\mathbb{N}$. We will not attempt to "prove" them or argue in what sense they are equivalent because in order to do so in a satisfying manner we would need to make clear our fundamental assumptions about $\mathbb{N}$. In any case, I hope that each of these facts is intuitively clear with a bit of thought.

Fact 2.1.1 (Principle of Mathematical Induction on $\mathbb{N}$)**.** *Let $X \subseteq \mathbb{N}$. Suppose that*

- $0 \in X$ *(the* base case*)*
- $n + 1 \in X$ *whenever $n \in X$ (the* inductive step*)*

We then have that $X = \mathbb{N}$.

Here's the intuitive argument for why induction is true. By the first assumption, we know that $0 \in X$. Since $0 \in X$, the second assumption tells us that $1 \in X$. Since $1 \in X$, the second assumption again tells us that $2 \in X$. By repeatedly applying the second assumption in this manner, each element of $\mathbb{N}$ is eventually determined to be in X.

The way that induction works above is that 5 is shown to be an element of X using only the assumption that 4 is an element of X. However, once you've arrived at 5, you've already shown that $0, 1, 2, 3, 4 \in X$, so why shouldn't you be able to make use of all of these assumptions when arguing that $5 \in X$? The answer is that you can, and this version of induction is sometimes called *strong induction*.

Fact 2.1.2 (Principle of Strong Induction on $\mathbb{N}$)**.** *Let $X \subseteq \mathbb{N}$. Suppose that $n \in X$ whenever $k \in X$ for all $k \in \mathbb{N}$ with $k < n$. We then have that $X = \mathbb{N}$.*

Thus, when arguing that $n \in X$, we are allowed to assume that we know all smaller numbers are in X. Notice that with this formulation we can even avoid the base case of checking 0 because of a technicality: If you have the above assumption, then $0 \in X$ because vacuously $k \in X$ whenever $k \in \mathbb{N}$ satisfies $k < 0$, simply because no such k exists. If that twist of logic makes you uncomfortable, feel free to argue a base case of 0 when doing strong induction.

The last of the three fundamental facts looks different from induction, but like induction is based on the concept that natural numbers start with 0 and are built by taking one discrete step at a time forward.

Fact 2.1.3 (Well-Ordering Property of $\mathbb{N}$)**.** *Suppose that $X \subseteq \mathbb{N}$ with $X \neq \emptyset$. There exists $k \in X$ such that $k \leq n$ for all $n \in X$.*

To see intuitively why this is true, suppose that $X \subseteq \mathbb{N}$ with $X \neq \emptyset$. If $0 \in X$, then we can take $k = 0$. Suppose not. If $1 \in X$, then since $1 \leq n$ for all $n \in \mathbb{N}$ with $n \neq 0$, we can take $k = 1$. Continue on. If we keep going until we eventually find $7 \in X$, then we know that $0, 1, 2, 3, 4, 5, 6 \notin X$, so we can take $k = 7$. If we keep going forever consistently finding that each natural number is not in X, then we have determined that $X = \emptyset$, which is a contradiction.

We now give an example of proof by induction. Notice that in this case we start with the base case of 1 rather than 0.

Theorem 2.1.4. *For any $n \in \mathbb{N}^+$, we have*

$$\sum_{k=1}^{n}(2k-1) = n^2$$

i.e.

$$1 + 3 + 5 + 7 + \cdots + (2n-1) = n^2$$

Proof. We prove the result by induction. If you want to formally apply the above statement of induction, we are letting

$$X = \{n \in \mathbb{N}^+ : \sum_{k=1}^{n}(2k-1) = n^2\}$$

and using the principle of induction to argue that $X = \mathbb{N}^+$. More formally still if you feel uncomfortable starting with 1 rather than 0, we are letting

$$X = \{0\} \cup \{n \in \mathbb{N}^+ : \sum_{k=1}^{n}(2k-1) = n^2\}$$

and using the principle of induction to argue that $X = \mathbb{N}$, then forgetting about 0 entirely. In the future, we will not bother to make these pedantic diversions to shoehorn our arguments into the technical versions expressed above, but you should know that it is always possible to do so.

- *Base Case:* Suppose that $n = 1$. We have

  $$\sum_{k=1}^{1}(2k-1) = 2 \cdot 1 - 1 = 1$$

  so the left hand-side is 1. The right-hand side is $1^2 = 1$. Therefore, the result is true when $n = 1$.

- *Inductive Step:* Suppose that for some fixed $n \in \mathbb{N}^+$ we know that

  $$\sum_{k=1}^{n}(2k-1) = n^2$$

  Notice that $2(n+1) - 1 = 2n + 2 - 1 = 2n + 1$, hence

  $$\begin{aligned}
  \sum_{k=1}^{n+1}(2k-1) &= [\sum_{k=1}^{n}(2k-1)] + [2(n+1) - 1] \\
  &= [\sum_{k=1}^{n}(2k-1)] + (2n+1) \\
  &= n^2 + (2n+1) && \text{(by induction)} \\
  &= (n+1)^2
  \end{aligned}$$

Since the result holds of 1 and it holds of $n+1$ whenever it holds of n, we conclude that the result holds for all $n \in \mathbb{N}^+$ by induction. □

Our other example of induction will be the Binomial Theorem, which tells us how to completely expand a binomial to a power, i.e. how to expand the expression $(x+y)^n$. For the first few values we have:

- $(x+y)^1 = x+y$
- $(x+y)^2 = x^2+2xy+y^2$
- $(x+y)^3 = x^3+3x^2y+3xy^2+y^3$
- $(x+y)^4 = x^4+4x^3y+6x^2y^2+4xy^3+y^4$

We seek to determine the coefficients of the various x^iy^j terms.

Definition 2.1.5. *Let $n, k \in \mathbb{N}$ with $k \leq n$. We define*

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

In combinatorics, you see that $\binom{n}{k}$ is the number of subsets of an n-element set of size k. In particular, the number $\binom{n}{k}$ is always an integer, which might be surprising from a naive glance at the formula. The following fundamental lemma gives a recursive way to calculate $\binom{n}{k}$.

Lemma 2.1.6. *Let $n, k \in \mathbb{N}$ with $k \leq n$. We have*

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$$

Proof. One extremely unenlightening proof is to expand out the formula on the right and do terrible algebraic manipulations on it. If you haven't done so, I encourage you to do it. If you believe the combinatorial description of $\binom{n}{k}$, here's a more meaningful combinatorial argument. Let $n, k \in \mathbb{N}$ with $k \leq n$. Consider a set X with $n+1$ many elements. To determine $\binom{n+1}{k+1}$, we need to count the number of subsets of X of size $k+1$. We do this as follows. Fix an arbitrary $a \in X$. Now an arbitrary subset of X of size $k+1$ fits into exactly one of the following types.

- The subset has a as an element. In this case, to completely determine the subset, we need to pick the remaining k elements of the subset from $X \backslash \{a\}$. Since $X \backslash \{a\}$ has n elements, the number of ways to do this is $\binom{n}{k}$.
- The subset does not have a as an element. In this case, to completely determine the subset, we need to pick all $k+1$ elements of the subset from $X \backslash \{a\}$. Since $X \backslash \{a\}$ has n elements, the number of ways to do this is $\binom{n}{k+1}$.

Putting this together, we conclude that the number of subsets of X of size $k+1$ equals $\binom{n}{k} + \binom{n}{k+1}$. □

Theorem 2.1.7 (Binomial Theorem)**.** *Let $x, y \in \mathbb{R}$ and let $n \in \mathbb{N}^+$. We have*

$$\begin{aligned}(x+y)^n &= \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n \\ &= \sum_{k=0}^{n} \binom{n}{k} x^{n-k}y^k\end{aligned}$$

Proof. We prove the result by induction. The base case is trivial. Suppose that we know the result for a given $n \in \mathbb{N}^+$. We have

$$\begin{aligned}
(x+y)^{n+1} &= (x+y)\cdot(x+y)^n \\
&= (x+y)\cdot(\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n) \\
&= \binom{n}{0}x^{n+1} + \binom{n}{1}x^n y + \binom{n}{2}x^{n-1}y^2 + \cdots + \binom{n}{n-1}x^2y^{n-1} + \binom{n}{n}xy^n \\
&\qquad + \binom{n}{0}x^n y + \binom{n}{1}x^{n-1}y^2 + \cdots + \binom{n}{n-2}x^2y^{n-1} + \binom{n}{n-1}xy^n + \binom{n}{n}y^{n+1} \\
&= x^{n+1} + (\binom{n}{1} + \binom{n}{0})\cdot x^n y + (\binom{n}{2} + \binom{n}{1})\cdot x^{n-1}y^2 + \cdots + (\binom{n}{n} + \binom{n}{n-1})\cdot xy^n + y^{n+1} \\
&= \binom{n+1}{0}x^{n+1} + \binom{n+1}{1}x^n y + \binom{n+1}{2}x^{n-1}y^2 + \cdots + \binom{n+1}{n}xy^n + \binom{n+1}{n+1}y^{n+1}
\end{aligned}$$

where we have used the lemma to combine each of the sums to get the last line. □

2.2 Divisibility

Definition 2.2.1. *Let $a, b \in \mathbb{Z}$. We say that a* divides *b, and write $a \mid b$, if there exists $m \in \mathbb{Z}$ with $b = am$.*

For example, we have $2 \mid 6$ because $2 \cdot 3 = 6$ and $-3 \mid 21$ because $-3 \cdot 7 = 21$. We also have that $2 \nmid 5$ since it is "obvious" that no such integer exists. If you are uncomfortable with that (and there is certainly reason to be), we will be much more careful about such statements in a couple of sections.

Notice that $a \mid 0$ for every $a \in \mathbb{Z}$ because $a \cdot 0 = 0$ for all $a \in \mathbb{Z}$. In particular, we have $0 \mid 0$ because as noted we have $0 \cdot 0 = 0$. Of course we also have $0 \cdot 3 = 0$ and in fact $0 \cdot m = 0$ for all $m \in \mathbb{Z}$, so every integer serves as a "witness" that $0 \mid 0$. Our definition says nothing about the $m \in \mathbb{Z}$ being unique.

Proposition 2.2.2. *If $a \mid b$ and $b \mid c$, then $a \mid c$.*

Proof. Since $a \mid b$, there exists $m \in \mathbb{Z}$ with $b = am$. Since $b \mid c$, there exists $n \in \mathbb{Z}$ with $c = bn$. We then have

$$c = bn = (am)n = a(mn)$$

Since $mn \in \mathbb{Z}$, it follows that $a \mid c$. □

Proposition 2.2.3.

1. *If $a \mid b$, then $a \mid bk$ for all $k \in \mathbb{Z}$.*
2. *If $a \mid b$ and $a \mid c$, then $a \mid (b+c)$.*
3. *If $a \mid b$ and $a \mid c$, then $a \mid (bm+cn)$ for all $m, n \in \mathbb{Z}$.*

Proof.

1. Let $k \in \mathbb{Z}$. Since $a \mid b$, there exists $m \in \mathbb{Z}$ with $b = am$. We then have

$$bk = (am)k = a(mk)$$

Since $mk \in \mathbb{Z}$, it follows that $a \mid bk$.

2. Since $a \mid b$, there exists $m \in \mathbb{Z}$ with $b = am$. Since $a \mid c$, there exists $n \in \mathbb{Z}$ with $c = an$. We then have

$$b + c = am + an = a(m + n)$$

Since $m + n \in \mathbb{Z}$, it follows that $a \mid b + c$.

3. This follows by combining 1 and 2 as follows. Let $m, n \in \mathbb{Z}$. Since $a \mid b$, we conclude from part 1 that $a \mid bm$. Since $a \mid c$, we conclude from part 1 again that $a \mid cn$. Using part 2, it follows that $a \mid (bm + cn)$.

□

Proposition 2.2.4. *Suppose that $a, b \in \mathbb{Z}$. If $a \mid b$ and $b \neq 0$, then $|a| \leq |b|$.*

Proof. Suppose that $a \mid b$ with $b \neq 0$. Fix $d \in \mathbb{Z}$ with $ad = b$. Since $b \neq 0$, we have $d \neq 0$. Thus, $|d| \geq 1$, and so

$$|b| = |ad| = |a| \cdot |d| \geq |a| \cdot 1 = |a|$$

□

Corollary 2.2.5. *Suppose that $a, b \in \mathbb{Z}$. If $a \mid b$ and $b \mid a$, then either $a = b$ or $a = -b$.*

Proof. Suppose first that $a \neq 0$ and $b \neq 0$. By the previous Proposition, we know that both $|a| \leq |b|$ and $|b| \leq |a|$. It follows that $|a| = |b|$, and hence either $a = b$ or $a = -b$.

Suppose now that $a = 0$. As above, since $a \mid b$, we may fix $m \in \mathbb{Z}$ with $b = am$. We then have $b = am = 0m = 0$ as well. Therefore, $a = b$.

Suppose fiinally that $b = 0$. Since $b \mid a$, we may fix $m \in \mathbb{Z}$ with $a = bm$. We then have $a = bm = 0m = 0$ as well. Therefore, $a = b$. □

2.3 Division with Remainder

The primary goal of this section is to prove the following deeply fundamental result.

Theorem 2.3.1. *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. There exist unique $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < |b|$. Uniqueness here means that if $a = q_1 b + r_1$ with $0 \leq r_1 < |b|$ and $a = q_2 b + r_2$ with $0 \leq r_2 < |b|$, then $q_1 = q_2$ and $r_1 = r_2$.*

Here are a bunch of examples illustrating existence:

- Let $a = 5$ and $b = 2$. We have $5 = 2 \cdot 2 + 1$
- Let $a = 135$ and $b = 45$. We have $135 = 3 \cdot 45 + 0$
- Let $a = 60$ and $b = 9$. We have $60 = 6 \cdot 9 + 6$
- Let $a = 29$ and $b = -11$. We have $29 = (-2)(-11) + 7$
- Let $a = -45$ and $b = 7$. We have $-45 = (-7) \cdot 7 + 4$
- Let $a = -21$ and $b = -4$. We have $-21 = 6 \cdot (-4) + 3$

We begin by proving existence via a sequence of lemmas, starting in the case where a, b are natural numbers rather than just integers.

Lemma 2.3.2. *Let $a, b \in \mathbb{N}$ with $b > 0$. There exist $q, r \in \mathbb{N}$ such that $a = qb + r$ and $0 \leq r < b$.*

Proof. Fix $b \in \mathbb{N}$ with $b > 0$. For this fixed b, we prove the existence of q, r for all $a \in \mathbb{N}$ by induction. That is, for this fixed b, we define

$$X = \{a \in \mathbb{N} : \text{ There exist} q, r \in \mathbb{N} \text{ with } a = qb + r\}$$

and show that $X = \mathbb{N}$ by induction.

- *Base Case:* Suppose that $a = 0$. We then have $a = 0 \cdot b + 0$ and clearly $0 < b$, so we may take $q = 0$ and $r = 0$.
- *Inductive Step:* Suppose that we know the result for a given $a \in \mathbb{N}$. Fix $q, r \in \mathbb{Z}$ with $0 \leq r < b$ such that $a = qb + r$. We then have $a + 1 = qb + (r + 1)$. Since $r, b \in \mathbb{N}$ with $r < b$, we know that $r + 1 \leq b$. If $r + 1 < b$, then we are done. Otherwise, we have $r + 1 = b$, hence

$$\begin{aligned} a + 1 &= qb + (r + 1) \\ &= qb + b \\ &= (q + 1)b \\ &= (q + 1)b + 0 \end{aligned}$$

  and so we may take $q + 1$ and 0.

The result follows by induction. □

Lemma 2.3.3. *Let $a, b \in \mathbb{Z}$ with $b > 0$. There exist $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < b$.*

Proof. If $a \geq 0$, we are done by the previous proposition. Suppose that $a < 0$. We then have $-a > 0$, so by the previous proposition we may fix $q, r \in \mathbb{N}$ with $0 \leq r < b$ such that $-a = qb + r$. We then have $a = -(qb + r) = (-q)b + (-r)$. If $r = 0$, the $-r = 0$ and we are done. Otherwise we $0 < r < b$ and

$$\begin{aligned} a &= (-q)b + (-r) \\ &= (-q)b - b + b + (-r) \\ &= (-q - 1)b + (b - r) \end{aligned}$$

Now since $0 < r < b$, we have $0 < b - r < b$, so this gives existence. □

Lemma 2.3.4. *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. There exist $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < |b|$.*

Proof. If $b > 0$, we are done by the previous corollary. Suppose that $b < 0$. We then have $-b > 0$, so there exist $q, r \in \mathbb{N}$ with $0 \leq r < -b$ and $a = q(-b) + r$. We then have $a = (-q)b + r$ and we are done because $|b| = -b$. □

With that sequence of lemmas building to existence now in hand, we finish off the proof of the theorem.

Proof of Theorem 2.3.1. The final lemma above gives us existence. Suppose that

$$q_1 b + r_1 = a = q_2 b + r_2$$

where $0 \leq r_1 < |b|$ and $0 \leq r_2 < |b|$. We then have

$$b(q_2 - q_1) = r_1 - r_2$$

hence $b \mid (r_2 - r_1)$. Now $-|b| < -r_1 \leq 0$, so adding this to $0 \leq r_2 < |b|$, we conclude that

$$-|b| < r_2 - r_1 < |b|$$

and therefore

$$|r_2 - r_1| < |b|$$

Now if $r_2 - r_1 \neq 0$, then since $b \mid (r_2 - r_1)$, we would conclude that $|b| \leq |r_2 - r_1|$, a contradiction. It follows that $r_2 - r_1 = 0$, and hence $r_1 = r_2$. Since

$$q_1 b + r_1 = q_2 b + r_2$$

and $r_1 = r_2$, we conclude that $q_1 b = q_2 b$. Now $b \neq 0$, so it follows that $q_1 = q_2$. □

Proposition 2.3.5. *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. Write $a = qb + r$ for the unique choice of $q, r \in \mathbb{Z}$ with $0 \leq r < |b|$. We then have that $b \mid a$ if and only if $r = 0$.*

Proof. If $r = 0$, then $a = qb + r = bq$, so $b \mid a$. Suppose conversely that $b \mid a$ and fix $m \in \mathbb{Z}$ with $a = bm$. We then have $a = mb + 0$ and $a = qb + r$, so by the uniqueness part of the above theorem, we must have $r = 0$. □

2.4 GCDs and the Euclidean Algorithm

Definition 2.4.1. *Suppose that $a, b \in \mathbb{Z}$. We say that $d \in \mathbb{Z}$ is a* common divisor *of a and b if both $d \mid a$ and $d \mid b$.*

The common divisors of 120 and 84 are $\{\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 12\}$ (we will see a careful argument below). The common divisors of 10 and 0 are $\{\pm 1, \pm 2, \pm 5, \pm 10\}$. Every element of $\mathbb{Z}$ is a common divisor of 0 and 0. The following little proposition is fundamental to this entire section.

Proposition 2.4.2. *Suppose that $a, b, q, r \in \mathbb{Z}$ and $a = qb + r$. For any $d \in \mathbb{Z}$, we have that d is a common divisor of a and b if and only if d is a common divisor of b and r, i.e.*

$$\{d \in \mathbb{Z} : d \text{ is a common divisor of } a \text{ and } b\} = \{d \in \mathbb{Z} : d \text{ is a common divisor of } b \text{ and } r\}$$

Proof. Suppose first that d is a common divisor of b and r. Since $d \mid b$, $d \mid r$, and $a = qb + r = bq + r1$, we may use Proposition 2.2.3 to conclude that $d \mid a$.

Conversely, suppose that d is a common divisor of a and b. Since $d \mid a$, $d \mid b$, and $r = a - qb = a1 + b(-q)$, we may use Proposition 2.2.3 to conclude that $d \mid r$. □

For example, suppose that we are trying to find the set of common divisors of 120 and 84 (we wrote them above, but now want to justify it). We repeatedly do division to reduce the problem as follows:

$$\begin{aligned} 120 &= 1 \cdot 84 + 36 \\ 84 &= 2 \cdot 36 + 12 \\ 36 &= 3 \cdot 12 + 0 \end{aligned}$$

The first line tells us that the set of common divisors of 120 and 84 equals the set of common divisors of 84 and 36. The next line tells us that the set of common divisors of 84 and 36 equals the set of common divisors of 36 and 12. The last line tells us that the set of common divisors of 36 and 12 equals the set of common divisors of 12 and 0. Now the set of common divisors of 12 and 0 is simply the set of divisors of 12 (because every number divides 0). Putting it all together, we conclude that the set of common divisors of 120 and 84 equals the set of divisors of 12.

Definition 2.4.3. *Let $a, b \in \mathbb{Z}$. We say that an element $d \in \mathbb{Z}$ is a* greatest common divisor *of a and b if:*

- $d \geq 0$

- *d is a common divisor of a and b.*
- *Whenever $c \in \mathbb{Z}$ is a common divisor of a and b, we have $c \mid d$.*

Notice that we are *not* defining the greatest common divisor of a and b to be the largest divisor of a and b. The primary reason we do not is because this description fails to capture the most fundamental property (namely that of being divisible by all other divisors, not just larger than them). Furthermore, if you were to take that definition, then 0 and 0 would fail to have a greatest common divisor because every integer is a common divisor of 0 and 0. With this definition however, it is a straightforward matter to check that 0 satisfies the above three conditions.

Since we require more of a greatest common divisor than just picking the largest, we first need to check that they do indeed exist. The proof is an inductive formulation of the above method of calculation.

Theorem 2.4.4. *Every pair of integers $a, b \in \mathbb{Z}$ has a unique greatest common divisor.*

We first sketch the idea of the proof in the case where $a, b \in \mathbb{N}$. If $b = 0$, we are done because it is simple to verify that a is a greatest common divisor of a and 0. Suppose then that $b \neq 0$. Fix $q, r \in \mathbb{N}$ with $a = qb + r$ and $0 \leq r < b$. Now the idea is to assert inductively the existence of a greatest common divisor of b and r because this pair is "smaller" than the pair a and b. The only issue is how to make this intuitive idea of "smaller" precise. There are several ways to do this, but perhaps the most straightforward is to only induct on b. Thus, our base case handles all pairs of form $(a, 0)$. Next, we handle all pairs of the form $(a, 1)$ and in doing this we can use the fact the we know the result for all pairs of the form $(a', 0)$. Notice that we can we even change the value of the first coordinate here which is why we used a'. Then, we handle all pairs of the form $(a, 2)$ and in doing this we can use the fact that we know the result for all pairs of the form $(a', 0)$ and $(a', 1)$. We now begin the formal argument.

Proof. We begin by proving existence only in the special case where $a, b \in \mathbb{N}$. We use (strong) induction on b to prove the result. That is, we let

$$X = \{b \in \mathbb{N} : \text{For all } a \in \mathbb{N}, \text{ there exists a greatest common divisor of } a \text{ and } b\}$$

and prove that $X = \mathbb{N}$ by strong induction.

- *Base Case:* Suppose that $b = 0$. Let $a \in \mathbb{N}$ be arbitrary. We then have that the set of common divisors of a and $b = 0$ equals the set of divisors of a (because every integer divides 0), so a satisfies the requirement of a greatest common divisor of a and 0. Since $a \in \mathbb{N}$ was arbitrary, we showed that there exists a greatest common divisor of a and 0 for every $a \in \mathbb{N}$, hence $0 \in X$.
- *Inductive Step:* Suppose then that $b \in \mathbb{N}^+$ and we know the result for all smaller natural numbers. In other words, we are assuming that $c \in X$ whenever $0 \leq c < b$. We prove that $b \in X$. Let $a \in \mathbb{N}$ be arbitrary. From above, we may fix $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < b$. Since $0 \leq r < b$, we know by strong induction that $r \in X$, hence b and r have a greatest common divisor d. By the Proposition 2.4.2, the set of common divisors of a and b equals the set of common divisors of b and r. It follows that d is a greatest common divisor of a and b. Since $a \in \mathbb{N}$ was arbitrary, we showed that there exists a greatest common divisor of a and b for every $a \in \mathbb{N}$, hence $b \in X$.

Therefore, we have shown that $X = \mathbb{N}$, which implies that whenever $a, b \in \mathbb{N}$, there exists a greatest common divisor of a and b.

To turn the argument into a proof for all $a, b \in \mathbb{Z}$, we simply note the set of divisors of an element $m \in \mathbb{Z}$ equals the set of divisors of $-m$. So, for example, if $a < 0$ but $b \geq 0$ we can simply take a greatest common divisor of $-a$ and b (which exists since $-a, b \in \mathbb{N}$) will also be a greatest common divisor of a and b. A similar argument works if $a \geq 0$ and $b < 0$, or if both $a < 0$ and $b < 0$.

For uniqueness, suppose that c and d are both greatest common divisors of a and b. Since d is a greatest common divisor and c is a common divisor, we know by the last condition that $c \mid d$. Similarly, since c is a greatest common divisor and d is a common divisor, we know by the last condition that $d \mid c$. Therefore, either $c = d$ or $c = -d$. Using the first requirement that a greatest common divisor must be nonnegative, we must have $c = d$. □

Definition 2.4.5. *Let $a, b \in \mathbb{Z}$. We let $\gcd(a, b)$ be the unique greatest common divisor of a and b.*

For example we have $\gcd(120, 84) = 12$ and $\gcd(0, 0) = 0$. The following corollary is immediate from Proposition 2.4.2.

Corollary 2.4.6. *Suppose that $a, b, q, r \in \mathbb{Z}$ and $a = qb + r$. We have $\gcd(a, b) = \gcd(b, r)$.*

The method of using repeated division and this corollary to reduce the problem of calculating greatest common divisors is known as the *Euclidean Algorithm*. We saw it in action of above with 120 and 84. Here is another example where we are trying to compute $\gcd(525, 182)$. We have

$$\begin{aligned} 525 &= 2 \cdot 182 + 161 \\ 182 &= 1 \cdot 161 + 21 \\ 161 &= 7 \cdot 21 + 14 \\ 21 &= 1 \cdot 14 + 7 \\ 14 &= 2 \cdot 7 + 0 \end{aligned}$$

Therefore, $\gcd(525, 182) = \gcd(7, 0) = 7$.

Theorem 2.4.7. *For all $a, b \in \mathbb{Z}$, there exist $k, \ell \in \mathbb{Z}$ with $\gcd(a, b) = ka + \ell b$.*

Proof. We begin by proving existence in the special case where $a, b \in \mathbb{N}$. We use induction on b to prove the result. That is, we let

$$X = \{b \in \mathbb{N} : \text{For all } a \in \mathbb{N}, \text{ there exist } k, \ell \in \mathbb{Z} \text{ with } \gcd(a, b) = ka + \ell b\}$$

and prove that $X = \mathbb{N}$ by strong induction.

- *Base Case:* Suppose that $b = 0$. Let $a \in \mathbb{N}$ be arbitrary. We then have that

  $$\gcd(a, b) = \gcd(a, 0) = a$$

  Since $a = 1 \cdot a + 0 \cdot b$, so we may let $k = 1$ and $\ell = 0$. Since $a \in \mathbb{N}$ was arbitrary, we conclude that $0 \in X$.

- *Inductive Step:* Suppose then that $b \in \mathbb{N}^+$ and we know the result for all smaller nonnegative values. In other words, we are assuming that $c \in X$ whenever $0 \leq c < b$. We prove that $b \in X$. Let $a \in \mathbb{N}$ be arbitrary. From above, we may fix $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < b$. We also know from above that $\gcd(a, b) = \gcd(b, r)$. Since $0 \leq r < b$, we know by strong induction that $r \in X$, hence there exist $k, \ell \in \mathbb{Z}$ with

  $$\gcd(b, r) = kb + \ell r$$

  Now $r = a - qb$, so

  $$\begin{aligned} \gcd(a, b) &= \gcd(b, r) \\ &= kb + \ell r \\ &= kb + \ell(a - qb) \\ &= kb + \ell a - qb\ell \\ &= \ell a + (k - q\ell)b \end{aligned}$$

  Since $a \in \mathbb{N}$ was arbitrary, we conclude that $b \in X$.

Therefore, we have shown that $X = \mathbb{N}$, which implies that whenever $a, b \in \mathbb{N}$, there exists $k, \ell \in \mathbb{Z}$ with $\gcd(a, b) = ka + \ell b$. □

Given $a, b \in \mathbb{Z}$, we can explicitly calculate $k, \ell \in \mathbb{Z}$ by "winding up" the work created from the Euclidean Algorithm. For example, we saw above that $\gcd(525, 182) = 7$ by calculating

$$\begin{aligned} 525 &= 2 \cdot 182 + 161 \\ 182 &= 1 \cdot 161 + 21 \\ 161 &= 7 \cdot 21 + 14 \\ 21 &= 1 \cdot 14 + 7 \\ 14 &= 2 \cdot 7 + 0 \end{aligned}$$

We now use these steps in reverse to calculate:

$$\begin{aligned} 7 &= 1 \cdot 7 + 0 \cdot 0 \\ &= 1 \cdot 7 + 0 \cdot (14 - 2 \cdot 7) \\ &= 0 \cdot 14 + 1 \cdot 7 \\ &= 0 \cdot 14 + 1 \cdot (21 - 1 \cdot 14) \\ &= 1 \cdot 21 + (-1) \cdot 14 \\ &= 1 \cdot 21 + (-1) \cdot (161 - 7 \cdot 21) \\ &= (-1) \cdot 161 + 8 \cdot 21 \\ &= (-1) \cdot 161 + 8 \cdot (182 - 1 \cdot 161) \\ &= 8 \cdot 182 + (-9) \cdot 161 \\ &= 8 \cdot 182 + (-9) \cdot (525 - 2 \cdot 182) \\ &= (-9) \cdot 525 + 26 \cdot 182 \end{aligned}$$

This wraps everything up perfectly, but it is easier to simply start at the fifth line.

Theorem 2.4.8. *Let $a, b \in \mathbb{Z}$ with at least one of a and b nonzero. We then have that $\gcd(a, b)$ is the least positive element of the set*

$$\{am + bn : m, n \in \mathbb{Z}\}$$

In particular, if $d = \gcd(a, b)$, then there exist $k, \ell \in \mathbb{Z}$ with $d = ka + \ell b$.

Proof. This can be proved using the above results, but we give a direct proof. Let

$$S = \{ma + nb : m, n \in \mathbb{Z}\} \cap \mathbb{N}^+$$

We first claim that $S \neq \emptyset$. If $a > 0$, then $a = 1 \cdot a + 0 \cdot b \in S$. Similarly, if $b > 0$, then $b \in S$. If $a < 0$, then $-a > 0$ and $-a = (-1) \cdot a + 0 \cdot b \in S$. Similarly, if $b < 0$, then $-b \in S$. Since at least one of a and b is nonzero, it follows that $S \neq \emptyset$. By the Well-Ordering property of $\mathbb{N}$, we know that S has a least element. Let $d = \min(S)$. Since $d \in S$, we may fix $k, \ell \in \mathbb{Z}$ with $d = ka + \ell b$. We claim that d is the greatest common divisor of a and b.

First, we need to check that d is a common divisor of a and b. We begin by showing that $d \mid a$. Fix $q, r \in \mathbb{Z}$ with $a = qd + r$ and $0 \leq r < d$. We want to show that $r = 0$. We have

$$\begin{aligned} r &= a - qd \\ &= a - q(ak + b\ell) \\ &= (1 - qk) \cdot a + (-q\ell) \cdot b \end{aligned}$$

Now if $r > 0$, then we have shown that $r \in S$, which contradicts the choice of d as the least element of S. Hence, we must have $r = 0$, and so $d \mid a$.

We next show that $d \mid b$. Fix $q, r \in \mathbb{Z}$ with $b = qd + r$ and $0 \leq r < d$. We want to show that $r = 0$. We have

$$\begin{aligned} r &= b - qd \\ &= b - q(ak + b\ell) \\ &= (-qk) \cdot a + (1 - q\ell) \cdot b \end{aligned}$$

Now if $r > 0$, then we have shown that $r \in S$, which contradicts the choice of d as the least element of S. Hence, we must have $r = 0$, and so $d \mid b$.

Finally, we need to check the last condition for d to be the greatest common divisor. Let c be ancommon divisor of a and b. Since $c \mid a$, $c \mid b$, and $d = ka + \ell b$, we may use Proposition 2.2.3 to conclude that $c \mid d$. □

Definition 2.4.9. *Two elements $a, b \in \mathbb{Z}$ are* relatively prime *if $\gcd(a, b) = 1$.*

Proposition 2.4.10. *Let $a, b, c \in \mathbb{Z}$. If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.*

Proof. Since $a \mid bc$, we may fix $m \in \mathbb{Z}$ with $bc = am$. Since $\gcd(a, b) = 1$, we may fix $k, \ell \in \mathbb{Z}$ with $ak + b\ell = 1$. Multiplying this last equation through by c we conclude that $akc + b\ell c = c$, so

$$\begin{aligned} c &= akc + \ell(bc) \\ &= akc + na\ell \\ &= a(kc + n\ell) \end{aligned}$$

It follows that $a \mid c$. □

2.5 Primes and Factorizations in $\mathbb{Z}$

Definition 2.5.1. *An element $p \in \mathbb{Z}$ is prime if $p > 1$ and the only positive divisors of p are 1 and p. If $n \in \mathbb{Z}$ with $n > 1$ is not prime, we say that n is* composite.

We begin with the following simple fact.

Proposition 2.5.2. *Every $n \in \mathbb{N}$ with $n > 1$ is a product of primes.*

Proof. We prove the result by induction on $\mathbb{N}$. If $n = 2$, we are done because 2 itself is prime. Suppose that $n > 2$ and we have proven the result for all k with $1 < k < n$. If n is prime, we are done. Suppose that n is not prime and fix a divisor $c \mid n$ with $1 < c < n$. Fix $d \in \mathbb{N}$ with $cd = n$. We then have that $1 < d < n$, so by induction, both c and d are products of primes, say $c = p_1 p_2 \cdots p_k$ and $d = q_1 q_2 \cdots q_\ell$ with each p_i and q_j prime. We then have

$$n = cd = p_1 p_2 \cdots p_k q_1 q_2 \cdots q_\ell$$

so n is a product of primes. The result follows by induction. □

Corollary 2.5.3. *Every $a \in \mathbb{Z}$ with $a \notin \{-1, 0, -1\}$ is either a product of primes, or -1 times a product of primes.*

Corollary 2.5.4. *Every $a \in \mathbb{Z}$ with $a \notin \{-1, 0, -1\}$ is divisible by at least one prime.*

Proposition 2.5.5. *There are infinitely many primes.*

Proof. We know that 2 is a prime, so there is at least one prime. We will take an arbitrary given finite list of primes and show that there exists a prime which is omitted. Suppose then that $p_1, p_2, \ldots, p_k$ is an arbitrary finite list of prime numbers with $k \geq 1$. We show that there exists a prime not in the list. Let

$$n = p_1 p_2 \cdots p_k + 1$$

We have $n \geq 3$, so by the above corollary we know that n is divisible by some prime q. If $q = p_i$, we would have that $q \mid n$ and also $q \mid p_1 p_2 \cdots p_k$, so $q \mid (n - p_1 p_2 \cdots p_k)$. This would imply that $q \mid 1$, so $|q| \leq 1$, a contradiction. Therefore $q \neq p_i$ for all i, and we have succeeded in finding a prime not in the list. □

The next proposition using our hard work on greatest common divisors. It really is the most useful and fundamental properties of prime numbers.

Proposition 2.5.6. *If p is prime and $p \mid ab$, then either $p \mid a$ or $p \mid b$.*

Proof. Suppose that $p \mid ab$ and $p \nmid a$. Since $\gcd(a, p)$ divides p and we know that $p \nmid a$, we have $\gcd(a, p) \neq p$. The only other positive divisor of p is 1, so $\gcd(a, p) = 1$. Therefore, by the Proposition 2.4.10, we conclude that $p \mid b$. □

Now that we've handled the product of two numbers, you get the following corollary about finite products by a trivial induction.

Corollary 2.5.7. *If p is prime and $p \mid a_1 a_2 \cdots a_n$, then $p \mid a_i$ for some i.*

We now have all the tools necessary to provide the uniqueness of prime factorizations.

Theorem 2.5.8 (Fundamental Theorem of Arithmetic)**.** *Every natural number greater than 1 factors uniquely (up to order) into a product of primes. In other words, if $n \geq 2$ and*

$$p_1 p_2 \cdots p_k = n = q_1 q_2 \cdots q_\ell$$

with $p_1 \leq p_2 \leq \cdots \leq p_k$ and $q_1 \leq q_2 \leq \cdots \leq q_\ell$ all primes, then $k = \ell$ and $p_i = q_i$ for $1 \leq i \leq k$.

Proof. Existence follows from above. We prove the result by (strong) induction on n. Suppose first that n is prime and

$$p_1 p_2 \cdots p_k = n = q_1 q_2 \cdots q_\ell$$

with $p_1 \leq p_2 \leq \cdots \leq p_k$ and $q_1 \leq q_2 \leq \cdots \leq q_\ell$ all primes. Notice that $p_i \mid n$ for all i and $q_j \mid n$ for all j. Since the only positive divisors of n are 1 and n, and 1 is not prime, we conclude that $p_i = n$ for all i and $q_j = n$ for all j. If $k \geq 2$, then $p_1 p_2 \cdots p_k \geq n^k > n$, a contradiction, so we must have $k = 1$. Similarly we must have $\ell = 1$. Thus, the result holds for all primes n. In particular, it holds if $n = 2$.

Suppose now that $n > 2$ is a natural number and the result is true for all smaller natural numbers. Suppose that

$$p_1 p_2 \cdots p_k = n = q_1 q_2 \cdots q_\ell$$

If n is prime, we are done by the above discussion. Suppose then that n is composite. In particular, we have $k \geq 2$ and $\ell \geq 2$. Now $p_1 \mid q_1 q_2 \cdots q_\ell$, so $p_1 \mid q_j$ for some j. Since q_j is prime and $p_1 \neq 1$, we must have $p_1 = q_j$. Similarly, we must have $q_1 = p_i$ for some i. We then have

$$p_1 = q_j \geq q_1 = p_i \geq p_1$$

hence all inequalities must be equalities, and we conclude that $p_1 = q_1$. Canceling, we get

$$p_2 \cdots p_k = q_2 \cdots q_\ell$$

and this common number is smaller than n. By induction, it follows that $k = \ell$ and $p_i = q_i$ for $2 \leq i \leq k$. □

Given a natural number $n \in \mathbb{N}$ with $n \geq 2$, when we write its prime factorization, we typically group together like primes and write

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$$

where the p_i are distinct primes. We often allow the insertion of "extra" primes in the factorization of n by permitting some α_i to equal to 0. This convention is particularly useful when comparing prime factorization of two numbers so that we can assume that both factorizations have the same primes occurring. It also allows us to write 1 in such a form by choosing all α_i to equal 0. Here is one example.

Proposition 2.5.9. *Suppose that $n, d \in \mathbb{N}^+$. Write the prime factorizations of n and d as*

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$$
$$d = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$$

where the p_i are distinct primes and possibly some α_i and β_j are 0. We then have that $d \mid n$ if and only if $0 \leq \beta_i \leq \alpha_i$ for all i.

Proof. Suppose first that $0 \leq \beta_i \leq \alpha_i$ for all i. We then have that $\alpha_i - \beta_i \geq 0$ for all i, so we may let

$$c = p_1^{\alpha_1 - \beta_1} p_2^{\alpha_2 - \beta_2} \cdots p_k^{\alpha_k - \beta_k} \in \mathbb{Z}$$

Notice that

$$\begin{aligned} dc &= p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k} \cdot p_1^{\alpha_1 - \beta_1} p_2^{\alpha_2 - \beta_2} \cdots p_k^{\alpha_k - \beta_k} \\ &= (p_1^{\beta_1} p_1^{\alpha_1 - \beta_1})(p_2^{\beta_2} p_2^{\alpha_2 - \beta_2}) \cdots (p_n^{\beta_n} p_n^{\alpha_1 - \beta_n}) \\ &= p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n} \\ &= n \end{aligned}$$

hence $d \mid n$.

Conversely, suppose that $d \mid n$ and fix $c \in \mathbb{Z}$ with $dc = n$. Notice that $c > 0$ because $d, n > 0$. Now we have $dc = n$, so $c \mid n$. If q is prime and $q \mid c$, then $q \mid n$ by transitively of divisibility so $q \mid p_i$ for some i by Corollary 2.5.7, and hence $q = p_i$ for some i because each p_i is prime. Thus, we can write the prime factorization of c as

$$c = p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k}$$

where again we may have some γ_i equal to 0. We then have

$$\begin{aligned} n &= dc \\ &= (p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k})(p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k}) \\ &= (p_1^{\beta_1} p_1^{\gamma_1})(p_2^{\beta_2} p_2^{\gamma_2}) \cdots (p_k^{\beta_k} p_k^{\gamma_k}) \\ &= p_1^{\beta_1 + \gamma_1} p_2^{\beta_2 + \gamma_2} \cdots p_k^{\beta_k + \gamma_k} \end{aligned}$$

By the Fundamental Theorem of Arithmetic, we have $\beta_i + \gamma_i = \alpha_i$ for all i. Since $\beta_i, \gamma_i, \alpha_i \geq 0$ for all i, we conclude that $\beta_i \leq \alpha_i$ for all i. □

Corollary 2.5.10. *Let $a, b \in \mathbb{N}^+$ with and write*

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$$
$$b = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$$

where the p_i are distinct primes. We then have

$$gcd(a, b) = p_1^{\min\{\alpha_1, \beta_1\}} p_2^{\min\{\alpha_2, \beta_2\}} \cdots p_k^{\min\{\alpha_k, \beta_k\}}$$

Corollary 2.5.11. *Suppose that $n > 1$ and $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ where the p_i are distinct primes. The number of nonnegative divisors of n is*

$$\prod_{i=1}^{k} (\alpha_i + 1)$$

and the number of integers divisors of n is

$$2 \cdot \prod_{i=1}^{k} (\alpha_i + 1)$$

Proof. We know that a nonnegative divisor d of n must factor as

$$d = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$$

where $0 \leq \beta_i \leq \alpha_i$ for all i. Thus, we have $\alpha_i + 1$ many choices for each β_i. Notice that different choices of β_i give rise to different values of d by the Fundamental Theorem of Arithmetic. □

Proposition 2.5.12. *Suppose that $m, n \in \mathbb{N}^+$, and write the prime factorization of m as $m = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ where the p_i are distinct primes. We then have that m is an n^{th} power in $\mathbb{N}$ if and only if $n \mid \alpha_i$ for all i.*

Proof. Suppose first that $n \mid \alpha_i$ for all i. For each i, fix β_i such that $\alpha_i = n\beta_i$. Since n and each α_i are nonnegative, it follows that each β_i is also nonnegative. Letting $\ell = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$, we then have

$$\ell^n = p_1^{n\beta_1} p_2^{n\beta_2} \cdots p_k^{n\beta_k} = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k} = m$$

so m is an n^{th} power in $\mathbb{N}$.

Suppose conversely that m is an n^{th} power in $\mathbb{N}$, and write $m = \ell^n$. Since $m > 1$, we have $\ell > 1$. Write the unique prime factorization of ℓ as

$$\ell = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$$

We then have

$$m = \ell^n = (p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k})^n = p_1^{n\beta_1} p_2^{n\beta_2} \cdots p_k^{n\beta_k}$$

By the Fundamental Theorem of Arithmetic, we have $\alpha_i = n\beta_i$ for all i, so $n \mid \alpha_i$ for all i. □

Theorem 2.5.13. *Let $m, n \in \mathbb{N}$ with $m, n \geq 2$. If the unique prime factorization of m does not have the property that every prime exponent is divisible by n, then $\sqrt[n]{m}$ is irrational.*

Proof. We proof the contrapostive. Suppose that $\sqrt[n]{m}$ is rational and write $\sqrt[n]{m} = \frac{a}{b}$ where $a, b \in \mathbb{Z}$. We may assume that $a, b > 0$ because $\sqrt[n]{m} > 0$. We then have

$$\frac{a^n}{b^n} = (\frac{a}{b})^n = q^n = m$$

hence

$$a^n = b^n m$$

Write a, b, m in their unique prime factorizations as

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$$
$$b = p_1^{\beta_1} p_2^{\beta_2} \cdots p_k^{\beta_k}$$
$$m = p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_k^{\gamma_k}$$

where the p_i are distinct (and possibly some $\alpha_i, \beta_i, \gamma_i$ are equal to 0). Since $a^n = b^n m$, we have

$$p_1^{n\alpha_1} p_2^{n\alpha_2} \cdots p_k^{n\alpha_k} = p_1^{n\beta_1 + \gamma_1} p_2^{n\beta_2 + \gamma_2} \cdots p_k^{n\beta_k + \gamma_k}$$

By the Fundamental Theorem of Arithmetic, we conclude that $n\alpha_i = n\beta_i + \gamma_i$ for all i. Therefore, for each i, we have $\gamma_i = n\alpha_i - n\beta_i = n(\alpha_i - \beta_i)$, and so $n \mid \gamma_i$ for each i. □

Chapter 3

Equivalence Relations and Modular Arithmetic

3.1 Equivalence Relations

Definition 3.1.1. *Let A be a set. A* relation *on A is a subset $R \subseteq A \times A$.*

Given a relation R on a set A, we typically use the same infix notation as in binary operations and write aRb rather than the more cumbersome $(a, b) \in R$.

Definition 3.1.2. *An* equivalence relation *on a set A is a relation $\sim$ on A such that*

- *$\sim$ is reflexive: $a \sim a$ for all $a \in A$.*
- *$\sim$ is symmetric: Whenever $a, b \in A$ satisfy $a \sim b$, we have $b \sim a$.*
- *$\sim$ is transitive: Whenever $a, b, c \in A$ satisfy $a \sim b$ and $b \sim c$, we have $a \sim c$.*

We begin with a couple of examples. In each of the following examples, it is straightforward to check that $\sim$ is an equivalence relation on A by checking each of the properties.

Example 3.1.3. *Let $A = \mathbb{R}$ where $a \sim b$ means $|a| = |b|$. We then have that $\sim$ is an equivalence relation on A.*

Example 3.1.4. *Let A be the set of all differentiable functions on R where $f(x) \sim g(x)$ means $f'(x) = g'(x)$ for all x. We then have that $\sim$ is an equivalence relation on A.*

We now consider a nonexample. Consider $A = \mathbb{Z}$ where $a \sim b$ means that $a \leq b$. Notice that $\sim$ is not an equivalence relation because $3 \leq 4$ but $4 \not\leq 3$, so $\sim$ is not symmetric. We now move on to some more interesting examples which we treat more carefully.

Example 3.1.5. *Let A be the set of all 2×2 matrices with real coefficients. Let $M \sim N$ mean that there exists an invertible 2×2 matrix P such that $M = PNP^{-1}$. We then have that $\sim$ is an equivalence relation on A.*

Proof. We need to check the three properties.

- Reflexive: Let $M \in A$. The 2×2 identity matrix I is invertible and satisfies $I^{-1} = I$, so we have $M = IMI^{-1}$. Therefore, $\sim$ is reflexive.

- Symmetric: Let $M, N \in A$ with $M \sim N$. Fix a 2×2 invertible matrix P with $M = PNP^{-1}$. Multiplying on the left by P^{-1} we get $P^{-1}M = NP^{-1}$, and now multiplying on the right by P we conclude that $P^{-1}MP = N$. We know from linear algebra that P^{-1} is also invertible and $(P^{-1})^{-1} = P$, so $N = P^{-1}M(P^{-1})^{-1}$ and hence $N \sim M$.

- Transitive: Let $L, M, N \in A$ with $L \sim M$ and $M \sim N$. Since $L \sim M$, we may fix a 2×2 invertible matrix P with $L = PMP^{-1}$. Since $M \sim N$, we may fix a 2×2 invertible matrix Q with $M = QNQ^{-1}$. We then have
$$L = PMP^{-1} = P(QNQ^{-1})P^{-1} = (PQ)N(Q^{-1}P^{-1})$$
Now by linear algebra, we know that the product of two invertible matrices is invertible, so PQ is invertible and furthermore we know that $(PQ)^{-1} = Q^{-1}P^{-1}$. Therefore, we have
$$L = (PQ)N(PQ)^{-1}$$
so $L \sim N$.

Putting it all together, we conclude that $\sim$ is an equivalence relation on A. □

Example 3.1.6. *Let A be the set $\mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$, i.e. A is the set of all pairs $(a, b) \in \mathbb{Z}^2$ with $b \neq 0$. Define a relation $\sim$ on A as follows. Given $a, b, c, d \in \mathbb{Z}$ with $b, d \neq 0$, we let $(a, b) \sim (c, d)$ mean $ad = bc$. We then have that $\sim$ is an equivalence relation on A.*

Proof. We check the three properties.

- Reflexive: Let $a, b \in \mathbb{Z}$ with $b \neq 0$. Since $ab = ba$, it follows that $(a, b) \sim (a, b)$.

- Symmetric: Let $a, b, c, d \in \mathbb{Z}$ with $b, d \neq 0$, and $(a, b) \sim (c, d)$. We then have that $ad = bc$. From this, we conclude that $cb = da$ so $(c, d) \sim (a, b)$.

- Transitive: Let $a, b, c, d, e, f \in \mathbb{Z}$ with $b, d, f \neq 0$ where $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. We then have that $ad = bc$ and $cf = de$. Multiplying the first equation by f we see that $adf = bcf$. Multiplying the second equation by b gives $bcf = bde$. Therefore, we know that $adf = bde$. Now $d \neq 0$ by assumption, so we may cancel it to conclude that $af = be$. It follows that $(a, b) \sim (e, f)$

Therefore, $\sim$ is an equivalence relation on A. □

Let's analyze the above situation more carefully. We have $(1, 2) \sim (2, 4)$, $(1, 2) \sim (4, 8)$, $(1, 2) \sim (-5, -10)$, etc. If you think of (a, b) as representing the fraction $\frac{a}{b}$, then the relation $(a, b) \sim (c, d)$ is saying exactly that the fractions $\frac{a}{c}$ and $\frac{b}{d}$ are equal. You probably never thought about equality of fractions as the result of imposing an equivalence relation on pairs of integers, but that is exactly what it is. We will be more precise about this below.

The next example is an important example in geometry. We introduce it now, and will return to it later.

Example 3.1.7. *Let A be the set $\mathbb{R}^2\backslash\{(0, 0)\}$. Define a relation $\sim$ on A by letting $(x_1, y_1) \sim (x_2, y_2)$ if there exists a real number $\lambda \neq 0$ with $(x_1, y_1) = (\lambda x_2, \lambda y_2)$. We then have that $\sim$ is an equivalence relation on A.*

Proof. We check the three properties.

- Reflexive: Let $(x, y) \in \mathbb{R}^2\backslash\{(0, 0)\}$ we have $(x, y) \sim (x, y)$ because using $\lambda = 1$ we see that $(x, y) = (1 \cdot x, 1 \cdot y)$. Therefore, $\sim$ is reflexive.

- Symmetric: Suppose now that $(x_1, y_1) \sim (x_2, y_2)$, and fix a nonzero real number λ such that $(x_1, y_1) = (\lambda x_2, \lambda y_1)$. We then have that $x_1 = \lambda x_2$ and $y_1 = \lambda y_2$, so $x_2 = \frac{1}{\lambda} \cdot x_1$ and $y_2 = \frac{1}{\lambda} \cdot y_1$ (notice that we are using $\lambda \neq 0$ so we can divide by it). Hence $(x_2, y_2) = (\frac{1}{\lambda} \cdot x_2, \frac{1}{\lambda} \cdot y_2)$, and so $(x_2, y_2) \sim (x_1, y_1)$. Therefore, $\sim$ is symmetric.

- Transitive: Suppose that $(x_1, y_1) \sim (x_2, y_2)$ and $(x_2, y_2) \sim (x_3, y_3)$. Fix a nonzero real number λ with $(x_1, y_1) = (\lambda x_2, \lambda y_2)$, and also fix a nonzero real number μ with $(x_2, y_2) = (\mu x_3, \mu y_3)$. We then have that $(x_1, y_1) = ((\lambda\mu)x_3, (\lambda\mu)y_3)$. Since both $\lambda \neq 0$ and $\mu \neq 0$, notice that $\lambda\mu \neq 0$ as well, so $(x_1, y_1) \sim (x_3, y_3)$. Therefore, $\sim$ is transitive.

Therefore, $\sim$ is an equivalence relation on A. □

Definition 3.1.8. *Let $\sim$ be an equivalence relation on a set A. Given $a \in A$, we let*

$$\overline{a} = \{b \in A : a \sim b\}$$

The set $\overline{a}$ is called the equivalence class *of a.*

Some sources use the notation $[a]$ instead of $\overline{a}$. This notation helps emphasize that the equivalence class of a is a *subset* of A rather than an element of A. However, it is cumbersome notation when you begin working with equivalence classes. We will stick with our notation, although it might take a little time to get used to. Notice that by the reflexive property of $\sim$, we have that $a \in \overline{a}$ for all $a \in A$.

For example, suppose we are working as above with $A = \mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$ where $(a, b) \sim (c, d)$ means that $ad = bc$. As discussed above, some elements of $\overline{(a, b)}$ are $(1, 2)$, $(2, 4)$, $(4, 8)$, $(-5, -10)$, etc. So

$$\overline{(a, b)} = \{(1, 2), (2, 4), (4, 8), (-5, -10), \dots\}$$

Again, I want to emphasize that $\overline{(a, b)}$ is a subset of A.

The following proposition is hugely fundamental. It says that if two equivalence classes overlap, then they must in fact be equal. In other words, if $\sim$ is an equivalence on A, then the equivalence classes *partition* the set A into pieces.

Proposition 3.1.9. *Let $\sim$ be an equivalence relation on a set A and let $a, b \in A$. If $\overline{a} \cap \overline{b} \neq \emptyset$, then $\overline{a} = \overline{b}$.*

Proof. Suppose that $\overline{a} \cap \overline{b} \neq \emptyset$. Fix $c \in \overline{a} \cap \overline{b}$. We then have $a \sim c$ and $b \sim c$. By symmetry, we know that $c \sim b$, and using transitivity we get that $a \sim b$. Using symmetry again, we conclude that $b \sim a$.

We first show that $\overline{a} \subseteq \overline{b}$. Let $x \in \overline{a}$. We then have that $a \sim x$. Since $b \sim a$, we can use transitivity to conclude that $b \sim x$, hence $x \in \overline{b}$.

We next show that $\overline{b} \subseteq \overline{a}$. Let $x \in \overline{b}$. We then have that $b \sim x$. Since $a \sim b$, we can use transitivity to conclude that $a \sim x$, hence $x \in \overline{a}$.

Putting this together, we get that $\overline{a} = \overline{b}$. □

With that proposition in hand, we are ready for the foundational theorem about equivalence relations.

Theorem 3.1.10. *Let $\sim$ be an equivalence relation on a set A and let $a, b \in A$.*

1. *$a \sim b$ if and only if $\overline{a} = \overline{b}$.*
2. *$a \not\sim b$ if and only if $\overline{a} \cap \overline{b} = \emptyset$.*

Proof. We first prove 1. Suppose first that $a \sim b$. We then have that $b \in \overline{a}$. Now we know that $b \sim b$ because $\sim$ is reflexive, so $b \in \overline{b}$. Thus, $b \in \overline{a} \cap \overline{b}$, so $\overline{a} \cap \overline{b} \neq \emptyset$. By the previous proposition, we conclude that $\overline{a} = \overline{b}$.

Suppose conversely that $\overline{a} = \overline{b}$. Since $b \sim b$ because $\sim$ is reflexive, we have that $b \in \overline{b}$. Therefore, $b \in \overline{a}$ and hence $a \sim b$.

We now use everything we've shown to get 2 with little effort. Suppose that $a \not\sim b$. Since we just proved 1, it follows that $\overline{a} \neq \overline{b}$, so by the previous proposition we must have $\overline{a} \cap \overline{b} = \emptyset$. Suppose conversely that $\overline{a} \cap \overline{b} = \emptyset$. We then have $\overline{a} \neq \overline{b}$ (because $a \in \overline{a}$ so $\overline{a} \neq \emptyset$), so $a \not\sim b$ by part 1. □

Let's revisit the example of $A = \mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$ where $(a, b) \sim (c, d)$ means $ad = bc$. The equivalence class of $(1, 2)$, namely the set $\overline{(1, 2)}$ is the set of all pairs of integers which are ways of representing the fraction $\frac{1}{2}$. In fact, this is how you "construct" the rational numbers from the integers. You simply *define* the rational numbers to be the set of equivalence classes of A under $\sim$. In other words, we let

$$\frac{a}{b} = \overline{(a, b)}$$

So when you write something like

$$\frac{1}{2} = \frac{4}{8}$$

you are simply saying that

$$\overline{(1, 2)} = \overline{(4, 8)}$$

which is true because $(1, 2) \sim (4, 8)$.

Example 3.1.11. *Recall the example above where $A = \mathbb{R}^2\backslash\{(0, 0)\}$ and where $(x_1, y_1) \sim (x_2, y_2)$ means that there exists a real number $\lambda \neq 0$ with $(x_1, y_1) = (\lambda x_2, \lambda y_2)$. The equivalence classes of $\sim$ are the lines through the origin (omitting the origin itself).*

Proof. Our first claim is that every point of A is equivalent to exactly one of the following points.

- $(0, 1)$
- $(1, m)$ for some $m \in \mathbb{R}$.

We first show that every point is equivalent to at least one of the above points. Suppose that $(x, y) \in A$ so $(x, y) \neq (0, 0)$. If $x = 0$, then we must have $y \neq 0$. Therefore $(x, y) \sim (0, 1)$ via $\lambda = y$. Now if $x \neq 0$, then $(x, y) \sim (1, \frac{y}{x})$ via $\lambda = x$. This gives existence.

To show uniqueness, it suffices to show that no two of the above points are equivalent to each other because we already know that $\sim$ is an equivalence relation. Suppose that $m \in \mathbb{R}$ and that $(0, 1) \sim (1, m)$. Fix $\lambda \in \mathbb{R}$ with $\lambda \neq 0$ such that $(0, 1) = (\lambda 1, \lambda m)$. Looking at the first coordinate, we conclude that $\lambda = 0$, a contradiction. Therefore, $(0, 1)$ is not equivalent to any point of the second type. Suppose now that $m, n \in \mathbb{R}$ with $(1, m) \sim (1, n)$. Fix $\lambda \in \mathbb{R}$ with $\lambda \neq 0$ such that $(1, m) = (\lambda 1, \lambda n)$. Looking at first coordinates, we must have $\lambda = 1$, so examining second coordinates gives $m = \lambda n = n$. Therefore $(1, m) \not\sim (1, n)$ whenever $m \neq n$. This finishes the claim.

Now we examine the equivalence classes of each of the above points. We first handle $\overline{(0, 1)}$ and claim that it equals the set of points in A on the line $x = 0$. Notice first that if $(x, y) \in \overline{(0, 1)}$, then $(0, 1) \sim (x, y)$, so fixing $\lambda \neq 0$ with $(0, 1) = (\lambda x, \lambda y)$ we conclude that $\lambda x = 0$ and hence $x = 0$. Thus, every element of $\overline{(0, 1)}$ is indeed on the line $x = 0$. Now taking an arbitrary point on the line $x = 0$, say $(0, y)$ with $y \neq 0$, we simply notice that $(0, 1) \sim (0, y)$ via $\lambda = \frac{1}{y}$. Hence, every point on the line $x = 0$ is an element of $\overline{(0, 1)}$.

Finally we fix $m \in \mathbb{R}$ and claim that $\overline{(1, m)}$ is the set of points in A on the line $y = mx$. Notice first that if $(x, y) \in \overline{(1, m)}$, then $(1, m) \sim (x, y)$, hence $(x, y) \sim (1, m)$. Fix $\lambda \neq 0$ with $(x, y) = (\lambda 1, \lambda m)$. We then have $x = \lambda$ by looking at first coordinates, so $y = \lambda m = mx$ by looking at second coordinates. Thus, every element of $\overline{(1, m)}$ lies on the line $y = mx$. Now take an arbitrary point in A on the line $y = mx$, say (x, mx). We then have that $x \neq 0$ because $(0, 0) \notin A$. Thus $(1, m) \sim (x, mx)$ via $\lambda = x$. Hence, every point on the line $y = mx$ is an element of $\overline{(1, m)}$. □

The set of equivalence classes of $\sim$ in the previous example is known as the *projective real line*.

3.2 Defining Functions on Equivalence Classes

Suppose that $\sim$ is a equivalence relation on the set A. When we look at the equivalence classes of A, we saw in the last section that we have shattered the set A into pieces. Just as in the case where we constructed the rationals, you often want to form a new set which consists of the equivalence classes themselves. Thus, the elements of this new set are themselves sets. Here is the formal definition.

Definition 3.2.1. *Let A be a set and let $\sim$ be an equivalence relation on A. The set whose elements are the equivalence classes of A under $\sim$ is called the* quotient *of A by $\sim$ and is denoted $A/\sim$.*

Thus, if let $A = \mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$ where $(a, b) \sim (c, d)$ means $ad = bc$, then the set of rations is the quotient $A/\sim$. Letting $Q = A/\sim$, we then have that the *set* $\overline{(a, b)}$ is an *element* of Q for every choice of $a, b \in \mathbb{Z}$ with $b \neq 0$. This quotient construction is extremely general, and we will see that it will play a fundamental role in our studies. Before we delve into our first main example of modular arithmetic in the next section, we first address an important and subtle question.

To begin with, notice that a given element of Q (namely a rational) is *represented* by many different pairs of integers. After all, we have

$$\frac{1}{2} = \overline{(1, 2)} = \overline{(2, 4)} = \overline{(-5, -10)} = \dots$$

Suppose that we want to define a function whose domain is Q. For example, we want to define a function $f\colon Q \to \mathbb{Z}$. Now we can try to write down something like:

$$f(\overline{(a, b)}) = a$$

Intuitively, we are trying to define $f\colon Q \to \mathbb{Z}$ by letting $f(\frac{a}{b}) = a$. From a naive glance, this might look perfectly reasonable. However, there is a real problem which arises from the fact that elements of Q have many representations. On the one hand, we should have

$$f(\overline{(1, 2)}) = 1$$

and on the other hand we should have

$$f(\overline{(2, 4)}) = 2$$

But we know that $\overline{(1, 2)} = \overline{(2, 4)}$, which contradicts the very definition of a function (after all, a function must have a unique output for any given input, but our description has imposed multiple different outputs for the same input). Thus, if we want to define a function on Q, we need to check that our definition does not depend on our choice of representatives.

For a positive example, consider the projective real line P. That is, let $A = \mathbb{R}^2\backslash\{(0, 0\}$ where $(x_1, y_1) \sim (x_2, y_2)$ means there there exists a real number $\lambda \neq 0$ such that $(x_1, y_1) = (\lambda x_2, \lambda y_2)$. We then have have that $P = A/\sim$. Consider trying to define the function $g\colon P \to \mathbb{R}$ by

$$g(\overline{(x, y)}) = \frac{5xy}{x^2 + y^2}$$

First we check a technicality: If $\overline{(x, y)} \in P$, then $(x, y) \neq (0, 0)$, so $x^2 + y^2 \neq 0$, and hence the domain of g really is all of P. Now we claim that g "makes sense", i.e. that it actually is a function. To see this, we take two elements (x_1, y_1) and (x_2, y_2) with $\overline{(x_1, y_1)} = \overline{(x_2, y_2)}$, and check that $g(\overline{(x_1, y_1)}) = g(\overline{(x_2, y_2)})$. In other words, we check that our definition of g does not actually depend on our choice of representative. Suppose then that $\overline{(x_1, y_1)} = \overline{(x_2, y_2)}$. We then have $(x_1, y_1) \sim (x_2, y_2)$, so we may fix $\lambda \neq 0$ with $(x_1, y_1) = (\lambda x_2, \lambda y_2)$.

Now

$$\begin{aligned} g(\overline{(x_2, y_2)}) &= \frac{5x_2y_2}{x_2^2 + y_2^2} \\ &= \frac{5(\lambda x_1)(\lambda y_1)}{(\lambda x_1)^2 + (\lambda y_1)^2} \\ &= \frac{5\lambda^2 x_1 y_1}{\lambda^2 x_1^2 + \lambda^2 y_1^2} \\ &= \frac{\lambda^2 \cdot 5x_1y_1}{\lambda^2 \cdot (x_1^2 + y_1^2)} \\ &= \frac{5x_1y_1}{x_1^2 + y_1^2} \\ &= g(\overline{(x_1, y_1)}) \end{aligned}$$

Hence, g does indeed make sense as a function on P.

Now technically we are being sloppy above when we start by defining a "function" and only checking after the fact that it makes sense and actually results in an honest function. However, this shortcut is completely standard. The process of checking that a "function" $f\colon A/\!\sim \to X$ defined via representatives of equivalence classes is independent of the actual representatives chosen, and so actually makes sense, is called checking that f is *well-defined*.

For a final example, let's consider a function from a quotient to another quotient. Going back to Q, we define a function $f\colon Q \to Q$ as follows:

$$f(\overline{(a, b)}) = \overline{(a^2 + 3b^2, 2b^2)}$$

We claim that this function is well-defined on Q. Intuitively, we want to define the following function on fractions:

$$f(\frac{a}{b}) = \frac{a^2 + 3b^2}{2b^2}$$

Let's check it that it does indeed make sense. Suppose that $a, b, c, d \in \mathbb{Z}$ with $b, d \neq 0$ and we have $\overline{(a, b)} = \overline{(c, d)}$, i.e. that $(a, b) \sim (c, d)$. We need to show that $f(\overline{(a, b)}) = f(\overline{(c, d)})$, i.e. that $\overline{(a^2 + 3b^2, 2b^2)} = \overline{(c^2 + 3d^2, 2d^2)}$ or equivalently that $(a^2+3b^2, 2b^2) \sim (c^2+3d^2, 2d^2)$. Since we are assuming that $(a, b) \sim (c, d)$, we know that $ad = bc$. Hence

$$\begin{aligned} (a^2 + 3b^2) \cdot 2d^2 &= 2a^2d^2 + 6b^2d^2 \\ &= 2(ad)^2 + 6b^2d^2 \\ &= 2(bc)^2 + 6b^2d^2 \\ &= 2b^2c^2 + 6b^2d^2 \\ &= 2b^2 \cdot (c^2 + 3d^2) \end{aligned}$$

Therefore, $(a^2 + 3b^2, 2b^2) \sim (c^2 + 3d^2, 2d^2)$, which is to say that $f(\overline{(a, b)}) = f(\overline{(c, d)})$. It follows that f is well-defined on Q.

3.3 Modular Arithmetic

Definition 3.3.1. *Let $n \in \mathbb{N}^+$. We define a relation $\equiv_n$ on $\mathbb{Z}$ by letting $a \equiv_n b$ mean that $n \mid (a - b)$. When $a \equiv_n b$ we say that a is congruent to b modulo n.*

Proposition 3.3.2. *Let $n \in \mathbb{N}^+$. The relation $\equiv_n$ is an equivalence relation on $\mathbb{Z}$.*

Proof. We need to check the three properties.

- Reflexive: Let $a \in \mathbb{Z}$. Since $a - a = 0$ and $n \mid 0$, we have that $n \mid (a - a)$, hence $a \equiv_n a$.
- Symmetric: Let $a, b \in \mathbb{Z}$ with $a \equiv_n b$. We then have that $n \mid (a - b)$. Thus $n \mid (-1)(a - b)$, which says that $n \mid (b - a)$, and so $b \equiv_n a$.
- Transitive: Let $a, b, c \in \mathbb{Z}$ with $a \equiv_n b$ and $b \equiv_n c$. We then have that $n \mid (a - b)$ and $n \mid (b - c)$. It follows that $n \mid [(a - b) + (b - c)]$, which is to say that $n \mid (a - c)$. Therefore, $a \equiv_n c$.

Putting it all together, we conclude that $\equiv_n$ is an equivalence relation on $\mathbb{Z}$. □

By our general theory about equivalence relations, we know that $\equiv_n$ partitions $\mathbb{Z}$ into equivalence classes. We next determine the number of such equivalence classes.

Proposition 3.3.3. *Let $n \in \mathbb{N}^+$ and let $a \in \mathbb{Z}$. There exists a unique $b \in \{0, 1, \ldots, n-1\}$ such that $a \equiv_n b$. In fact, if we write $a = qn + r$ for the unique choice of $q, r \in \mathbb{Z}$ with $0 \leq r < n$, then $b = r$.*

Proof. As in the statment, fix $q, r \in \mathbb{Z}$ with $a = qn + r$ and $0 \leq r < n$. We then have $a - r = nq$, so $n \mid (a - r)$. It follows that $a \equiv_n r$, so we have proven existence.

Suppose now that $b \in \{0, 1, \ldots, n-1\}$ and $a \equiv_n b$. We then have that $n \mid (a - b)$, so we may fix $k \in \mathbb{Z}$ with $nk = a - b$. This gives $a = kn + b$. Since $0 \leq b < n$, we may use the uniqueness part of Theorem 2.3.1 to conclude that $k = q$ (which is unneccesary) and $b = r$. □

Therefore, the quotient $\mathbb{Z}/\equiv_n$ has n elements, and we can obtain unique representatives for these equivalence classes by taking the representatives from the set $\{0, 1, \ldots, n-1\}$. For example, if $n = 5$, we have that $\mathbb{Z}/\equiv_n$ consists of the following five sets.

- $\overline{0} = \{\ldots, -10, -5, 0, 5, 10, \ldots\}$
- $\overline{1} = \{\ldots, -9, -4, 1, 6, 11, \ldots\}$
- $\overline{2} = \{\ldots, -8, -3, 2, 7, 12, \ldots\}$
- $\overline{3} = \{\ldots, -7, -2, 3, 8, 13, \ldots\}$
- $\overline{4} = \{\ldots, -6, -1, 4, 9, 14, \ldots\}$

Now that we've used $\equiv_n$ to break $\mathbb{Z}$ up into n pieces, our next goal is to show how to add and multiply elements of this quotient. The idea is to define addition/multiplication of elements of $\mathbb{Z}/\equiv_n$ by simply adding/multiplying representatives. In other words, we would like to define

$$\overline{a} + \overline{b} = \overline{a + b} \qquad \text{and} \qquad \overline{a} \cdot \overline{b} = \overline{a \cdot b}$$

Of course, whenever you define functions on equivalence classes via representatives, you need to be careful to ensure that your function is well-defined, i.e. that it does not depend of the choice of representatives. That is the content of the next result.

Proposition 3.3.4. *Suppose that $a, b, c, d \in \mathbb{Z}$ with $a \equiv_n c$ and $b \equiv_n d$. We then have*

1. *$a + b \equiv_n c + d$*
2. *$ab \equiv_n cd$*

Proof. Since $a \equiv_n c$ and $b \equiv_n d$, we have $n \mid (a - c)$ and $n \mid (b - d)$.

1. Notice that
$$(a+b)-(c+d)=(a-c)+(b-d)$$
Since $n \mid (a-c)$ and $n \mid (b-d)$, it follows that $n \mid [(a-c)+(b-d)]$ and so $n \mid [(a+b)-(c+d)]$. Therefore, $a+b \equiv_n c+d$.

2. Notice that
$$\begin{aligned} ab-cd &= ab-bc+bc-cd \\ &= (a-c)\cdot b+(b-d)\cdot c \end{aligned}$$
Since $n \mid (a-c)$ and $n \mid (b-d)$, it follows that $n \mid [(a-c)\cdot b+(b-d)\cdot c]$ and so $n \mid (ab-cd)$. Therefore, $ab \equiv_n cd$.

□

3.4 The Groups $\mathbb{Z}/n\mathbb{Z}$ and $U(\mathbb{Z}/n\mathbb{Z})$

Proposition 3.4.1. *Let $n \in \mathbb{N}^+$ and let $G = \mathbb{Z}/\equiv_n$, i.e. the elements of G are the equivalence classes of $\mathbb{Z}$ under the equivalence relation $\equiv_n$. Define a binary operation $+$ on G by letting $\overline{a}+\overline{b}=\overline{a+b}$. We then have that $(G,+,\overline{0})$ is an abelian group.*

Proof. We have already shown in Proposition 3.3.4 that $+$ is well-defined on G. With that in hand, the definition makes it immediate that $+$ is a binary operation on G. We now check the axioms for an abelian group.

- Associative: For any $a,b,c \in \mathbb{Z}$ we have
$$\begin{aligned} (\overline{a}+\overline{b})+\overline{c} &= (\overline{a+b})+\overline{c} \\ &= \overline{(a+b)+c} \\ &= \overline{a+(b+c)} && \text{(since } + \text{ is associative on } \mathbb{Z}) \\ &= \overline{a}+\overline{b+c} \\ &= \overline{a}+(\overline{b}+\overline{c}) \end{aligned}$$
so $+$ is associative on G.

- Identity: For any $a \in \mathbb{Z}$ we have
$$\overline{a}+\overline{0}=\overline{a+0}=\overline{a}$$
and
$$\overline{0}+\overline{a}=\overline{0+a}=\overline{a}$$

- Inverses: For any $a \in \mathbb{Z}$ we have
$$\overline{a}+\overline{-a}=\overline{a+(-a)}=\overline{0}$$
and
$$\overline{-a}+\overline{a}=\overline{(-a)+a}=\overline{0}$$

- Commutative: For any $a,b \in \mathbb{Z}$ we have
$$\begin{aligned} \overline{a}+\overline{b} &= \overline{a+b} \\ &= \overline{b+a} && \text{(since } + \text{ is commutative on } \mathbb{Z}) \\ &= \overline{b}+\overline{a} \end{aligned}$$
so $+$ is commutative on G.

Therefore, $(G, +, \overline{0})$ is an abelian group. □

Definition 3.4.2. *Let $n \in \mathbb{N}^+$. We denote the above abelian group by $\mathbb{Z}/n\mathbb{Z}$. We call the group "$\mathbb{Z}$ mod $n\mathbb{Z}$".*

This notation is unmotivated at the moment, but will make sense when we discuss general quotient groups (and be consistent with the more general notation we establish there). Let's examine one example in details. Consider $\mathbb{Z}/5\mathbb{Z}$. As discussed in the last section, the equivalence classes $\overline{0}$, $\overline{1}$, $\overline{2}$, $\overline{3}$ and $\overline{4}$ are all distinct and give all elements of $\mathbb{Z}/5\mathbb{Z}$. By definition, we have $\overline{3} + \overline{4} = \overline{7}$. This is perfectly correct, but 7 is not one of the special representatives we chose above. Since $\overline{7} = \overline{2}$, we can also write $\overline{3} + \overline{4} = \overline{2}$ and now we have removed mention of all representatives other than the chosen ones. Working it all out with only those representatives, we get the following Cayley table for $\mathbb{Z}/5\mathbb{Z}$:

$+$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{0}$	$\overline{1}$
$\overline{3}$	$\overline{3}$	$\overline{4}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{4}$	$\overline{4}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$

We also showed in Proposition 3.3.4 that multiplication is well-defined on the quotient. Now it is straightforward to mimic the above arguments to show that multiplication is associative, commutative, and that $\overline{1}$ is an identity. However, it seems unlikely that we always have inverse because $\mathbb{Z}$ itself fails to have multiplicative inverses for all elements other than ± 1. But let's look at what happens in the case $n = 5$. For example, we have $\overline{4} \cdot \overline{4} = \overline{16} = \overline{1}$, so in particular $\overline{4}$ does have a multiplicative inverse. Working out the computations, we get the following table.

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{1}$	$\overline{3}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{1}$	$\overline{4}$	$\overline{2}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

Examining the table, we are pleasantly surprised that every element other that $\overline{0}$ does have a multiplicative inverse! In hindsight, there was no hope of $\overline{0}$ having a multiplicative inverse because $\overline{0} \cdot \overline{a} = \overline{0 \cdot a} = \overline{0} \neq \overline{1}$ for all $a \in \mathbb{Z}$. With this one example you might have the innocent hope that you always get multiplicative inverses for elements other $\overline{0}$ when you change n. Let's dash those hopes now by looking at the case when $n = 6$.

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{0}$	$\overline{2}$	$\overline{4}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{2}$	$\overline{0}$	$\overline{4}$	$\overline{2}$
$\overline{5}$	$\overline{0}$	$\overline{5}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

Looking at the table, we see that only $\overline{1}$ and $\overline{5}$ have multiplicative inverses. There are other curiosities in the above table. For example, you can have two nonzero elements whose product is zero as shown by $\overline{3} \cdot \overline{4} = \overline{0}$. This is an interesting fact, and will return to such considerations when we get to ring theory.

However, let's get to our primary concern of forming a group under multiplication. The idea is simply to "trim down" to the elements which do happen to have inverses. In order to do this, it would be helpful to have a decent characterization of such elements.

Proposition 3.4.3. *Let $n \in \mathbb{N}^+$ and let $a \in \mathbb{Z}$. The following are equivalent.*

1. *There exists $c \in \mathbb{Z}$ with $ac \equiv_n 1$.*
2. *$\gcd(a, n) = 1$.*

Proof. We first suppose that there exists $c \in \mathbb{Z}$ with $ac \equiv_n 1$. Fix such a $c \in \mathbb{Z}$. We then have $n \mid (ac - 1)$, so we may fix $k \in \mathbb{Z}$ with $nk = ac - 1$. Rearranging, we see that $ac + n(-k) = 1$. Hence, there is an integer combination of a and n which gives 1. Since $\gcd(a, n)$ is the least positive such integer, and there is no positive integer less that 1, we conclude that $\gcd(a, n) = 1$.

Suppose conversely that $\gcd(a, n) = 1$. Fix $k, \ell \in \mathbb{Z}$ with $ak + n\ell = 1$. Rearranging gives $n(-\ell) = ak - 1$, so $n \mid (ak - 1)$. It follows that $ak \equiv_n 1$, so we choose $c = k$. □

Thus, in the case $n = 6$, the fundamental reason why $\overline{1}$ and $\overline{5}$ have multiplicative inverses is that $\gcd(1, 6) = 1 = \gcd(5, 6)$. In the case of $n = 5$, the reason why every element other than $\overline{0}$ had a multiplicative inverse is because every possible number less than 5 is relatively prime with 5, which is essentially just saying that 5 is prime. Notice that the above argument can be turned into an explicit algorithm for finding such a c as follows. Given $n \in \mathbb{N}^+$ and $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$, use the Euclidean algorithm to produce $k, \ell \in \mathbb{Z}$ with $ak + n\ell = 1$. As the argument shows, you can then choose $c = k$.

As a consequence of the above result and the fact that multiplication is well-defined (Proposition 3.3.4), we see that if $a \equiv_n b$, then $\gcd(a, n) = 1$ if and only if $\gcd(b, n) = 1$. Of course we could have proved this without this result. Suppose that $a \equiv_n b$. We then have that $n \mid (a - b)$, so we may fix $k \in \mathbb{Z}$ with $nk = a - b$. This gives $a = kn + b$ so $\gcd(a, n) = \gcd(b, n)$ by Corollary 2.4.6.

Now that we know which elements have multiplicative inverses, the idea is simply to slim down to these elements to form a group.

Proposition 3.4.4. *Let $n \in \mathbb{N}^+$ and let $G = \mathbb{Z}/\equiv_n$, i.e. G the elements of G are the equivalence classes of $\mathbb{Z}$ under the equivalence relation $\equiv_n$. Let U be the following subset of G:*

$$U = \{\overline{a} : a \in \mathbb{Z} \text{ and } \gcd(a, n) = 1\}$$

Define a binary operation $\cdot$ on U by letting $\overline{a} \cdot \overline{b} = \overline{a \cdot b}$. We then have that $(U, \cdot, \overline{1})$ is an abelian group.

Proof. We have already shown in Proposition 3.3.4 that $\cdot$ is well-defined on G. However, before we dive into checking the axioms, there is one important fact that we need to check, which is whether $\cdot$ is a binary operation on U, i.e. if you multiply two elements of U do you get another element of U. Suppose then that $a, b \in \mathbb{Z}$ and that $\gcd(a, n) = 1 = \gcd(b, n)$. By the above characterization, we may fix $c, d \in \mathbb{Z}$ with $ac \equiv_n 1$ and $bd \equiv_n 1$. It follows that $acbd \equiv_n 1$, hence $(ab)(cd) \equiv_n 1$. Therefore, by the above characterization in the opposite direction, we conclude that $\gcd(ab, n) = 1$ (for those paying attention, you proved directly on homework that if $\gcd(a, n) = 1 = \gcd(b, n)$, then $\gcd(ab, n) = 1$). Therefore, the operation $\cdot$ is well-defined on U.

With the binary operation business taken care of, we can move on to the group axioms. The proof that $\cdot$ is associative and commutative on U follows exactly as in the case of $\mathbb{Z}/n\mathbb{Z}$ using the fact that $\cdot$ is associative and commutative on $\mathbb{Z}$. Also, $\overline{1}$ is an identity because 1 is a multiplicative identity on $\mathbb{Z}$.

Finally, we need to check that every element has an inverse. Suppose that $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$ so that $\overline{a} \in U$. By the above characterization, we may fix $c \in \mathbb{Z}$ with $ac \equiv_n 1$. We then have

$$\overline{a} \cdot \overline{c} = \overline{ac} = \overline{1}$$

and

$$\overline{c} \cdot \overline{a} = \overline{ca} = \overline{1}$$

The last little thing that we need to check is that $\overline{c} \in U$. However, since $ac \equiv_n 1$, we have $ca \equiv_n 1$, so $\gcd(c, n) = 1$ by our characterization and hence $\overline{c} \in U$. Therefore, every element of U has an inverse in U. □

Definition 3.4.5. *Let $n \in \mathbb{N}^+$. We denote the above abelian group by $U(\mathbb{Z}/n\mathbb{Z})$.*

Again, this notation may seems a bit odd, but we will come back and explain it in context when we get to ring theory. To see some examples of Cayley tables of these groups, here is the Cayley table of $U(\mathbb{Z}/5\mathbb{Z})$:

$\cdot$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{2}$	$\overline{2}$	$\overline{4}$	$\overline{1}$	$\overline{3}$
$\overline{3}$	$\overline{3}$	$\overline{1}$	$\overline{4}$	$\overline{2}$
$\overline{4}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

and here is the Cayley table of $U(\mathbb{Z}/6\mathbb{Z})$:

$\cdot$	$\overline{1}$	$\overline{5}$
$\overline{1}$	$\overline{1}$	$\overline{5}$
$\overline{5}$	$\overline{5}$	$\overline{1}$

Finally, we give the Cayley table of $U(\mathbb{Z}/8\mathbb{Z})$ because we will return to this group later as an important example:

$\cdot$	$\overline{1}$	$\overline{3}$	$\overline{5}$	$\overline{7}$
$\overline{1}$	$\overline{1}$	$\overline{3}$	$\overline{5}$	$\overline{7}$
$\overline{3}$	$\overline{3}$	$\overline{1}$	$\overline{7}$	$\overline{5}$
$\overline{5}$	$\overline{5}$	$\overline{7}$	$\overline{1}$	$\overline{3}$
$\overline{7}$	$\overline{7}$	$\overline{5}$	$\overline{3}$	$\overline{1}$

Chapter 4

Subgroups, Cyclic Groups, and Generation

4.1 Notation and Conventions

With a few examples of groups in hand, we now dive into the general theory. Recall the definition of a group.

Definition 4.1.1. *A* group *is a set G equipped with a binary operation $\cdot$ and an element $e \in G$ such that*

1. *(Associativity) For all $a, b, c \in G$, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.*
2. *(Identity) For all $a \in G$, we have $a \cdot e = a$ and $e \cdot a = a$.*
3. *(Inverses) For all $a \in G$, there exists $b \in G$ with $a \cdot b = e$ and $b \cdot a = e$.*

We now establish some notation. First, we often just say "Let G be a group" rather than the more precise "Let $(G, \cdot, e)$ be a group". In other words, we typically do not explicit mention the binary operation and will just write $\cdot$ afterward if we feel the need to explicitly mention it. The identity element is unique as shown in Chapter 1, so we don't feel the need to be so explicit about it and will just call it e later if we need to refer to it. If some confusion may arises (for example, there are several natural binary operations on the given set), then we will need to be explicit.

Let G be a group. If $a, b \in G$, we typically write ab rather than explicitly writing the binary operation in $a \cdot b$. Of course, sometimes it makes sense to insert the dot. For example, if G contains the integers 2 and 4 and 24, writing 24 would certainly be confusing. Also, if the group operation is standard addition or some other operation with a conventional symbol, we will switch up and use that symbol to avoid confusion. However, when there is no confusion, and when we are dealing with an abstract group without explicit mention of what the operation actually is, we will tend to omit it.

With this convention in hand, associativity tells us that $(ab)c = a(bc)$ for all $a, b, c \in G$. Now using associativity repeatedly we obtain what can be called "generalized associativity". For example, on the first homework you showed that $(ab)(cd) = (a(bc))d$ for all $a, b, c, d \in G$ by using associativity a few times. In general, such rearrangements are always possible by iteratively moving the parentheses around as long as the order of the elements doesn't change. In other words, no matter how we insert parentheses in $abcd$ so that it makes sense, you always obtain the same result. For a sequence of 4 elements it is straightforward to try them all, and for 5 elements it is tedious but feasible. However, it is true no matter how long the sequence is. A careful proof of this fact requires careful definitions of what a "permissible insertion of parentheses" means and at this level such an involved tangent is more distracting from our primary aims that it is enlightening. We will simply take the result as true.

Keep in mind that the order of the elements occurring does matter. There is no reason at all to think that *abcd* equals *dacb* unless the group is abelian. However, if G is abelian then upon any insertion of parentheses into these expressions so that they make sense, they will evaluate to the same value. Thus, assuming that G is commutative, you obtain a kind of "generalized commutativity" just like we have a "generalized associativity".

Definition 4.1.2. *Let G be a group. If G is a finite set, then the order of G, denoted $|G|$, is the number of elements in G. If G is infinite, we simply write $|G| = \infty$.*

For example, we have $|\mathbb{Z}/n\mathbb{Z}| = n$ for all $n \in \mathbb{N}^+$. Thus, we have shown that there exists a group (in fact an abelian group) of every order. We now provide some notations for dealing with the group $U(\mathbb{Z}/n\mathbb{Z})$,

Definition 4.1.3. *We define a function $\varphi\colon \mathbb{N}^+ \to \mathbb{N}^+$ as follows. For each $n \in \mathbb{N}^+$, we let*

$$\varphi(n) = |\{a \in \{1, 2, 3, \ldots, n\} : \gcd(a, n) = 1\}|$$

The function φ is called the Euler φ-function or Euler totient function.

Directly from our definition of $U(\mathbb{Z}/n\mathbb{Z})$, we see that $|U(\mathbb{Z}/n\mathbb{Z})| = \varphi(n)$ for all $n \in \mathbb{N}^+$. Calculating $\varphi(n)$ is a nontrivial task, although it is straightforward if you know the prime factorization of n (consult your local number theory course). Notice that $\varphi(p) = p - 1$ for all primes p, so $|U(\mathbb{Z}/p\mathbb{Z})| = p - 1$ for all primes p.

We now have a decent supply of groups at our disposal. We just discussed the examples $\mathbb{Z}/n\mathbb{Z}$ and $U(\mathbb{Z}/n\mathbb{Z})$ for each $n \in \mathbb{N}^+$. From Chapter 1 we have the groups $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{R}, +)$. We also have $(\mathbb{Q}\backslash\{0\}, \cdot)$ and $(\mathbb{R}\backslash\{0\}, \cdot)$, together with a few other isolated examples. Now all of the groups just mentioned are abelian. Our primary example of nonabelian groups at the moment are the following (we discussed the case $n = 2$ in Chapter 1).

Definition 4.1.4. *Let $n \in \mathbb{N}^+$. The set of $n \times n$ invertible matrices under matrix multiplication is a group. We denote it by $GL_n(\mathbb{R})$ and call it the* general linear group *of degree n.*

For each $n \geq 2$, the group $GL_n(\mathbb{R})$ is nonabelian so we have an infinite family of such groups (it is worthwhile to explicitly construct two $n \times n$ invertible matrices which do not commute with each other). Now at the moment, we haven't completely verified the existence of finite nonabelian groups, but we will do so in the next chapter.

4.2 Orders of Elements

Definition 4.2.1. *Let G be a group and let $a \in G$. For each $n \in \mathbb{N}^+$, we define*

$$a^n = aaa \cdots a$$

where there are n total a's in the above product. If you want to be more formal, we define a^n recursively by letting $a^1 = a$ and $a^{n+1} = a^n a$ for all $n \in \mathbb{N}^+$. We also define

$$a^0 = e$$

Finally, we extend the notation a^n for $n \in \mathbb{Z}$ as follows. If $n < 0$, we let

$$a^n = (a^{-1})^{|n|}$$

It other words, if $n < 0$, we let

$$a^n = a^{-1}a^{-1}a^{-1} \cdots a^{-1}$$

where there are n total (a^{-1})'s in the above product.

Let G be a group and let $a \in G$. Notice that

$$a^3a^2 = (aaa)(aa) = aaaaa = a^5$$

and

$$(a^3)^2 = a^3a^3 = (aaa)(aaa) = aaaaaa = a^6$$

For other examples, notice that

$$a^5a^{-2} = (aaaaa)(a^{-1}a^{-1}) = aaaaaa^{-1}a^{-1} = aaaaa^{-1} = aaa = a^3$$

and for any n we have

$$(a^4)^{-1} = (aaaa)^{-1} = a^{-1}a^{-1}a^{-1}a^{-1} = (a^{-1})^4 = a^{-4}$$

In general, we have the following. A precise proof would involve induction using the formal recursive definition of a^n given above. However, just like the above discussion, these special cases explain where they come from and such a formal argument adds little enlightenment so will be omitted.

Proposition 4.2.2. *Let G be a group. Let $a \in G$ and let $m, n \in \mathbb{Z}$. We have*

1. $a^{m+n} = a^m a^n$
2. $a^{mn} = (a^m)^n$

In some particular groups, this notation is very confusing if used in practice. For example, when working in the group $(\mathbb{Z}, +, 0)$, we would have $2^4 = 2 + 2 + 2 + 2 = 8$ because the operation is addition rather than multiplication. However, with the standard operation of exponentiation on $\mathbb{Z}$ we have $2^4 = 16$. Make sure you understand what operation is in use whenever we use that notation a^n in a given group. It might be better to use a different notation if confusion can arise.

Definition 4.2.3. *Let G be a group and let $a \in G$. We define the* order *of a as follows. Let*

$$S = \{n \in \mathbb{N}^+ : a^n = e\}$$

If $S \neq \emptyset$, we let $|a| = \min(S)$ (which exists by well-ordering), and if $S = \emptyset$, we define $|a| = \infty$. In other words, $|a|$ is the least positive n such that $a^n = e$ provided such an n exists.

The reason why we choose to overload the word *order* for two apparently very different concepts (when applied to a group versus when applied to an element of a group) will be explained in the next section on subgroups.

Example 4.2.4. *Here are some examples of computing orders of elements in a group.*

- *In any group G, we have $|e| = 1$.*
- *In the group $(\mathbb{Z}, +)$, we have $|0| = 1$ as noted, but $|n| = \infty$ for all $n \neq 0$.*
- *In the group $\mathbb{Z}/n\mathbb{Z}$, we have $|\overline{1}| = n$.*
- *In the group $\mathbb{Z}/12\mathbb{Z}$, we have $|\overline{9}| = 4$ because*

$$\overline{9}^2 = \overline{2 \cdot 9} = \overline{18} = \overline{6} \qquad \overline{9}^3 = \overline{3 \cdot 9} = \overline{27} = \overline{3} \qquad \overline{9}^4 = \overline{4 \cdot 9} = \overline{36} = \overline{0}$$

- *In the group $U(\mathbb{Z}/7\mathbb{Z})$ we have $|\overline{4}| = 3$ because*

$$\overline{4}^2 = \overline{16} = \overline{2} \qquad \overline{4}^3 = \overline{64} = \overline{1}$$

The order of an element $a \in G$ is the *least* positive $m \in \mathbb{Z}$ with $a^m = e$. There may be many other larger positive powers of a that give the identity, or even negative powers which do. For example, consider $\overline{1} \in \mathbb{Z}/2\mathbb{Z}$. We have $|\overline{1}| = 2$, but $\overline{1}^4 = \overline{0}$, $\overline{1}^6 = \overline{0}$, and $\overline{1}^{-14} = \overline{0}$. In general, give the order of an element, we can characterize all powers of that element which give the identity as follows.

Proposition 4.2.5. *Let G be a group and let $a \in G$.*

1. *Suppose that $|a| = m \in \mathbb{N}^+$. For any $n \in \mathbb{Z}$, we have $a^n = e$ if and only if $m \mid n$.*
2. *Suppose that $|a| = \infty$. For any $n \in \mathbb{Z}$, we have $a^n = e$ if and only if $n = 0$.*

Proof. We first prove 1. Let $|a| = m \in \mathbb{Z}$. We then have in particular that $a^m = e$. Suppose first that $n \in \mathbb{Z}$ is such that $m \mid n$. Fix $k \in \mathbb{Z}$ with $n = mk$ We then have

$$a^n = a^{mk} = (a^m)^k = e^k = e$$

so $a^n = e$. Suppose conversely that $n \in \mathbb{Z}$ and that $a^n = e$. Since $m > 0$, we may write $n = qm + r$ where $0 \leq r < m$. We then have

$$\begin{aligned} e &= a^n \\ &= a^{qm+r} \\ &= a^{qm}a^r \\ &= (a^m)^q a^r \\ &= e^q a^r \\ &= a^r \end{aligned}$$

Now by definition we know that m is the least positive power of a which gives the identity. Therefore, since $0 \leq r < m$ and $a^r = e$, we must have that $r = 0$. It follows that $n = qm$ so $m \mid n$.

We now prove 2. Suppose that $|a| = \infty$. If $n = 0$, then we have $a^n = a^0 = e$. If $n > 0$, then we have $a^n \neq e$ because by definition no positive power of a equals the identity. Suppose then that $n < 0$ and that $a^n = e$. We then have

$$e = e^{-1} = (a^n)^{-1} = a^{-n}$$

Now $-n > 0$ because $n < 0$, but this is a contradiction because no positive power of a gives the identity. It follows that if $n < 0$, then $a^n \neq e$. Therefore, $a^n = e$ if and only if $n = 0$. □

Suppose now that we have an element a of a group G and we know its order. How do we compute the orders of the power of G? For an example, consider $\overline{3} \in \mathbb{Z}/30\mathbb{Z}$. It is straightforward to check that $|\overline{3}| = 10$. Let's look at the orders of some powers of $\overline{3}$. Now $\overline{3}^2 = \overline{6}$ and a simple check shows that $|\overline{6}| = 5$ so $|\overline{3}^2| = 5$. Now consider $\overline{3}^3 = \overline{9}$. We have $\overline{9}^2 = \overline{18}$, then $\overline{9}^3 = \overline{27}$, and then $\overline{9}^4 = \overline{36} = \overline{6}$. Since 30 is not a multiple of 9 we see that we "cycle around" and it is not quite as clear when we will hit $\overline{0}$ as we continue. However, if you keep at it, you will find that $|\overline{3}^3| = 10$. If you keep calculating away, you will find that $|\overline{3}^4| = 5$ and $|\overline{3}^5| = 2$. We would like to have a better way to determine these values without resorting to tedious calculations, and that is what the next proposition supplies.

Proposition 4.2.6. *Let G be a group and let $a \in G$.*

1. *Suppose that $|a| = m \in \mathbb{N}^+$. For any $n \in \mathbb{Z}$, we have*

$$|a^n| = \frac{m}{\gcd(m, n)}$$

2. *Suppose that* $|a| = \infty$. *We have* $|a^n| = \infty$ *for all* $n \in \mathbb{Z}\backslash\{0\}$.

Proof. We first prove 1. Fix $n \in \mathbb{Z}$ and let $d = \gcd(m, n)$. Since d is a common divisor of m and n, we may fix $s, t \in \mathbb{Z}$ with $m = ds$ and $n = dt$. Notice that $s > 0$ because both $m > 0$ and $d > 0$. With this notation, we need to show that $|a^n| = s$.

Notice that

$$(a^n)^s = a^{ns} = a^{dts} = a^{mt} = (a^m)^t = e^t = e$$

Thus, s is a positive power of a which gives the identity, and so $|a| \leq s$.

Suppose now that $k \in \mathbb{N}^+$ with $(a^n)^k = e$. We need to show that $s \leq k$. We have $a^{nk} = e$, so by the previous proposition we know that $m \mid nk$. Fix $\ell \in \mathbb{Z}$ with $m\ell = nk$. We then have that $ds\ell = dtk$, so canceling $d > 0$ we conclude that $s\ell = tk$, and hence $s \mid tk$. Now by the homework problem about least common multiples, we know that $\gcd(s, t) = 1$. Using Proposition 2.4.10, we conclude that $s \mid k$. Since $s, k > 0$, it follows that $s \leq k$.

Therefore, s is the least positive value of k such that $(a^n)^k = e$, and so we conclude that $|a^k| = s$.

We now prove 2. Suppose that $n \in \mathbb{Z}\backslash\{0\}$. Let $k \in \mathbb{N}^+$. We then have $(a^n)^k = a^{nk}$. Now $nk \neq 0$ because $n \neq 0$ and $k > 0$, hence $(a^n)^k = a^{nk} \neq 0$ by the previous proposition. Therefore, $(a^n)^k \neq 0$ for all $k \in \mathbb{N}^+$ and hence $|a^n| = \infty$. □

Corollary 4.2.7. *Let* $n \in \mathbb{N}^+$ *and let* $k \in \mathbb{Z}$. *In the group* $\mathbb{Z}/n\mathbb{Z}$, *we have*

$$|\overline{k}| = \frac{n}{\gcd(k, n)}$$

Proof. Working in the group $\mathbb{Z}/n\mathbb{Z}$, we know that $|\overline{1}| = n$. Now $\overline{k} = \overline{1}^k$, so the result follows by the previous proposition. □

4.3 Subgroups

If we have a group, we can consider subsets of G which also happen to be a group under the same operation. If we take a subset of a group and restrict the group operation to that subset, we trivially have the operation is associative on H because it is associative on G. The only issues are whether the operation remains a binary operation on H (it could conceivably combine two elements of H and return an element not in H), whether the identity is there, and whether the inverse of every element of H is also in H. This gives the following definition.

Definition 4.3.1. *Let* G *be a group and let* $H \subseteq G$. *We say that* H *is a* subgroup *of* G *if*

1. $e \in H$ *(*H *contains the identity of* G*)*
2. $ab \in H$ *whenever* $a \in H$ *and* $b \in H$ *(*H *is closed under the group operation)*
3. $a^{-1} \in H$ *whenever* $a \in H$ *(*H *is closed under inverses)*

Example 4.3.2. *Here are some examples of subgroups.*

- *For any group* G, *we also have two trivial examples of subgroups. Namely* G *is always a subgroup of itself, and* $\{e\}$ *is always a subgroup of* G.
- $\mathbb{Z}$ *is a subgroup of* $(\mathbb{Q}, +)$ *and* $(\mathbb{R}, +)$. *This follows because* $0 \in \mathbb{Z}$, *the sum of two integers is an integer, and the additive inverse of an integer is an integer.*
- *The set* $2\mathbb{Z} = \{2n : n \in \mathbb{Z}\}$ *is a subgroup of* $(\mathbb{Z}, +)$. *This follows from the fact that* $0 = 2 \cdot 0 \in 2\mathbb{Z}$, *that* $2m + 2n = 2(m + n)$ *so the sum of two evens is even, and that* $-(2n) = 2(-n)$ *so the the additive inverse of an even number is even.*

- *The set $H = \{\overline{0}, \overline{3}\}$ is a subgroup of $\mathbb{Z}/6\mathbb{Z}$. To check that H is a subgroup, we need to check the three conditions. We have $\overline{0} \in H$, so H contains the identity. Also $\overline{0}^{-1} = \overline{0} \in H$ and $\overline{3}^{-1} = \overline{3} \in H$, so the inverse of each element of H lies in H. Finally, to check that H is closed under the group operation, we simply have to check the four possibilities. For example, we have $\overline{3} + \overline{3} = \overline{6} = \overline{0} \in H$. The other 3 possible sums are even easier.*

Example 4.3.3. *Let $G = (\mathbb{Z}, +)$. Here are some examples of subsets of $\mathbb{Z}$ which are* **not** *subgroups of G.*

- *The set $H = \{2n + 1 : n \in \mathbb{Z}\}$ is not a subgroup of G because $0 \notin H$.*
- *The set $H = \{0\} \cup \{2n + 1 : n \in \mathbb{Z}\}$ is not a subgroup of G because even though it contains 0 and is closed under inverses, it is not closed under the group operation. For example, $1 \in H$ and $3 \in H$, but $1 + 3 \notin H$.*
- *The set $\mathbb{N}$ is not a subgroup of G because even though it contains 0 and is closed under the group operation, it is not closed under inverse. For example, $1 \in H$ but $-1 \notin H$*

Now it is possible that a subset of a group G forms a group under a completely different binary operation that the one used in G, but whenever we talk about a subgroup H of G we only think of restricting the group operation of G down to H. For example, let $G = (\mathbb{Q}, +)$. The set $H = \mathbb{Q}\backslash\{0\}$ is *not* a subgroup of G (since it does not contain the identity) even though H can be made into with the completely different operation of multiplication. When we consider a subset H of a group G, we only call it a subgroup of G if it is a group with respect to the exact same binary operation.

Proposition 4.3.4. *Let $n \in \mathbb{N}^+$. The set $H = \{A \in GL_n(\mathbb{R}) : \det(A) = 1\}$ is a subgroup of $GL_n(\mathbb{R})$.*

Proof. Letting I_n be the $n \times n$ identity matrix (which is the identity of $GL_n(\mathbb{R})$, we have that $\det(I_n) = 1$ so $I_n \in H$. Suppose that $M, N \in H$ so $\det(M) = 1 = \det(N)$. We then have

$$\det(MN) = \det(M) \cdot \det(N) = 1 \cdot 1 = 1$$

so $MN \in H$. Suppose finally that $M \in H$. We have $MM^{-1} = I_n$, so $\det(MM^{-1}) = \det(I_n) = 1$, hence $\det(M) \cdot \det(M^{-1}) = 1$. Since $M \in H$ we have $\det(M) = 1$, hence $\det(M^{-1}) = 1$. It follows that $M^{-1} \in H$. We have checked the three properties of a subgroup, so H is a subgroup of G. □

Definition 4.3.5. *Let $n \in \mathbb{N}^+$. We let $SL_n(\mathbb{R})$ be the above subgroup of $GL_n(\mathbb{R})$. That is,*

$$SL_n(\mathbb{R}) = \{A \in GL_n(\mathbb{R}) : \det(A) = 1\}$$

The group $SL_n(\mathbb{R})$ is called the special linear group *of degree n.*

The following proposition occasionally makes the process of checking whether a subset of a group is indeed a subgroup a bit easier.

Proposition 4.3.6. *Let G be a group and let $H \subseteq G$. The following are equivalent:*

- *H is a subgroup of G*
- *$H \neq \emptyset$ and $ab^{-1} \in H$ whenever $a, b \in H$.*

Proof. Suppose first that H is a subgroup of G. By definition, we must have $e \in H$, so $H \neq \emptyset$. Suppose that $a, b \in H$. Since H is a subgroup and $b \in H$, we must have $b^{-1} \in H$. Now using the fact that $a \in H$ and $b^{-1} \in H$, together with the second part of the definition of a subgroup, it follows that $ab^{-1} \in H$.

Now suppose conversely that $H \neq \emptyset$ and $ab^{-1} \in H$ whenever $a, b \in H$. We need to check the three defining characteristics of a subgroup. Since $H \neq \emptyset$, we may fix $c \in H$. Using our condition and the fact that

$c \in H$, it follows that $e = cc^{-1} \in H$, so we have checked the first property. Now using the fact that $e \in H$, given any $a \in H$ we have $a^{-1} = ea^{-1} \in H$ by our condition, so we have checked the third property. Suppose now that $a, b \in H$. From what we just showed, we know that $b^{-1} \in H$. Therefore, using our condition, we conclude that $ab = a(b^{-1})^{-1} \in H$, so we have verified the second property. We have shown that all 3 properties hold for H, so H is a subgroup of G. □

Definition 4.3.7. *Let G be a group and let $c \in G$. We define*

$$\langle c \rangle = \{c^n : n \in \mathbb{Z}\}$$

The set $\langle c \rangle$ is called the subgroup of G generated by c.

The next proposition explains our choice of definition for $\langle c \rangle$. Namely, the set $\langle c \rangle$ is the smallest subgroup of G containing c as an element.

Proposition 4.3.8. *Let G be a group and let $c \in G$. Let $H = \langle c \rangle$.*

1. *H is a subgroup of G with $c \in H$.*
2. *If K is a subgroup of G with $c \in K$, then $H \subseteq K$.*

Proof. We first prove 1. We begin by noting that $c = c^1 \in H$. In particular, we have that $H \neq \emptyset$. Suppose now that $a, b \in H$ and fix $m, n \in \mathbb{Z}$ with $a = c^m$ and $b = c^n$. We then have that

$$ab^{-1} = c^m(c^n)^{-1} = c^m c^{-n} = c^{m-n} \in H$$

so $ab^{-1} \in H$. Therefore, H is indeed a subgroup of G by the preceding proposition.

We now prove 2. Suppose that K is a subgroup of G with $c \in K$. We first prove by induction on $n \in \mathbb{N}^+$ that $c^n \in K$. We clearly have $c^1 = c \in K$ by assumption. Suppose that $n \in \mathbb{N}^+$ and we know that $c^n \in K$. Since $c^n \in K$ and $c \in K$, and K is a subgroup of G, it follows that $c^{n+1} = c^n c \in K$. Therefore, by induction, we know that $c^n \in K$ for all $n \in \mathbb{N}^+$. Now $c^0 = e \in K$ because K is a subgroup of G, so $c^n \in K$ for all $n \in \mathbb{N}$. Finally, if $n \in \mathbb{Z}$ with $n < 0$, then $c^{-n} \in K$ because $-n \in \mathbb{N}^+$ and hence $c^n = (c^{-n})^{-1} \in K$ because inverses of elements of K must be in K. Therefore, $c^n \in K$ for all $n \in \mathbb{Z}$, which is to say that $H \subseteq K$. □

This next proposition finally justifies our overloading of the word *order*. Given a group G and an element $c \in G$, it says that $|c|$ (the order of c as an element of G) equals $|\langle c \rangle|$ (the number of elements in the subgroup of G generated by c).

Proposition 4.3.9. *Suppose that G is a group and that $c \in G$. Let $H = \langle c \rangle$.*

1. *Suppose that $|c| = m$. We then have that $H = \{c^i : 0 \leq i < m\} = \{e, c, c^2, \ldots, c^{m-1}\}$ and $c^k \neq c^\ell$ whenever $0 \leq k < \ell < m$. Thus, $|H| = m$*
2. *Suppose that $|c| = \infty$. We then have that $c^k \neq c^\ell$ whenever $k, \ell \in \mathbb{Z}$ with $k < \ell$, so $|H| = \infty$*

In particular, we have $|c| = |\langle c \rangle|$.

Proof. We first prove 1. By definition we have $H = \{c^n : n \in \mathbb{Z}\}$, so we trivially have that

$$\{c^i : 0 \leq i < m\} \subseteq H$$

Now let $n \in \mathbb{Z}$. Write $n = qm + r$ where $0 \leq r < n$. We then have

$$c^n = c^{qm+r} = (c^m)^q c^r = e^1 c^r = c^r$$

so $c^n = c^r \in \{c^i : 0 \leq i < m\}$. Therefore, $H \subseteq \{c^i : 0 \leq i < m\}$ and combining this with the reverse inclusion above we conclude that $H = \{c^i : 0 \leq i < m\}$.

Suppose now that $0 \leq k < \ell < m$. Assume for the sake of obtaining a contradiction that $c^k = c^\ell$. Multiplying both sides by c^{-k} on the right, we see that $c^k c^{-k} = c^\ell c^{-k}$, hence

$$e = c^0 = c^{k-k} = c^k c^{-k} = c^\ell c^{-k} = c^{\ell-k}$$

Now we have $0 \leq k < \ell < m$, so $0 < \ell - k < m$. This contradicts the assumption that $m = |c|$ is the least positive power of c giving the identity. Hence, we must have $c^k \neq c^\ell$.

We now prove 2. Suppose that $k, \ell \in \mathbb{Z}$ with $k < \ell$. Assume that $c^k = c^\ell$. As in part 1 we can multiply both sides on the right by c^{-k} to conclude that $c^{\ell-k} = e$. Now $\ell - k > 0$, so this contradicts the assumption that $|c| = \infty$. Therefore, we must have $c^k \neq c^\ell$. □

Corollary 4.3.10. *If G is a finite group, then every element of G has finite order. Moreover, for each $a \in G$, we have $|a| \leq |G|$.*

Proof. Let $a \in G$. We then have that $\langle a \rangle \subseteq G$, so $|\langle a \rangle| \leq |G|$. The result follows because $|a| = |\langle a \rangle|$. □

In fact, much more is true. We will see as a consequence of Lagrange's Theorem that the order of every element of finite group G is actually a divisor of $|G|$.

4.4 Cyclic Groups

Definition 4.4.1. *A group G is cyclic if there exists $c \in G$ such that $G = \langle c \rangle$. An element $c \in G$ with $G = \langle c \rangle$ is called a* generator *of G.*

For example, for each $n \in \mathbb{N}^+$, the group $\mathbb{Z}/n\mathbb{Z}$ is cyclic because $\overline{1}$ is a generator (since $\overline{1}^k = \overline{k}$ for all k). Also, $\mathbb{Z}$ is cyclic because 1 is a generator (remember than $\langle c \rangle$ is the set of all powers of c, both positive and negative). In general, a cyclic group has many generators. For example, -1 is a generator of $\mathbb{Z}$, and $\overline{3}$ is a generator of $\mathbb{Z}/4\mathbb{Z}$.

Proposition 4.4.2. *Let G be a finite group with $|G| = n$. An element $c \in G$ is a generator of G if and only if $|c| = n$. In particular, G is cyclic if and only if it has an element of order n.*

Proof. Suppose first that c is a generator of G so that $G = \langle c \rangle$. We know from above that $|c| = |\langle c \rangle|$, so $|c| = |G| = n$. Suppose conversely that $|c| = n$. We then know that $|\langle c \rangle| = n$. Since $\langle c \rangle \subseteq G$ and each has n elements, we must have that $G = \langle c \rangle$. □

Proposition 4.4.3. *All cyclic groups are abelian.*

Proof. Suppose that G is a cyclic group and fix $c \in G$ with $G = \langle c \rangle$. Let $a, b \in G$. Since $G = \langle c \rangle$, we may fix $m, n \in \mathbb{Z}$ with $a = c^m$ and $b = c^n$. We then have

$$ab = c^m c^n = c^{m+n} = c^{n+m} = c^n c^m = ba$$

Therefore, $ab = ba$ for all $a, b \in G$, so G is abelian. □

The converse of the preceding proposition is false. In other words, there exist abelian groups which are not cyclic. For example, consider the group $U(\mathbb{Z}/8\mathbb{Z})$. We have $U(\mathbb{Z}/8\mathbb{Z}) = \{\overline{1}, \overline{3}, \overline{5}, \overline{7}\}$, so $|U(\mathbb{Z}/8\mathbb{Z})| = 4$, but every nonidentity element has order 2. For another example, the infinite abelian group $(\mathbb{Q}, +)$ is also not cyclic: We clearly have $\langle 0 \rangle \neq \mathbb{Q}$, and if $q \neq 0$, then $\frac{q}{2} \notin \{nq : n \in \mathbb{Z}\} = \langle q \rangle$.

Proposition 4.4.4. *Let G be a cyclic group.*

- *If $|G| = n \in \mathbb{N}^+$, then G has exactly $\varphi(n)$ distinct generators.*
- *If $|G| = \infty$, then G has exactly 2 generators.*

Proof. If $|G| = 1$, then G has exactly $\varphi(1) = 1$ generator, so the result is true. Suppose then that G is a finite cyclic group with $|G| \geq 2$, and fix $c \in G$ with $G = \langle c \rangle$. We have

$$G = \{e, c, c^2, \ldots, c^{n-1}\}$$

and we know that $c^k \neq c^\ell$ whenever $0 \leq k < \ell < n$. Thus, we need only determine how many of these elements are generators. Suppose that $0 \leq k < n$. We then know that c^k is a generator of G if and only if $|c^k| = n$. Now for any $k \in \mathbb{Z}$, Proposition 4.2.6 tells us that

$$|c^k| = \frac{n}{\gcd(n, k)}$$

so c^k is a generator of G if and only if $\frac{n}{\gcd(n,k)} = n$, which is if and only if $\gcd(k, n) = 1$. Since $\varphi(n)$ is the number of integers less than or equal to n which are relatively prime to n, it follows that there are $\varphi(n)$ many such elements.

Suppose finally that $|G| = \infty$ and fix $c \in G$ with $G = \langle c \rangle$. It is straightforward to check that both c and c^{-1} are generators, and that $c \notin \langle c^n \rangle$ for any $n \in \mathbb{Z} \backslash \{1, -1\}$. □

We now completely determine the subgroup structure of a cyclic group. We begin with the following important fact.

Proposition 4.4.5. *Every subgroup of a cyclic group is cyclic.*

Proof. Suppose that G is a cyclic group and that H is a subgroup of G. Fix a generator c of G so that $G = \langle c \rangle$. If $H = \{e\}$ then $H = \langle e \rangle$, so H is cyclic. Suppose then that $H \neq \{e\}$. Since $H \neq \{e\}$, there exists $n \in \mathbb{Z} \backslash \{0\}$ such that $c^n \in H$. Notice that we also have $c^{-n} = (c^n)^{-1} \in H$ because H is a subgroup of G, so

$$\{n \in \mathbb{N}^+ : c^n \in H\} \neq \emptyset$$

Let m be the least element of this set (which exists by well-ordering). We show that $H = \langle c^m \rangle$. Since $c^m \in H$ and H is a subgroup of G, we know that $\langle c^m \rangle \subseteq H$ by Proposition 4.3.8. Suppose that $a \in H$. Since $H \subseteq G$, we also have $a \in G$, so we may fix $k \in \mathbb{Z}$ with $a = c^k$. Write $k = qm + r$ where $0 \leq r < m$. We then have

$$a = c^k = c^{qm+r} = c^{mq} c^r$$

Hence

$$c^r = (c^{mq})^{-1} \cdot a = c^{-mq} \cdot a = (c^m)^{-q} \cdot a$$

Now $c^m \in H$ and $a \in H$. Since H is a subgroup of G, we know that it is closed under inverses and the group operation, hence $(c^m)^{-q} \cdot a \in H$ and so $c^r \in H$. By choice of m as the smallest positive power of c which lies in H, we conclude that we must have $r = 0$. Therefore, $a = c^k = c^{qm} = (c^m)^q \in \langle c^m \rangle$. Since $a \in H$ was arbitrary, it follows that $H \subseteq \langle c^m \rangle$. Combining this with the reverse containment above, we conclude that $H = \langle c^m \rangle$, so H is cyclic. □

Lemma 4.4.6. *Let G be a cyclic with $|G| = n \in \mathbb{N}^+$ and let c be a generator of G. Let $m \in \mathbb{Z}$ and set $d = \gcd(m, n)$. We then have that $\langle c^m \rangle = \langle c^d \rangle$.*

Proof. Since $d \mid m$, we may fix $t \in \mathbb{Z}$ with $m = dt$. We then have

$$c^m = c^{dt} = (c^d)^t$$

Therefore, $c^m \in \langle c^d \rangle$ and hence $\langle c^m \rangle \subseteq \langle c^d \rangle$. We now prove the reverse containment. Since $d = \gcd(m, n)$, we may fix $k, \ell \in \mathbb{Z}$ with $d = mk + n\ell$. Now

$$c^d = c^{mk+n\ell} = c^{mk}c^{n\ell} = (c^m)^k(c^n)^\ell = (c^m)^k e^\ell = (c^m)^k$$

so $c^d \in \langle c^m \rangle$. It follows that $\langle c^d \rangle \subseteq \langle c^m \rangle$. Combing the two containments, we see that $\langle c^m \rangle = \langle c^d \rangle$. □

Theorem 4.4.7. *Let G be a cyclic group and fix $c \in G$ with $G = \langle c \rangle$.*

- *Suppose that $|G| = n \in \mathbb{N}^+$. Every subgroup of G has order dividing n. Furthermore, for each positive divisor $d \mid n$, there exists a unique subgroup of G of order d, namely $\langle c^{n/d} \rangle$.*
- *Suppose that $|G| = \infty$. Every subgroup of G equals $\langle c^n \rangle$ for some $n \in \mathbb{N}$, and each of these subgroups are distinct.*

Proof. Suppose first that $|G| = n \in \mathbb{N}^+$. Let $d \in \mathbb{N}^+$ be such that $d \mid n$. We then have that $\frac{n}{d} \in \mathbb{Z}$ and $\frac{n}{d} \mid n$. Therefore

$$|c^{n/d}| = \frac{n}{\gcd(n, n/d)} = \frac{n}{n/d} = d$$

so the subgroup $\langle c^{n/d} \rangle$ has order d. This proves the existence part of the result.

Suppose now that H is a an subgroup of G. Since G is cyclic, we know by Proposition 4.4.5 that H is cyclic, so we may fix $m \in \mathbb{Z}$ with $H = \langle c^m \rangle$. Letting $k = \gcd(m, n)$, we know from the previous lemma that $H = \langle c^m \rangle = \langle c^k \rangle$. Now $k \mid n$, so we may fix $k \in \mathbb{Z}$ with $n = dk$. Then $k = \frac{n}{d}$ and hence $H = \langle c^{n/d} \rangle$. As above, it follows that $|H| = d$. Thus, we have taken an arbitrary subgroup of G and shown that it has order d for a divisor of n, and in fact that $H = \langle c^{n/d} \rangle$. The result follows.

Suppose now that $|G| = \infty$. Let H be a subgroup of G. By Proposition 4.4.5, we know that H is cyclic, so we may fix $n \in \mathbb{Z}$ with $H = \langle c^n \rangle$. If $n < 0$, then notice that $\langle c^n \rangle = \langle c^{-n} \rangle$ because $c^{-n} = (c^n)^{-1} \in \langle c^n \rangle$ and $c^n = (c^{-n})^{-1} \in \langle c^{-n} \rangle$. Thus every subgroup of G equals $\langle c^n \rangle$ for some $n \in \mathbb{N}$.

Suppose now that $m, n \in \mathbb{N}$ with $m < n$. We claim that $\langle c^m \rangle \neq \langle c^n \rangle$. Now if $m = 0$, then $\langle c^m \rangle = \{e\}$ so $c^n \neq e$ is an element of $\langle c^n \rangle$ which is not an element of $\langle c^m \rangle$. Suppose then that $0 < m < n$. Assume for the sake of obtaining a contradiction that $c^m \in \langle c^n \rangle$. We may then fix $k \in \mathbb{Z}$ with $c^m = (c^n)^k = c^{nk}$. By Proposition 4.3.9, we must have $m = nk$, hence $n \mid m$. However, we have $0 < m < n$, so this is a contradiction. Therefore, we have that c^m is an element of $\langle c^m \rangle$ which is not an element of $\langle c^n \rangle$. It follows that the subgroups $\langle c^n \rangle$ for $n \in \mathbb{N}$ are all distinct. □

Corollary 4.4.8. *The subgroups of $\mathbb{Z}$ are precisely $n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$ for each $n \in \mathbb{N}$.*

4.5 Generating Subgroups

Suppose that G is a group. We have seen how to take an element $c \in G$ and form the smallest subgroup of G containing c by simply taking the set $\{c^n : n \in \mathbb{Z}\}$. Suppose now that you have many elements of G and want to form the smallest subgroup of G which contains them. For definiteness now, suppose that you have $a, b \in G$. How would you construct the smallest possible subgroup of G containing both a and b, which we denote by $\langle a, b \rangle$? A natural guess would be

$$\{a^m b^n : m, n \in \mathbb{Z}\}$$

but unless G is abelian there is no reason to think that this set is closed under multiplication. For example, we must have $aba \in \langle a, b \rangle$, but it doesn't obviously appear there. Whatever $\langle a, b \rangle$ is, it must contain the following elements:

$$abababa \qquad a^{-1}ba^{-1}b^{-1}a \qquad a^3b^6a^{-2}b^7$$

If you take 3 elements a, b, c, it gets even more complicated because we can alternate the 3 elements in such sequences without such a repetitive pattern. If you have infinitely many elements, it gets even worse. Since the constructions get messy, we define the subgroup generated by an arbitrary set in a much less explicit manner.

Definition 4.5.1. *Let G be a group and let $A \subseteq G$. We define $\langle A \rangle$ to be the intersection of all subgroups of G which contain A. If $A = \{a_1, a_2, \ldots, a_n\}$, we write $\langle a_1, a_2, \ldots, a_n \rangle$ rather than $\langle \{a_1, a_2, \ldots, a_n\} \rangle$.*

Since the intersection of subgroups is always a subgroup, we have that $\langle A \rangle$ is indeed a subgroup of G, and in fact it equals it is the smallest such subgroup explicitly from the definition. We will discuss this in more detail as we proceed and see some examples.

Chapter 5

Functions, Symmetric Groups, and Permutation Groups

5.1 Injections, Surjections, and Bijections

Definition 5.1.1. *Suppose that $f\colon A \to B$ and $g\colon B \to C$ are functions. The* composition *of g and f, denoted $g \circ f$, is the function $g \circ f\colon A \to C$ defined by $(g \circ f)(a) = g(f(a))$ for all $a \in A$.*

We begin with the following simple proposition which suggests that composition of functions might form the binary operation of some group.

Proposition 5.1.2. *Let A, B, C, D be sets. Suppose that $f\colon A \to B$, that $g\colon B \to C$, and that $h\colon C \to D$ are functions. We then have that $(h \circ g) \circ f = h \circ (g \circ f)$. Stated more simply, function composition is associative whenever it is defined.*

Proof. Let $a \in A$. We then have

$$\begin{aligned}((h \circ g) \circ f)(a) &= (h \circ g)(f(a)) \\ &= h(g(f(a)) \\ &= h((g \circ f)(a)) \\ &= (h \circ (g \circ f))(a)\end{aligned}$$

Therefore $((h \circ g) \circ f)(a) = (h \circ (g \circ f))(a)$ for all $a \in A$. It follows that $(h \circ g) \circ f = h \circ (g \circ f)$. □

We now possess a natural operation which is associative, which is great news because so far we have not come across very many natural operations (aside from addition and multiplication of numbers) which were associative. It is even more interesting because this binary operation is very much noncommutative in general. For example, if $f\colon \mathbb{R} \to \mathbb{R}$ is $f(x) = \sin x$ and $g\colon \mathbb{R} \to \mathbb{R}$ is $g(x) = x^2$, then

$$(f \circ g)(x) = f(g(x)) = f(x^2) = \sin(x^2)$$

while

$$(g \circ f)(x) = g(f(x)) = g(\sin x) = (\sin x)^2$$

so $f \circ g \neq g \circ f$. It looks like we're in a great position to build nonabelian groups. Before we get ahead of ourselves, we need to think about identities and inverses.

Definition 5.1.3. *Let A be a set. The function $id_A\colon A \to A$ defined by $id_A(a) = a$ for all $a \in A$ is called the* identity function *on A.*

The identity function does leave other functions alone when you compose with it. However, you have to be careful that you are composing with the identity function on the correct set.

Proposition 5.1.4. *For any function $f\colon A \to B$, we have $f \circ id_A = f$ and $id_B \circ f = f$.*

Proof. Let $f\colon A \to B$. For any $a \in A$, we have

$$(f \circ id_A)(a) = f(id_A(a)) = f(a)$$

Since $a \in A$ was arbitrary, it follows that $f \circ id_A = f$. For any $b \in B$, we have

$$(id_B \circ f)(a) = id_B(f(a)) = f(a)$$

because $f(a)$ is some element in B. Since $b \in B$ was arbitrary, it follows that $id_B \circ f = f$. □

Now that we have some notion of what will serve as an identity, we can move on to think about inverses.

Definition 5.1.5. *Let $f\colon A \to B$ be a function. An* inverse *of f is a function $g\colon B \to A$ such that* **both** *$g \circ f = id_A$ and $f \circ g = id_B$.*

Notice that we need inverses on both sides and the identity functions differ if $A \neq B$. Now consider the function $f\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) = e^x$. Does f have an inverse? As defined, we are asking whether there exists a function $g\colon \mathbb{R} \to \mathbb{R}$ such that both $g \circ f = id_{\mathbb{R}}$ and $f \circ g = id_{\mathbb{R}}$. A natural guess after a calculus course would be to consider the function $h\colon \mathbb{R}^+ \to \mathbb{R}$ given by $h(x) = \ln x$. Now h is *not* an inverse of f because it is not even defined for all real numbers as would be required. Suppose that you try to correct this minor defect by instead considering the function $g\colon \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} \ln x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now at least g is defined on all of $\mathbb{R}$. Is g an inverse of f? Let's check the definitions. For any $x \in \mathbb{R}$, we have $e^x > 0$, so

$$(g \circ f)(x) = g(f(x)) = g(e^x) = \ln(e^x) = x = id_{\mathbb{R}}(x)$$

so we have shown that $g \circ f = id_{\mathbb{R}}$. Let's examine $f \circ g$. Now if $x > 0$ then we have

$$(f \circ g)(x) = f(g(x)) = f(\ln x) = e^{\ln x} = x = id_{\mathbb{R}}(x)$$

so everything is good there. However, if $x < 0$, then

$$(f \circ g)(x) = f(g(x)) = f(0) = e^0 = 1 \neq x$$

Therefore, g is *not* an inverse of f, but merely a "left inverse". In fact, there is no function $g\colon \mathbb{R} \to \mathbb{R}$ which serves as a "right inverse" for f, i.e. for which $f \circ g = id_{\mathbb{R}}$.

Notice that if we had instead considered as f as a function from $\mathbb{R}$ to $\mathbb{R}^+$, then f does have an inverse. In fact, the function $h\colon \mathbb{R}^+ \to \mathbb{R}$ defined by $h(x) = \ln x$ does satisfy $h \circ f = id_{\mathbb{R}}$ and $f \circ h = id_{\mathbb{R}^+}$. Thus, when restricting the target set, you can obtain an inverse for a function which did not have one originally.

To understand these concepts more deeply, recall the following fundamental definitions.

Definition 5.1.6. *Let $f\colon A \to B$ be a function.*

- *We say that f is injective (or one-to-one) if whenever $f(a_1) = f(a_2)$ we have $a_1 = a_2$.*

- *We say that f is a surjective (or onto) if for all $b \in B$ there exists $a \in A$ such that $f(a) = b$. In other words, $range(f) = B$ where*

$$range(f) = \{b \in B : \text{There exists } a \in A \text{ with } f(a) = b\}$$

- *We say that f is a bijection if both f is an injection and a surjection.*

An equivalent condition for f to be injective is obtained by simply taking the contrapositive, i.e. $f\colon A \to B$ is injective if and only if whenever $a_1 \neq a_2$, we have $f(a_1) \neq f(a_2)$. Stated in more colloquial language, f is injective if every element of B is hit by at most one element of a via f. In this manner, f is surjective if every element of B is hit by at least one element of a via f, and f is bijective if every element of B is hit by exactly one element of a via f.

If you want to prove that a function $f\colon A \to B$ is injective, it is usually better to use our official definition than the contrapositive one with negations. Thus, you want to start by assuming that you are given $a_1, a_2 \in A$ which satisfy $f(a_1) = f(a_2)$, and using this assumption you want to prove that $a_1 = a_2$. The reason why this approach is often preferable is because it is typically easier to work with and manipulate a statement involving equality than it is to derive statements from a non-equality.

Proposition 5.1.7. *Suppose that $f\colon A \to B$ and $g\colon B \to C$ are both functions.*

1. *If both f and g are injective, then $g \circ f$ is injective.*
2. *If both f and g are surjective, then $g \circ f$ is surjective.*
3. *If both f and g are bijective, then $g \circ f$ is bijective.*

Proof.

1. Suppose that $a_1, a_2 \in A$ satisfy $(g \circ f)(a_1) = (g \circ f)(a_2)$. We then have that $g(f(a_1)) = g(f(a_2))$. Using the fact that g is injective, we conclude that $f(a_1) = f(a_2)$. Now using the fact that f is injective, it follows that $a_1 = a_2$. Therefore, $g \circ f$ is injective.
2. Let $c \in C$. Since $g\colon B \to C$ is surjective, we may fix $b \in B$ with $g(b) = c$. Since $f\colon A \to B$ is surjective, we may fix $a \in A$ with $f(a) = b$. We then have

$$(g \circ f)(a) = g(f(a)) = g(b) = c$$

   Therefore, $g \circ f$ is surjective.
3. This follows from combining 1 and 2.

□

Proposition 5.1.8. *Let $f\colon A \to B$ be a function.*

1. *f is injective if and only if there exists a function $g\colon B \to A$ with $g \circ f = id_A$.*
2. *f is surjective if and only if there exists a function $h\colon B \to A$ with $f \circ h = id_B$.*
3. *f is bijective if and only if there exists a function $g\colon B \to A$ with both $g \circ f = id_A$ and $f \circ g = id_B$.*

Proof.

1. Suppose first that such a function exists, and fix $g\colon B \to A$ with $g \circ f = id_A$. Suppose that $a_1, a_2 \in A$ satisfy $f(a_1) = f(a_2)$. Applying the function g to both sides we see that $g(f(a_1)) = g(f(a_2))$, and hence $(g \circ f)(a_1) = (g \circ f)(a_2)$. We now have

$$\begin{aligned} a_1 &= id_A(a_1) \\ &= (g \circ f)(a_1) \\ &= (g \circ f)(a_2) \\ &= id_A(a_2) \\ &= a_2 \end{aligned}$$

   It follows that f is injective.

   Suppose conversely that f is injective. If $A = \emptyset$, then f is the empty function, and we are done by letting g be the empty function (if empty functions annoy you, just ignore this case). Let's assume then that $A \neq \emptyset$ and fix $a_0 \in A$. We define $g\colon B \to A$ as follows. Given $b \in B$, we define $g(b)$ as follows:

   - If $b \in \text{range}(f)$, then there exists a unique $a \in A$ with $f(a) = b$ (because f is injective), and we let $g(b) = a$ for this unique choice.
   - If $b \notin \text{range}(f)$, we let $g(b) = a_0$.

   This completes the definition of $g\colon B \to A$. We need to check that $g \circ f = id_A$. Given any $a \in A$, we have $g(f(a)) = a$ because in the definition of g on the value $b = f(a)$ we simply notice that $b = f(a) \in \text{range}(f)$ trivially, so we defined $g(b) = a$. Therefore, for every $a \in A$, we have that $(g \circ f)(a) = g(f(a)) = a = id_A(a)$. It follows that $g \circ f = id_A$.

2. Suppose first that such a function exists, and fix $h\colon B \to A$ with $f \circ h = id_A$. Suppose that $b \in B$. We then have that

$$b = id_B(b) = (f \circ g)(b) = f(g(b))$$

   hence there exists $a \in A$ with $f(a) = b$, namely $a = g(b)$. Since $b \in B$ was arbitrary, it follows that f is surjective.

   Suppose conversely that f is surjective. We define $h\colon B \to A$ as follows. For every $b \in B$, we know that there exists (possibly many) $a \in A$ with $f(a) = b$ because f is surjective. Given $b \in B$, we then define $h(b) = a$ for some (any) $a \in A$ for which $f(a) = b$. Now given any $b \in B$, notice that $h(b)$ satisfies $f(h(b)) = b$ by definition of h, so $(f \circ h)(b) = b = id_B(b)$. Since $b \in B$ was arbitrary, it follows that $f \circ h = id_B$.

3. The right to left direction is immediate from parts 1 and 2. For the left to right direction, we need only note that if f is a bijection, then the function g defined in the left to right direction in the proof of 1 equals the function h defined in the left to right direction in the proof of 2.

□

5.2 The Symmetric Group

We now have everything we need to construct a group. Function composition is an associative operation, and if we restrict to just bijections we've seen that we will cut down to only those functions with inverses. The only thing we need to be careful about is that even if $f\colon A \to B$ and $g\colon B \to C$ are functions so that $g \circ f$ makes sense, we would still have that $f \circ g$ doesn't make sense unless $A = C$. Even more importantly, $f \circ f$ doesn't make sense unless $A = B$. Thus, if we want to make sure that composition always works, we should stick only to those functions which are a bijection from a set X to itself.

Definition 5.2.1. *Let X be a set. A* permutation *of X is a bijection $f\colon X \to X$.*

Rather than use the standard symbols $f, g, h, \ldots$ for general functions, we typically employ letters late in the greek alphabet, like $\sigma, \tau, \pi, \mu, \ldots$ for permutations. For an example, suppose that $X = \{1, 2, 3, 4, 5, 6\}$. To define a permutation $\sigma\colon X \to X$, we need to say what σ does on each element of X, which can do this by a simple list. Define $\sigma\colon X \to X$ as follows.

- $\sigma(1) = 5$
- $\sigma(2) = 6$
- $\sigma(3) = 3$
- $\sigma(4) = 1$
- $\sigma(5) = 4$
- $\sigma(6) = 2$

We then have that $\sigma\colon X \to X$ is a permutation (note every element of X is hit by exactly one element). Now it's unnecessarily cumbersome to write out the value of $\sigma(k)$ on each k in a bulleted list like that. Instead, we can use the slightly more compact notation

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

Interpret the above matrix by taking the top row as the inputs and each number below as the corresponding output. We now have all we need to define one of our most important groups.

Definition 5.2.2. *Let X be a set. The set of all permutations of X together with the operation of function composition is a group called the* symmetric group *on X. We denote this group by S_X. If we're given $n \in \mathbb{N}^+$, we simply write S_n rather than the more cumbersome $S_{\{1,2,\ldots,n\}}$.*

Let's work with one of these groups concretely. Consider the following two elements of S_6.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 3 & 1 & 4 & 2 \end{pmatrix} \qquad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 6 & 2 & 4 \end{pmatrix}$$

Let's compute $\sigma \circ \tau$. Remember that function composition happens from right to left. That is, the composition $\sigma \circ \tau$ is obtained by performing τ first and following after by performing σ. For example, we have

$$(\sigma \circ \tau)(2) = \sigma(\tau(2)) = \sigma(1) = 5$$

Working through the 6 inputs, we obtain:

$$\sigma \circ \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 4 & 2 & 6 & 1 \end{pmatrix}$$

On the other hand, we have

$$\tau \circ \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 3 & 6 & 1 \end{pmatrix}$$

Notice that $\sigma \circ \tau \neq \tau \circ \sigma$. We have just shown that S_6 is a nonabelian group. Let's next determine just how large the group is.

Proposition 5.2.3. *If $|X| = n$, then $|S_X| = n!$. In particular, $|S_n| = n!$ for all $n \in \mathbb{N}^+$.*

Proof. We count the number of elements of the set S_X by constructing all possible permutations $\sigma\colon X \to X$ via a sequence of choices. To construct such a permutation, we can first determine the value $\sigma(1)$. Since no numbers have been claimed, we have n possibilities for this value because we can choose any element of X. Now we move on to determine the value $\sigma(2)$. Notice that we can make $\sigma(2)$ any value in X *except* for the chosen value of $\sigma(1)$ (because we need to ensure that σ is an injection). Thus, we have $n-1$ many possibilities. Now to determine the value of $\sigma(3)$, we can choose any value in X other than $\sigma(1)$ and $\sigma(2)$ because those have already been claimed, so we have $n-2$ many choices. In general, when trying to define $\sigma(k+1)$, we have already claimed k necessarily distinct elements of X, so we have exactly $n-k$ possibly choices for the value. Thus, the number of permutations of X equals

$$n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n!$$

□

Let's return to our element σ in S_6:

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

This method of representing σ is reasonably compact, but it hides the fundamental structure of what is happening. For example, it is very difficult to "see" what the order of σ would be without repeatedly composing σ with itself. We now develop another method for representing an permutation on X called *cycle notation*. The basic idea is to take an element of X and follow its path through σ. For example, let's start with 1. We have $\sigma(1) = 5$. Now instead of moving on to deal with 2, let's continue this strand and determine the value $\sigma(5)$. Looking above, we see that $\sigma(5) = 4$. If we continue on this path to investigate 4, we see that $\sigma(4) = 1$, and we have found a "cycle" $1 \to 5 \to 4 \to 1$ hidden inside σ. We will denote this cycle with the notation (1 5 4). Now that those numbers are taken care of, we start again with the smallest number not yet claimed, which in this case is 2. We have $\sigma(2) = 6$ and following up gives $\sigma(6) = 2$. Thus, we have found the cycle $2 \to 6 \to 2$ and we denote this by (2 6). We have now claimed all numbers other than 3, and when we investigate 3 we see that $\sigma(3) = 3$, so we form the sad lonely cycle (3). Putting this all together, we write σ in cycle notation as

$$\sigma = (1\ 5\ 4)(2\ 6)(3)$$

One might wonder whether we ever get "stuck" when trying to build these cycles. What would happen if we follow 1 and we repeat a number before coming back to 1? For example, what if we see $1 \to 3 \to 6 \to 2 \to 6$? Don't fear because this can never happen. The only way the example above could crop up is if the purported permutation sent both 3 and 2 to 6, which would violate the fact that the purported permutation is injective. Also, if we finish a few cycles and start up a new one, then it is not possible that our new cycle has any elements in common with previous ones. For example, if we already have the cycle $1 \to 3 \to 2 \to 1$ and we start with 4, we can' find $4 \to 5 \to 3$ because then both 1 and 5 would map to 3.

Our conclusion is that this process of writing down a permutation in cycle notation never gets stuck and results in writing the given permutation as a product of disjoint cycles. Working through the same process with the permutation

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

we see that in cycle notation we have

$$\tau = (1\ 3\ 5\ 2)(4\ 6)$$

Now you can determine $\sigma \circ \tau$ in cycle notation directly from the cycle notations of σ and τ. For example, suppose we want to calculate the following:

$$(1\ 2\ 4)(3\ 6)(5) \circ (1\ 6\ 2)(3\ 5\ 4)$$

We want to determine the cycle notation of the resulting function, so we first need to determine where it sends 1. Again, function composition happens from right to left. Looking at the function represented on the right, we see the cycle containing 1 is (1 6 2), so the right function sends 1 to 6. We then go to the function on the left and see where it sends 6 The cycle containing 6 there is (3 6), so it takes 6 and sends it to 3. Thus, the composition sends 1 to 3. Thus, our result starts out as

$$(1\ 3$$

Now we need to see what happens to 3. The function on the right sends 3 to 5, and the function on the left takes 5 and leave it alone, so we have

$$(1\ 3\ 5$$

When we move on to see what happens to 5, we notice that the right function sends it to 4 and then the left function takes 4 to 1. Since 1 is the first element the cycle we started, we now close the loop and have

$$(1\ 3\ 5)$$

We now pick up the least element not in the cycle and continue. Working it out, we end with:

$$(1\ 2\ 4)(3\ 6)(5) \circ (1\ 6\ 2)(3\ 5\ 4) = (1\ 3\ 5)(2)(4\ 6)$$

Finally, we make our notation a bit more compact with a few conventions. First, we simply omit any cycles of length 1, so we just write (1 2 4)(3 6) instead of (1 2 4)(3 6)(5). Of course, this requires an understanding of which n we are using to avoid ambiguity as the notation (1 2 4)(3 6) doesn't specify whether we are viewing is as an element of S_6 or S_8 (in the latter case, the corresponding function fixes both 7 and 8). Also, as with most group operations, we simply omit the $\circ$ when composing functions. Thus, we would write the above as:

$$(1\ 2\ 4)(3\ 6)(1\ 6\ 2)(3\ 5\ 4) = (1\ 3\ 5)(2)(4\ 6)$$

Now there is potential for some conflict here. Looking at the first two cycles above, we meant to think of (1 2 4)(3 6) as one particular function on $\{1, 2, 3, 4, 5, 6\}$, but by omitting the group operation it could also be interpreted as (1 2 4) $\circ$ (3 6). Fortunately, these are exactly the same function because the cycles are disjoint. Thus, there is no ambiguity.

Let's work out everything about S_3. First, we know that $|S_3| = 3! = 6$. Working through the possibilities, we determine that

$$S_3 = \{id, (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)\}$$

so S_3 has the identity function, three 2-cycles, and two 3-cycles. Here is the Cayley table of S_3.

$\circ$	id	(1 2)	(1 3)	(2 3)	(1 2 3)	(1 3 2)
id	id	(1 2)	(1 3)	(2 3)	(1 2 3)	(1 3 2)
(1 2)	(1 2)	id	(1 3 2)	(1 2 3)	(2 3)	(1 3)
(1 3)	(1 3)	(1 2 3)	id	(1 3 2)	(1 2)	(2 3)
(2 3)	(2 3)	(1 3 2)	(1 2 3)	id	(1 3)	(1 2)
(1 2 3)	(1 2 3)	(1 3)	(2 3)	(1 2)	(1 3 2)	id
(1 3 2)	(1 3 2)	(2 3)	(1 2)	(1 3)	id	(1 2 3)

Notice that S_3 is a nonabelian group of order 6. In fact, it is the smallest possible nonabelian group, as we shall see later.

As alluded to above, viewing an element of $\sigma \in S_X$ in cycle notation allows a cleaer understanding of the group theoretic properties of σ. For example, let's think about how to compute $|\sigma|$. When computing powers of a permutation σ, the following simple fact is very useful.

Proposition 5.2.4. *Disjoint cycles commutes. That is, if X is a set and $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_\ell \in X$ are all distinct, then*

$$(a_1\ a_2\ \cdots\ a_k)(b_1\ b_2\ \cdots\ b_\ell) = (b_1\ b_2\ \cdots\ b_\ell)(a_1\ a_2\ \cdots\ a_k)$$

Proof. Simply work through where each a_i and b_j are sent on each side. Since $a_i \neq b_j$ for all i, j, each a_i is fixed by the cycle containing the b_j's and vice versa. □

Next, we need to determine the order of a cycle.

Proposition 5.2.5. *Let X be a set and let $a_1, a_2, \ldots, a_k \in X$ be distinct. Let $\sigma = (a_1\ a_2\ \cdots\ a_k)$. We then have that $|\sigma| = k$. In other words, the order of a cycle is its length.*

Proof. If $1 \leq i < k$, we have $\sigma^i(a_1) = a_i \neq a_1$, so $\sigma^i \neq id$. For each i, we have $\sigma^k(a_i) = a_i$, so σ^k fixes each a_i. Since σ fixes all other elements of X, it follows that σ^k fixes all other elements of X. Therefore, $\sigma^k = id$ and we have shown that $|\sigma| = k$. □

Proposition 5.2.6. *Let X be a set and let $a_1, a_2, \ldots, a_k \in X$ be distinct. Let $\sigma = (a_1\ a_2\ \cdots\ a_k)$. We then have that*

$$\begin{aligned}\sigma^{-1} &= (a_k\ a_{k-1}\ \cdots\ a_2\ a_1)\\ &= (a_1\ a_k\ a_{k-1}\ \cdots\ a_2)\end{aligned}$$

Proof. Let $\tau = (a_1\ a_k\ a_{k-1}\ \cdots\ a_2)$. For any i with $1 \leq i \leq k-1$, we have

$$(\tau \circ \sigma)(a_i) = \tau(\sigma(a_i)) = \tau(a_{i+1}) = a_i$$

and also

$$(\tau \circ \sigma)(a_k) = \tau(\sigma(a_k)) = \tau(a_1) = a_k$$

We have

$$(\sigma \circ \tau)(a_1) = \sigma(\tau(a_1)) = \sigma(a_k) = a_1$$

and for any $i \geq 2$

$$(\sigma \circ \tau)(a_i) = \sigma(\tau(a_i)) = \sigma(a_{i-1}) = a_i$$

□

Theorem 5.2.7. *Let X be a set and let $\sigma \in S_X$. We then have that $|\sigma|$ is the least common multiple of the cycle lengths occurring in the cycle notation of σ.*

Proof. Suppose that $\sigma = \tau_1\tau_2\cdots\tau_\ell$ where the τ_i are disjoint cycles. For each i, let $m_i = |\tau_i|$, and notice from above that m_i is the length of the cycle τ_i. Since disjoint cycles commute, for any $n \in \mathbb{N}^+$ we have

$$\sigma^n = \tau_1^n\tau_2^n\cdots\tau_\ell^n$$

Now if $m_i \mid n$ for each i, then $\tau_i^n = id$ for each i, so $\sigma^n = id$. Conversely, suppose that $n \in \mathbb{N}^+$ is such that there exists i with $m_i \nmid n$. We then have that $\tau_i^n \neq id$, so we may fix $a \in X$ with $\tau_i^n(a) \neq a$. Now both a and $\tau_i^n(a)$ are fixed by each τ_j with $j \neq i$ (because the cycles are disjoint). Therefore $\sigma^n(a) = \tau_i^n(a) \neq a$, and hence $\sigma^n \neq id$.

It follows that $\sigma^n = id$ if and only if $m_i \mid n$ for all i. Since $|\sigma|$ is the least $n \in \mathbb{N}^+$ with $\sigma^n = id$, it follows that $|\sigma|$ is the least $n \in \mathbb{N}^+$ satisfying $m_i \mid n$ for all i, which is to say that $|\sigma|$ is the least common multiple of the m_i. □

5.3 Transpositions and the Alternating Group

Definition 5.3.1. *Let X be a set. A* transposition *is an element of S_X whose cycle notation has equals $(a\ b)$ with $a \neq b$. in other words, a transposition is a bijection which flips two elements of X and leaves all other elements of X fixed.*

Proposition 5.3.2. *Every element of S_n can be written as a product of transpositions.*

Proof. We've seen that every permutation is a product of disjoint cycles, so it suffices to show that every cycle can be written as a product of transpositions. If $a_1, a_2, \ldots, a_k \in \{1, 2, \ldots, n\}$ are distinct, we have

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

The result follows. □

To illustrate the above construction in a special case, we have

$$(1\ 7\ 2\ 5)(3\ 9)(4\ 8\ 6) = (1\ 5)(1\ 2)(1\ 7)(3\ 9)(4\ 6)(4\ 8)$$

Notice that although we have proven that every element of S_n can be written as a product of transpositions and have given a description of how to do it, this decomposition is far from unique. It is even possible the chance the number of transpositions. For example, following our description, we have

$$(1\ 2\ 3) = (1\ 3)(1\ 2)$$

However, we can also write

$$(1\ 2\ 3) = (1\ 3)(2\ 3)(1\ 2)(1\ 3)$$

Although we can change the number and order of transpositions, it is a somewhat surprising fact that you can not change the parity of the number of transpositions. That is, it is impossible to find a $\sigma \in S_n$ which you can write simultaneously as a product of an even number of transpositions and also as a product of an odd number of transpositions. There are many proofs of this fundamental fact. We give one using the notions from linear algebra.

Definition 5.3.3. *Given $\sigma \in S_n$, we define an $n \times n$ matrix $M(\sigma)$ by letting*

$$M(\sigma)_{i,j} = \begin{cases} 1 & \text{if } \sigma(j) = i \\ 0 & \text{otherwise} \end{cases}$$

As an example, suppose that we are working in S_4. Let

$$\sigma = (1\ 2\ 3) \qquad \tau = (1\ 2)(3\ 4)$$

We then have

$$M(\sigma) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad M(\tau) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Notice that

$$\sigma\tau = (1\ 3\ 4)$$

so

$$M(\sigma\tau) = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

If you check the matrix above, you will see that $M(\sigma\tau) = M(\sigma) \cdot M(\tau)$. This result is true in general:

Proposition 5.3.4. *Given $\sigma, \tau \in S_n$, we have $M(\sigma\tau) = M(\sigma) \cdot M(\tau)$.*

Proof. We examine the entry $(M(\sigma)\cdot M(\tau))_{i,j}$. Now the i^{th} row of $M(\sigma)$ will have all zeros except for exactly one 1, say in position (i,k). By definition of $M(\sigma)$, we then have $\sigma(k) = i$. Similarly, the j^{th} column of $M(\tau)$ will have all zeros except for exactly one 1, say in position (ℓ, j). By definition of $M(\tau)$, we then have $\tau(j) = \ell$. From the definition of matrix multiplication, the entry $(M(\sigma) \cdot M(\tau))_{i,j}$ is the dot product of row i of $M(\sigma)$ with column j of $M(\tau)$. Therefore

$$(M(\sigma) \cdot M(\tau))_{i,j} = \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{otherwise} \end{cases}$$

Now we claim that $k = \ell$ if and only if $(\sigma\tau)(j) = i$. First notice that if $k = \ell$, then

$$(\sigma\tau)(j) = \sigma(\tau(j)) = \sigma(\ell) = \sigma(k) = i$$

Conversely, suppose that $(\sigma\tau)(j) = i$. We then have $\sigma(\tau(j)) = i$, so $\sigma(\ell) = i$. Since $\sigma(k) = i$ and σ is injective, it follows that $k = \ell$. Therefore

$$(M(\sigma) \cdot M(\tau))_{i,j} = \begin{cases} 1 & \text{if } (\sigma\tau)(j) = i \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $(M(\sigma) \cdot M(\tau))_{i,j} = M(\sigma\tau)_{i,j}$. Since i and j were arbitrary, it follows that $M(\sigma\tau) = M(\sigma) \cdot M(\tau)$. □

Definition 5.3.5. *Let $n \in \mathbb{N}^+$. We define a function $\varepsilon\colon S_n \to \{\pm 1\}$ be letting $\varepsilon(\sigma) = \det(M(\sigma))$. The value $\varepsilon(\sigma)$ is called the* sign *of σ. If $\varepsilon(\sigma) = 1$, then we say that σ is an* even *permutation. If $\varepsilon(\sigma) = -1$, then we say that σ is an* odd *permutation.*

Proposition 5.3.6. *If μ is a transposition, then $\varepsilon(\mu) = -1$.*

Proof. Since μ is a transposition, we can get from $M(\mu)$ to I_n via one row interchange. Since a row interchange introduces a minus sign in the determinant, and $\deg(I_n) = 1$, it follows that

$$\varepsilon(\mu) = \det(M(\mu)) = -\det(I_n) = -1$$

□

Proposition 5.3.7. *For all $\sigma, \tau \in S_n$, we have $\varepsilon(\sigma\tau) = \varepsilon(\sigma) \cdot \varepsilon(\tau)$*

Proof. We have

$$\begin{aligned} \varepsilon(\sigma\tau) &= \det(M(\sigma\tau)) \\ &= \det(M(\sigma) \cdot M(\tau)) \\ &= \det(M(\sigma)) \cdot \det(M(\tau)) \\ &= \varepsilon(\sigma) \cdot \varepsilon(\tau) \end{aligned}$$

□

Proposition 5.3.8. *Let $\sigma \in S_n$.*

- *$\varepsilon(\sigma) = 1$ if and only if it is possible to write σ as a product of an even number of transpositions.*
- *$\varepsilon(\sigma) = -1$ if and only if it is possible to write σ as a product of an odd number of transpositions.*

In particular, given an arbitrary permutation $\sigma \in S_n$, it is not possible to write σ both as an even number of permutations and also as an odd number of permutations.

Proof. Suppose that $\sigma = \mu_1\mu_2\cdots\mu_m$ where the μ_i are transpositions. We then have that

$$\varepsilon(\sigma) = \varepsilon(\mu_1)\varepsilon(\mu_2)\cdots\varepsilon(\mu_m) = (-1)^m$$

Therefore,

$$\varepsilon(\sigma) = \begin{cases} 1 & \text{if } m \text{ is even} \\ -1 & \text{if } m \text{ is odd} \end{cases}$$

□

Proposition 5.3.9. *Suppose that σ is a k-cycle.*

- *If k is an even number, then σ is an odd permutation, i.e. $\varepsilon(\sigma) = -1$.*
- *If k is an odd number, then σ is an even permutation, i.e. $\varepsilon(\sigma) = 1$.*

Proof. Write $\sigma = (a_1\ a_2\ a_3\ \cdots\ a_{k-1}\ a_k)$ where the a_i are distinct. As above, we have

$$\sigma = (a_1\ a_k)(a_1\ a_{k-1})\cdots(a_1\ a_3)(a_1\ a_2)$$

Thus, σ is the product of $k-1$ many cycles. If k is an even number, then $k-1$ is an odd number, and hence σ is an odd permutation. If k is an odd number, then $k-1$ is an even number, and hence σ is an even permutation. □

For example, suppose that

$$\sigma = (1\ 9\ 2\ 6\ 4\ 11)(3\ 8\ 5)(7\ 10)$$

We then have

$$\begin{aligned} \varepsilon(\sigma) &= \varepsilon((1\ 9\ 2\ 6\ 4\ 11)) \cdot \varepsilon((3\ 8\ 5)) \cdot \varepsilon((7\ 10)) \\ &= (-1)\cdot 1 \cdot (-1) \\ &= 1 \end{aligned}$$

so σ is an even permutation.

Proposition 5.3.10. *Let $n \in \mathbb{N}^+$ and define*

$$A_n = \{\sigma \in S_n : \varepsilon(\sigma) = 1\}$$

We then have that A_n is a subgroup of S_n.

Proof. We have $\varepsilon(id) = \det(M(id)) = \det(I_n) = 1$, so $id \in A_n$. Suppose that $\sigma, \tau \in A_n$ so that $\varepsilon(\sigma) = 1 = \varepsilon(\tau)$. we then have

$$\varepsilon(\sigma\tau) = \varepsilon(\sigma)\cdot\varepsilon(\tau) = 1\cdot 1 = 1$$

so $\sigma\tau \in A_n$. Finally, suppose that $\sigma \in A_n$. We then have $\sigma\sigma^{-1} = id$ so

$$\begin{aligned} \varepsilon(\sigma^{-1}) &= 1\cdot\varepsilon(\sigma^{-1}) \\ &= \varepsilon(\sigma)\cdot\varepsilon(\sigma^{-1}) \\ &= \varepsilon(\sigma\sigma^{-1}) \\ &= \varepsilon(id) \\ &= 1 \end{aligned}$$

so $\sigma^{-1} \in A_n$. □

Definition 5.3.11. *The subgroup $A_n = \{\sigma \in S_n : \varepsilon(\sigma) = 1\}$ of S_n is called the* alternating group *of degree n.*

We next determine the order of A_n.

Proposition 5.3.12. *Given $n \geq 2$, we have $|A_n| = \frac{n!}{2}$.*

Proof. Define a function $f\colon A_n \to S_n$ by letting $f(\sigma) = \sigma(1\ 2)$. We first claim that f is injective. To see this, suppose that $\sigma, \tau \in A_n$ and that $f(\sigma) = f(\tau)$. We then have $\sigma(1\ 2) = \tau(1\ 2)$. Multiplying on the right by $(1\ 2)$, we conclude that $\sigma = \tau$. Therefore, f is injective.

We next claim that $\text{range}(f) = S_n \backslash A_n$. Suppose first that $\sigma \in A_n$. We then have $\varepsilon(\sigma) = 1$, so

$$\varepsilon(f(\sigma)) = \varepsilon(\sigma(1\ 2)) = \varepsilon(\sigma) \cdot \varepsilon((1\ 2)) = 1 \cdot (-1) = -1$$

so $f(\sigma) \in S_n \backslash A_n$. Conversely, suppose that $\tau \in S_n \backslash A_n$ so that $\varepsilon(\tau) = -1$. We then have

$$\varepsilon(\tau(1\ 2)) = \varepsilon(\tau) \cdot \varepsilon((1\ 2)) = (-1) \cdot (-1) = 1$$

so $\tau(1\ 2) \in A_n$. Now

$$f(\tau(1\ 2)) = \tau(1\ 2)(1\ 2) = \tau$$

so $\tau \in \text{range}(f)$. It follows that $\text{range}(f) = S_n \backslash A_n$. Therefore, f maps A_n bijectively onto $S_n \backslash A_n$ and hence $|A_n| = |S_n \backslash A_n|$. Since S_n is the disjoint union of these two sets and $|S_n| = n!$, it follows that $|A_n| = \frac{n!}{2}$. $\square$

5.4 Dihedral Groups

We now define another subgroup of S_n which has a very geometric flavor. By way of motivation, consider a regular n-gon, i.e. a convex polygon with n edges in which all edges have the same length and all interior angles are congruent. Now we want to consider the symmetries of this polygon obtained by rigid motions. That is, we want to think about all results obtained by picking up the polygon, moving it around in space, and then placing it back down so that it looks exactly the same. For example, one possibility is that we rotate the polygon around the center by $360/n$ degrees. We can also flip the polygon over across an imaginary line through the center so long as we take vertices to vertices.

To put these geometric ideas in more algebraic terms, first label the vertices clockwise with the number $1, 2, 3, \ldots, n$. Now a symmetry of the polygon will move vertices to vertices, so will correspond to a bijection of $\{1, 2, 3, \ldots, n\}$, i.e. an element of S_n. For example, the permutation $(1\ \ 2\ \ 3\ \cdots\ n)$ corresponds to rotation by $360/n$ degrees because it sends vertex 1 to the position originally occupied by vertex 2, sends vertex 2 to the position originally occupied by vertex 3, etc.

Definition 5.4.1. *Fix $n \geq 3$. We define the following elements of S_n.*

- $r = (1\ \ 2\ \ 3\ \cdots\ n)$
- $s = (2\ \ n)(3\ \ n-1)(4\ \ n-2)\cdots$. *More formally, we have the following*
  - *If $n = 2k$, let $s = (2\ \ n)(3\ \ n-1)(4\ \ n-2)\cdots(k\ \ k+2)$*
  - *If $n = 2k+1$, let $s = (2\ \ n)(3\ \ n-1)(4\ \ n-2)\cdots(k+1\ \ k+2)$*

Thus, r corresponds to clockwise rotation by $360/n$ degrees (as describe above), and s corresponds to flipping the polygon across the line through the vertex 1 and the center. Now given two symmetries of the polygon, we can do one followed by the other to get another symmetry. In other words, the composition of two symmetries is a symmetry. We can also reverse any rigid motion, so the inverse of a symmetry is a symmetry. Finally, the identity function is a trivial symmetry, so the set of all symmetries of the regular n-gon forms a subgroup of S_n.

Definition 5.4.2. *Let $n \geq 3$. We define D_n to be the subgroup generated by r and s, i.e. D_n is the smallest subgroup of S_n containing both r and s. The group D_n is called the* dihedral group *of degree n.*

In other words, we simply *define* D_n to be everything symmetry you can obtain through a simple rotation and a simple flip. This is a nice precise algebraic description. We will explain below why every possible rigid motion of a regular n-gon is given by an element of D_n.

As an example, the following is an element of D_n.

$$r^7 s^5 r^{-4} s^{-4} r s^{-3} r^{14}$$

We would like to find a way to simplify such expressions, and the next proposition is the primary tool to do so.

Proposition 5.4.3. *Let $n \geq 3$. We have the following*

1. $|r| = n$
2. $|s| = 2$
3. $sr = r^{-1}s = r^{n-1}s$
4. *For all $k \in \mathbb{N}^+$, we have $sr^k = r^{-k}s$*

Proof. We have $|r| = n$ because r is an n-cycle and $|s| = 2$ because s is a product of 2-cycles. We now check that $sr = r^{-1}s$. First notice that

$$r^{-1} = (n \;\; n-1 \;\; \cdots \;\; 3 \;\; 2 \;\; 1) = (1 \;\; n \;\; n-1 \;\; \cdots \;\; 3 \;\; 2)$$

Suppose first that $n = 2k$. We then have

$$\begin{aligned} sr &= (2 \;\; n)(3 \;\; n-1)(4 \;\; n-2)\cdots(k \;\; k+2)(1 \;\; 2 \;\; 3 \;\; \cdots \;\; n) \\ &= (1 \;\; n)(2 \;\; n-1)(3 \;\; n-2)\cdots(k \;\; k+1) \end{aligned}$$

and

$$\begin{aligned} r^{-1}s &= (n \;\; n-1 \;\; \cdots \;\; 3 \;\; 2 \;\; 1)(2 \;\; n)(3 \;\; n-1)(4 \;\; n-2)\cdots(k \;\; k+2) \\ &= (1 \;\; n)(2 \;\; n-1)(3 \;\; n-2)\cdots(k \;\; k+1) \end{aligned}$$

so $sr = r^{-1}s$. Suppose now that $n = 2k+1$. We then have

$$\begin{aligned} sr &= (2 \;\; n)(3 \;\; n-1)(4 \;\; n-2)\cdots(k+1 \;\; k+2)(1 \;\; 2 \;\; 3 \;\; \cdots \;\; n) \\ &= (1 \;\; n)(2 \;\; n-1)(3 \;\; n-2)\cdots(k \;\; k+2) \end{aligned}$$

and

$$\begin{aligned} r^{-1}s &= (n \;\; n-1 \;\; \cdots \;\; 3 \;\; 2 \;\; 1)(2 \;\; n)(3 \;\; n-1)(4 \;\; n-2)\cdots(k+1 \;\; k+2) \\ &= (1 \;\; n)(2 \;\; n-1)(3 \;\; n-2)\cdots(k \;\; k+2) \end{aligned}$$

so $sr = r^{-1}s$. The last statement now follows by induction. □

We now give an example of how to use this proposition to simply the above expression in the case $n = 5$. We have

$$\begin{aligned} r^7s^5r^{-4}s^{-4}rs^{-3}r^{14} &= r^2srrsr^4 && \text{(using } |r| = 5 \text{ and } |s| = 2\text{)} \\ &= r^2sr^2sr^4 \\ &= r^2sr^2r^{-4}s \\ &= r^2sr^{-2}s \\ &= r^2sr^3s \\ &= r^2r^{-3}ss \\ &= r^{-1}s^2 \\ &= r^4 \end{aligned}$$

Theorem 5.4.4. *Let $n \geq 3$.*

1. *$D_n = \{r^is^k : 0 \leq i \leq n-1, 0 \leq k \leq 1\}$*
2. *If $r^is^k = r^js^\ell$ with $0 \leq i, j \leq n-1$ and $0 \leq k, \ell \leq 1$, then $i = j$ and $k = \ell$.*

In particular, we have $|D_n| = 2n$.

Proof. Using the fundamental relations that $|r| = n$, that $|s| = 2$, and that $sr^k = r^{-k}s$ for all $k \in \mathbb{N}^+$, the above argument shows that any product of r, s, and their inverses equals r^is^k for some i, k with $0 \leq i \leq n-1$ and $0 \leq k \leq 1$. To be more precise, you use the above relations to show that the set $\{r^is^k : 0 \leq i \leq n-1, 0 \leq k \leq 1\}$ is closed under multiplication and under inverses, but I will leave such a check for you if you would like to work through it. This gives part 1.

Suppose now that $r^is^k = r^js^\ell$ with $0 \leq i, j \leq n-1$ and $0 \leq k, \ell \leq 1$. Multiplying on the left by r^{-j} and on the right by s^{-k}, we see that

$$r^{i-j} = s^{\ell-k}$$

Suppose for the sake of obtaining a contradiction that $k \neq \ell$. Since $k, \ell \in \{0, 1\}$, we must have $\ell - k \in \{-1, 1\}$, so as $s^{-1} = s$ it follows that $r^{i-j} = s^{\ell-k} = s$. Now we have $s(1) = 1$, so we must have $r^{i-j}(1) = 1$ as well. This implies that $n \mid (i-j)$, so as $-n < i - j < n$, we conclude that $i - j = 0$. Thus $r^{i-j} = id$, and we conclude that $s = id$, a contradiction. Therefore, we must have $k = \ell$. It follows that $s^{\ell-k} = s^0 = id$, so $r^{i-j} = id$. Using the fact that $|r| = n$ now, we see that $n \mid (i-j)$, and as above this implies that $i - j = 0$ so $i = j$. □

Corollary 5.4.5. *Let $n \geq 3$. We then have that D_n is a nonabelian group with $|D_n| = 2n$.*

Proof. We claim that $rs \neq sr$. Suppose instead that $rs = sr$. Since we know that $sr = r^{-1}s$, it follows that $rs = r^{-1}s$. Canceling the s on the right, we see that $r = r^{-1}$. Multiplying on the left by r we see that $r^2 = id$, but this is a contradiction because $|r| = n$. Therefore, $rs \neq sr$ and hence D_n is nonabelian. Now by the previous theorem we know that

$$D_n = \{r^is^k : 0 \leq i \leq n-1, 0 \leq k \leq 1\}$$

and by the second part of the theorem that the $2n$ elements described in the set are distinct. □

We now come back around to giving a geometric justification for why the elements of D_n exactly correspond to the symmetries of the regular n-gon. We have described why every element of D_n does indeed give a symmetry (because both r and s do, and the set of symmetries must be closed under composition and inversion), so we need only understand why all possible symmetries of the regular n-gon arise from an

element of D_n. To determine a symmetry, we first need to send the vertex labeled by 1 to another vertex. We have n possible choices for where to send it, and suppose we send it to the original position of vertex k. Once we have sent vertex 1 to the position of vertex k, we now need to determine were the vertex 2 is sent. Now vertex 2 must go to one of the vertices adjacent to k, so we only have 2 choices for where to send it. Finally, once we've determined these two vertices (where vertex 1 and vertex 2 go), the rest of the n-gon is determined because we have completely determined where an entire edge goes. Thus, there are a total of $n \cdot 2 = 2n$ many possible symmetries. Since $|D_n| = 2n$, it follows that all symmetries are given by elements of D_n.

Finally notice that $D_3 = S_3$ simply because D_3 is a subgroup of S_3 and $|D_3| = 6 = |S_3$. In other words, any permutation of the vertices of an equilateral triangle is obtainable via a rigid motion of the triangle. However, if $n \geq 4$, then $|D_n|$ is much smaller than $|S_n|$ as most permutations of the vertices of a regular n-gon can not be obtained from a rigid motion.

We end with the Cayley table of D_4:

$\circ$	id	r	r^2	r^3	s	rs	r^2s	r^3s
id	id	r	r^2	r^3	s	rs	r^2s	r^3s
r	r	r^2	r^3	id	rs	r^2s	r^3s	s
r^2	r^2	r^3	id	r	r^2s	r^3s	s	rs
r^3	r^3	id	r	r^2	r^3s	s	rs	r^2s
s	s	r^3s	r^2s	rs	id	r^3	r^2	r
rs	rs	s	r^3s	r^2s	r	id	r^3	r^2
r^2s	r^2s	rs	s	r^3s	r^2	r	id	r^3
r^3s	r^3s	r^2s	rs	s	r^3	r^2	r	id

Chapter 6

Cosets and Lagrange's Theorem

6.1 Cosets

Suppose that G is a group and that H is a subgroup of G. The idea we want to explore is how to collapse the elements of H by considering them all to be "trivial" like the identity e. If we want this idea to work, we would then want to identify two elements $a, b \in G$ if we can get from one to the other via multiplication by a "trivial" element. In other words, we want to identify elements a and b if there exists $h \in H$ with $ah = b$.

For example, suppose that G is the group $\mathbb{R}^2$ under addition (so $(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$) and that $H = \{(0, b) : b \in \mathbb{R}\}$ is the y-axis. Notice that H is subgroup of G. We want to consider everything on the y-axis, that is every pair of the form $(0, b)$, as trivial. Now if we want the y-axis to be considered "trivial", then we would want to consider two points to be the "same" if we can get from one to the other by adding an element of the y-axis. Thus, we would want to identify (a_1, b_1) with (a_2, b_2) if and only if $a_1 = a_2$ (because then we can add the "trivial" point $(0, b_2 - b_1)$ to (a_1, b_1) to get (a_2, b_2)).

Let's move on to the group $G = \mathbb{Z}$. Let $n \in \mathbb{N}^+$ and consider $H = n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$. In this situation, we want to consider all multiples of n to be "equal" to 0, and in general we want to consider $a, b \in \mathbb{Z}$ to be equal if we can add some multiple of n to a to obtain b. In other words, we want to identify a and b if and only if there exists $k \in \mathbb{Z}$ with $a + kn = b$. Working it out, we want to identify a and b if and only if $b - a$ is a multiple of n, i.e. if and only if $a \equiv_n b$. Thus, in the special case of the subgroup $n\mathbb{Z}$ of $\mathbb{Z}$, we recover the fundamental ideas of modular arithmetic and our eventual definition of $\mathbb{Z}/n\mathbb{Z}$.

Left Cosets

Definition 6.1.1. *Let G be a group and let H be a subgroup of G. We define a relation $\sim_H$ on G by letting $a \sim_H b$ mean that there exists $h \in H$ with $ah = b$.*

Proposition 6.1.2. *Let G be a group and let H be a subgroup of G. The relation $\sim_H$ is an equivalence relation on G.*

Proof. We check the three properties.

- Reflexive: Let $a \in G$. We have that $ae = a$ and we know that $e \in H$ because H is a subgroup of G, so $a \sim_H a$.
- Symmetric: Let $a, b \in G$ with $a \sim_H b$. Fix $h \in H$ with $ah = b$. Multiplying on the right by h^{-1}, we see that $a = bh^{-1}$, so $bh^{-1} = a$. Now $h^{-1} \in H$ because H is a subgroup of G, so $b \sim_H a$

- Transitive: Let $a, b, c \in G$ with $a \sim_H b$ and $b \sim_H c$. Fix $h_1, h_2 \in H$ with $ah_1 = b$ and $bh_2 = c$. We then have
$$a(h_1h_2) = (ah_1)h_2 = bh_2 = c$$
Now $h_1h_2 \in H$ because H is a subgroup of G, so $a \sim_H c$.

Therefore, $\sim_H$ is an equivalence relation on G. □

The next proposition is a useful little rephrasing of when $a \sim_H b$.

Proposition 6.1.3. *Let G be a group and let H be a subgroup of G. Given $a, b \in G$, we have $a \sim_H b$ if and only if $a^{-1}b \in H$.*

Proof. Suppose first that $a, b \in G$ satisfy $a \sim_H b$. Fix $h \in H$ with $ah = b$. Multiplying on the left by a^{-1}, we conclude that $h = a^{-1}b$, so $a^{-1}b = h \in H$.

Suppose conversely that $a^{-1}b \in H$. Since
$$a(a^{-1}b) = (aa^{-1})b = eb = b$$
it follows that $a \sim_H b$. □

Definition 6.1.4. *Let G be a group and let H be a subgroup of G. Under the equivalence relation $\sim_H$, we have*
$$\begin{aligned}\overline{a} &= \{b \in G : a \sim_H b\} \\ &= \{b \in G : \text{ There exists } h \in H \text{ with } b = ah\} \\ &= \{ah : h \in H\} \\ &= aH.\end{aligned}$$
These equivalence class are called left cosets *of H in G.*

Since the left cosets of H in G are the equivalence classes of the equivalence relation $\sim_H$, all of our theory about equivalence relations apply. For example, if two left cosets of H in G intersect nontrivially, then they are in fact equal. Also, as is the case for general equivalence relations, the left cosets of H in G partition G into pieces. Using Proposition 6.1.3, we obtain the following fundamental way of determining when two left cosets are equal:
$$aH = bH \iff a^{-1}b \in H \iff b^{-1}a \in H$$
Let's work through an example.

Example 6.1.5. *Let $G = S_3$ and let $H = \langle (1\ 2) \rangle = \{id, (1\ 2)\}$. Determine the left cosets of H in G.*

Proof. The left cosets are the sets σH for $\sigma \in G$. For example, we have the left coset
$$idH = \{id \circ id, id \circ (1\ 2)\} = \{id, (1\ 2)\}$$
We can also consider the left coset
$$(1\ 2)H = \{id \circ (1\ 2), (1\ 2) \circ (1\ 2)\} = \{(1\ 2), id\} = \{id, (1\ 2)\}$$
Thus, the two left cosets idH and $(1\ 2)H$ are equal. This should not be surprising because
$$id^{-1} \circ (1\ 2) = id \circ (1\ 2) = (1\ 2) \in H$$
Alternatively, we can simply note that $(1\ 2) \in idH$ and $(1\ 2) \in (1\ 2)H$, so since the left cosets idH and $(1\ 2)H$ intersect nontrivially, we know immediately that they must be equal. Working through all the examples, we compute σH for each of the six $\sigma \in S_3$:

1. $idH = (1\ 2)H = \{id, (1\ 2)\}$
2. $(1\ 3)H = (1\ 2\ 3)H = \{(1\ 3), (1\ 2\ 3)\}$
3. $(2\ 3)H = (1\ 3\ 2)H = \{(2\ 3), (1\ 3\ 2)\}$

Thus, there are 3 distinct left cosets of H in G. □

Intuitively, if H is a subgroup of G, then a left coset aH is simply a "translation" of the subgroup H within G using the element $a \in G$. In the above example, $(1\ 3)H$ is simply the "shift" of the subgroup H using $(1\ 3)$ because we obtain it by hitting every element of H on the left by $(1\ 3)$.

To see this geometrically, consider again the group $G = \mathbb{R}^2$ under addition with subgroup

$$H = \{(0, b) : b \in \mathbb{R}\}$$

equal to the y-axis (or equivalently the line $x = 0$). Since the operation in G is $+$, we will denote the left coset of $(a, b) \in G$ by $(a, b) + H$ (rather than $(a, b)H$). Let's consider the left coset $(3, 0) + H$. We have

$$\begin{aligned}(3, 0) + H &= \{(3, 0) + (0, b) : b \in \mathbb{R}\} \\ &= \{(3, b) : b \in \mathbb{R}\}\end{aligned}$$

so $(3, 0) + H$ gives line $x = 3$. Thus, the left coset $(3, 0) + H$ is the translation of the line $x = 0$. Let's now consider the coset $(3, 5) + H$. We have

$$\begin{aligned}(3, 5) + H &= \{(3, 5) + (0, b) : b \in \mathbb{R}\} \\ &= \{(3, 5 + b) : b \in \mathbb{R}\} \\ &= \{(3, c) : c \in \mathbb{R}\} \\ &= (3, 0) + H.\end{aligned}$$

Notice that we could have obtained this with less work by noting that the inverse of $(3, 5)$ in G is $(-3, -5)$ and $(-3, -5) + (3, 0) = (0, -5) \in H$, hence $(3, 5) + H = (3, 0) + H$. Therefore, the left coset $(3, 5) + H$ also gives the line $x = 3$. Notice that every element of H when hit by $(3, 5)$ translates right 3 and shifts up 5, but as a set this latter shift of 5 is washed away.

Finally, let's consider $G = \mathbb{Z}$ (under addition) and $H = n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$ where $n \in \mathbb{N}^+$. Notice that given $a, b \in \mathbb{Z}$, we have

$$\begin{aligned}a \sim_{n\mathbb{Z}} b &\iff \text{There exists } h \in n\mathbb{Z} \text{ with } a + h = b \\ &\iff \text{There exists } k \in \mathbb{Z} \text{ with } a + nk = b \\ &\iff \text{There exists } k \in \mathbb{Z} \text{ with } nk = b - a \\ &\iff n \mid (b - a) \\ &\iff b \equiv_n a \\ &\iff a \equiv_n b.\end{aligned}$$

Therefore the two relations $\sim_{n\mathbb{Z}}$ and $\equiv_n$ are precisely the same, and we have recovered congruence modulo n as a special case of our general construction. Since the relations are the same, they have the same equivalence classes. Hence, the equivalence class of a under the equivalence relation $\equiv_n$, which in the past we denoted by $\overline{a}$, equals the equivalence class of a under the equivalence relation $\sim_{n\mathbb{Z}}$, which is the left coset $a + n\mathbb{Z}$.

Right Cosets

In the previous section, we defined $a \sim_H b$ to mean that there exists $h \in H$ with $ah = b$. Thus, we considered two elements of G to be equivalent if we could get from a to b through multiplication by an element of H on the right of a. In particular, with this definition, we saw that when we consider $G = S_3$ and $H = \langle(1\ 2)\rangle$, we have $(1\ 3) \sim_H (1\ 2\ 3)$ because $(1\ 3)(1\ 2) = (1\ 2\ 3)$.

What happens if we switch things up? For the rest of this section, completely ignore the definition of $\sim_H$ defined above because we will redefine it on the other side now.

Definition 6.1.6. *Let G be a group and let H be a subgroup of G. We define a relation $\sim_H$ on G by letting $a \sim_H b$ mean that there exists $h \in H$ with $ha = b$.*

The following results are proved exactly as above just working on the other side.

Proposition 6.1.7. *Let G be a group and let H be a subgroup of G. The relation $\sim_H$ is an equivalence relation on G.*

Proposition 6.1.8. *Let G be a group and let H be a subgroup of G. Given $a, b \in G$, we have $a \sim_H b$ if and only if $ab^{-1} \in H$.*

Now it would be nice if this new equivalence relation was the same as the original equivalence relation. Too bad. In general, they are different! For example, with this new equivalence relation, we do **not** have $(1\ 3) \sim_H (1\ 2\ 3)$ because $id \circ (1\ 3) = (1\ 3)$ and $(1\ 2)(1\ 3) = (1\ 3\ 2)$. Now that we know that the two relations differ in general, we should think about the equivalence classes of this new equivalence relation.

Definition 6.1.9. *Let G be a group and let H be a subgroup of G. Under the equivalence relation $\sim_H$, we have*

$$\begin{aligned}\overline{a} &= \{b \in G : a \sim_H b\} \\ &= \{b \in G : \text{ There exists } h \in H \text{ with } b = ha\} \\ &= \{ha : h \in H\} \\ &= Ha.\end{aligned}$$

These equivalence class are called right cosets *of H in G.*

Let's work out the right cosets of our previous example.

Example 6.1.10. *Let $G = S_3$ and let $H = \langle(1\ 2)\rangle = \{id, (1\ 2)\}$. Determine the right cosets of H in G.*

Proof. Working them all out, we obtain.

- $Hid = H(1\ 2) = \{id, (1\ 2)\}$
- $H(1\ 3) = H(1\ 3\ 2) = \{(1\ 3), (1\ 3\ 2)\}$
- $H(2\ 3) = H(1\ 2\ 3) = \{(2\ 3), (1\ 2\ 3)\}$

Thus, there are 3 distinct right cosets of H in G. □

Notice that although we obtained both 3 left cosets and 3 right cosets, these cosets were different. In particular, we have $(1\ 3)H \neq H(1\ 3)$. In other words, it is not true in general that the left coset aH equals the right coset Ha. Notice that this is fundamentally an issue because S_3 is nonabelian. If we were working in an abelian group G with a subgroup H of G, then $ah = ha$ for all $h \in H$, so $aH = Ha$.

As in the left coset section, using Proposition 6.1.8, we have the following fundamental way of determining when two left cosets are equal:

$$Ha = Hb \iff ab^{-1} \in H \iff ba^{-1} \in H$$

Index of a Subgroup

As we saw in the previous sections, it is not in general true that left cosets are right cosets and vice versa. However, in the one example we saw above, we at least had the same number of left cosets as we had right cosets. This is a general and important fact, which we now establish.

First, let's once again state the fundamental way to tell when two left cosets are equal and when two right cosets are equal.

$$aH = bH \iff a^{-1}b \in H \iff b^{-1}a \in H$$

$$Ha = Hb \iff ab^{-1} \in H \iff ba^{-1} \in H$$

Suppose now that G is a group and that H is a subgroup of G. Let $\mathcal{L}_H$ be the set of left cosets of H in G and let $\mathcal{R}_H$ be the set of right cosets of H in G. We will show that $|\mathcal{L}_H| = |\mathcal{R}_H|$ by defining a bijection $f\colon \mathcal{L}_H \to \mathcal{R}_H$. Now the natural idea is to define f by letting $f(aH) = Ha$. However, we need to be very careful. Recall that the left cosets of H in G are the equivalence classes of a certain equivalence relation. By "defining" f as above we are giving a definition based on particular representatives of these equivalence classes, and it may be possible that $aH = bH$ but $Ha \neq Hb$. In other words, we must determine whether f is well-defined.

In fact, in general the above f is not well-defined. Consider our standard example of $G = S_3$ and $H = \langle(1\ 2)\rangle$. Checking our above computations, we have $(1\ 3)H = (1\ 2\ 3)H$ but $H(1\ 3) \neq H(1\ 2\ 3)$. Therefore, in this particular case, that choice of f is not well-defined. We need to define f differently to make it well-defined in general, and the following lemma is the key to do so.

Lemma 6.1.11. *Let H be a subgroup of G and let $a, b \in G$. The following are equivalent.*

1. $aH = bH$
2. $Ha^{-1} = Hb^{-1}$

Proof. Suppose that $aH = bH$. We then have that $a^{-1}b \in H$ so using the fact that $(b^{-1})^{-1} = b$, we see that $a^{-1}(b^{-1})^{-1} \in H$. It follows that $Ha^{-1} = Hb^{-1}$.

Suppose conversely that $Ha^{-1} = Hb^{-1}$. We then have that $a^{-1}(b^{-1})^{-1} \in H$ so using the fact that $(b^{-1})^{-1} = b$, we see that $a^{-1}b \in H$. It follows that $aH = bH$. □

We can now prove our result.

Proposition 6.1.12. *Let G be a group and let H be a subgroup of G. Let $\mathcal{L}_H$ be the set of left cosets of H in G and let $\mathcal{R}_H$ be the set of right cosets of H in G. Define $f\colon \mathcal{L}_H \to \mathcal{R}_H$ by letting $f(aH) = Ha^{-1}$. We then have that f is a well-defined bijection from $\mathcal{L}_H$ onto $\mathcal{R}_H$. In particular, $|\mathcal{L}_H| = |\mathcal{R}_H|$.*

Proof. Notice that f is well-defined by the above lemma because if $aH = bH$, then

$$f(aH) = Ha^{-1} = Hb^{-1} = f(bH).$$

We next check that f is injective. Suppose that $f(aH) = f(bH)$, so that $Ha^{-1} = Hb^{-1}$. By the other direction of the lemma, we have that $aH = bH$. Therefore, f is injective.

Finally, we need to check that f is surjective. Fix an element of $\mathcal{R}_H$, say Hb. We then have that $b^{-1}H \in \mathcal{L}_H$ and

$$f(b^{-1}H) = H(b^{-1})^{-1} = Hb.$$

Hence, $\text{range}(f) = \mathcal{R}_H$, so f is surjective.

Putting it all together, we conclude that f is a well-defined bijection from $\mathcal{L}_H$ onto $\mathcal{R}_H$. □

Definition 6.1.13. *Let G be a group and let H be a subgroup of G. We define $[G : H]$ to be the number of left cosets of H in G (or equivalently the number of right cosets of H in G). That is, $[G : H]$ is the number of equivalence classes of G under the equivalence relation $\sim_H$. If there are infinitely many left cosets (or equivalently infinitely many right cosets), we write $[G : H] = \infty$. We call $[G : H]$ the* index *of H in G.*

For example, we saw above that $[S_3 : \langle(1\ 2)\rangle] = 3$. For any $n \in \mathbb{N}^+$, we have $[\mathbb{Z} : n\mathbb{Z}] = n$ because the left cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ are $0 + n\mathbb{Z}, 1 + n\mathbb{Z}, \ldots, (n-1) + n\mathbb{Z}$.

6.2 Lagrange's Theorem and Consequences

We are now in position to prove one of the most fundamental theorems about finite groups. We start with a proposition.

Proposition 6.2.1. *Let G be a group and let H be a subgroup of G. Let $a \in G$. Define a function $f : H \to aH$ by letting $f(h) = ah$. We then have that f is a bijection, so $|H| = |aH|$. In other words, all left cosets of H in G have the same size.*

Proof. Notice that f is surjective because if $b \in aH$, then we may fix $h \in H$ with $b = ah$, and notice that $f(h) = ah = b$, so $b \in \text{range}(f)$. Suppose that $h_1, h_2 \in H$ and that $f(h_1) = f(h_2)$. We then have that $ah_1 = ah_2$, so by canceling the a's on the left (i.e. multiplying on the left by a^{-1}), we conclude that $h_1 = h_2$. Therefore, f is injective. Putting this together with the fact that f is surjective, we conclude that f is a bijection. The result follows. □

Theorem 6.2.2 (Lagrange's Theorem)**.** *Let G be a finite group and let H be a subgroup of G. We have*

$$|G| = [G : H] \cdot |H|.$$

In particular, $|H|$ divides $|G|$ and

$$[G : H] = \frac{|G|}{|H|}.$$

Proof. The previous proposition shows that the $[G : H]$ many left cosets of H in G each have cardinality $|H|$. Since G is partitioned into $[G : H]$ many sets each of size $|H|$, we conclude that $|G| = [G : H] \cdot |H|$. □

For example, instead of finding all of the left cosets of $\langle(1\ 2)\rangle$ in S_3 to determine that $[S_3 : \langle(1\ 2)\rangle] = 3$, we could have simply calculated

$$[S_3 : \langle(1\ 2)\rangle] = \frac{|S_3|}{|\langle(1\ 2)\rangle|} = \frac{6}{2} = 3.$$

Notice the assumption in Lagrange's Theorem that G is finite. It makes no sense to calculate $[\mathbb{Z} : n\mathbb{Z}]$ in this manner (if you try to write $\frac{\infty}{\infty}$ you will make me very angry). We end this section with several simple consequences of Lagrange's Theorem.

Corollary 6.2.3. *Let G be a finite group. Let K be a subgroup of G and let H be a subgroup of K. We then have*

$$[G : H] = [G : K] \cdot [K : H].$$

Proof. Since G is finite (and hence trivially both H and K are finite) we may use Lagrange's Theorem to note that

$$[G : K] \cdot [K : H] = \frac{|G|}{|K|} \cdot \frac{|K|}{|H|} = \frac{|G|}{|H|} = [G : H].$$

□

Corollary 6.2.4. *Let G be a finite group and let $a \in G$. We then have that $|a|$ divides $|G|$.*

Proof. Let $H = \langle a \rangle$. By Proposition 4.3.9, we know that H is a subgroup of G and $|a| = |H|$. Therefore, by Lagrange's Theorem, we may conclude that $|a|$ divides $|G|$. □

Corollary 6.2.5. *Let G be a finite group. We then have that $a^{|G|} = e$ for all $a \in G$.*

Proof. Let $m = |a|$. By the previous corollary, we know that $m \mid |G|$. Therefore, by Proposition 4.2.5, it follows that $a^{|G|} = e$. □

Theorem 6.2.6. *Every group of prime order is cyclic (so in particular every group of prime order is abelian). In fact, if G is a group of prime order, then every nonidentity element is a generator of G.*

Proof. Let $p = |G|$ be the prime order of G. Suppose that $c \in G$ with $c \neq e$. We know that $|c|$ divides p, so as p is prime we must have that either $|c| = 1$ or $|c| = p$. Since $c \neq e$, we must have $|c| = p$. By Proposition 4.4.2, we conclude that c is a generator of G. Hence, every nonidentity element of G is a generator of G. □

Theorem 6.2.7 (Euler's Theorem)**.** *Let $n \in \mathbb{N}^+$ and let $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. We then have $a^{\varphi(n)} \equiv_n 1$.*

Proof. We apply Corollary 6.2.4 to the group $U(\mathbb{Z}/n\mathbb{Z})$. Since $\gcd(a, n) = 1$, we have that $\overline{a} \in U(\mathbb{Z}/n\mathbb{Z})$. Since $|U(\mathbb{Z}/n\mathbb{Z})| = \varphi(n)$, Corollary 6.2.4 tells us that $\overline{a}^{\varphi(n)} = \overline{1}$. Therefore, $a^{\varphi(n)} \equiv_n 1$. □

Corollary 6.2.8 (Fermat's Little Theorem)**.** *Let $p \in \mathbb{N}^+$ be prime.*

1. *If $a \in \mathbb{Z}$ with $p \nmid a$, then $a^{p-1} \equiv_p 1$.*
2. *If $a \in \mathbb{Z}$, then $a^p \equiv_p a$.*

Proof. We first prove 1. Suppose that $a \in \mathbb{Z}$ with $p \nmid a$. We then have that $\gcd(a, p) = 1$ (because $\gcd(a, p)$ divides p, so it must be either 1 or p, but it can not be p since $p \nmid a$). Now using the fact that $\varphi(p) = p - 1$, we conclude from Euler's Theorem that $a^{p-1} \equiv_p 1$.

We next prove 2. Suppose that $a \in \mathbb{Z}$. If $p \nmid a$, then $a^{p-1} \equiv_p 1$ by part 1, so multiplying both sides by a gives $a^p \equiv_p a$. Now if $p \mid a$, then we trivially have $p \mid a^p$ as well, so $p \mid (a^p - a)$ and hence $a^p \equiv_p a$. □

Chapter 7

New Groups From Old

In this section, we show how to construct new groups from given ones. We already have the notion of a subgroup which provides examples of "smaller" groups sitting inside of a group G. We show how to put groups together using the operation of direct product. With that quickly disposed of, we move into the more subtle and important concept of taking the quotient of a group.

7.1 Direct Products

As mentioned, we first give a construction for putting groups together.

Definition 7.1.1. *Suppose that $(G_i, \star_i)$ for $1 \leq i \leq n$ are all groups. Consider the Cartesian product of the sets $G_1, G_2, \ldots, G_n$, i.e.*

$$G_1 \times G_2 \times \cdots \times G_n = \{(a_1, a_2, \ldots, a_n) : a_i \in G_i \text{ for } 1 \leq i \leq n\}$$

Define an operation $\cdot$ on $G_1 \times G_2 \times \cdots \times G_n$ by letting

$$(a_1, a_2, \ldots, a_n) \cdot (b_1, b_2, \ldots, b_n) = (a_1 \star_1 b_1, a_2 \star_2 b_2, \ldots, a_n \star_n b_n)$$

Then $(G_1 \times G_2 \times \cdots \times G_n, \cdot)$ is a group which is called the (external) direct product *of $G_1, G_2, \ldots, G_n$.*

We first verify the claim that $(G_1 \times G_2 \times \cdots \times G_n, \cdot)$ is a group.

Proposition 7.1.2. *Suppose that $(G_i, \star_i)$ for $1 \leq i \leq n$ are all groups. The direct product defined above is a group with the following properties:*

- *The identity is $(e_1, e_2, \ldots, e_n)$ where e_i is the unique identity of G_i.*
- *Given $a_i \in G_i$ for all i, we have*

  $$(a_1, a_2, \ldots, a_n)^{-1} = (a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}).$$

  where a_i^{-1} is the inverse of a_i in the group G_i.
- $|G_1 \times G_2 \times \cdots \times G_n| = \prod_{i=1}^{n} |G_i|$, *i.e. the order of the direct product of the G_i is the product of the orders of the G_i.*

Proof. We first check that $\cdot$ is associative. Suppose that $a_i, b_i, c_i \in G_i$ for $1 \leq i \leq n$. We have

$$\begin{aligned}((a_1, a_2, \ldots, a_n) \cdot (b_1, b_2, \ldots, b_n)) \cdot (c_1, c_2, \ldots, c_n) &= (a_1 \star_1 b_1, a_2 \star_2 b_2, \ldots, a_n \star_n b_n) \cdot (c_1, c_2, \ldots, c_n) \\ &= ((a_1 \star_1 b_1) \star_1 c_1, (a_2 \star_2 b_2) \star_2 c_2, \ldots, (a_n \star_n b_n) \star_n c_n) \\ &= (a_1 \star_1 (b_1 \star_1 c_1), a_2 \star_2 (b_2 \star_2 c_2), \ldots, a_n \star_n (b_n \star_n c_n)) \\ &= (a_1, a_2, \ldots, a_n) \cdot (b_1 \star_1 c_1, b_2 \star_2 c_2, \ldots, b_n \star_n c_n) \\ &= (a_1, a_2, \ldots, a_n) \cdot ((b_1, b_2, \ldots, b_n) \cdot (c_1, c_2, \ldots, c_n)).\end{aligned}$$

Let e_i be the identity of G_i. We now check that $(e_1, e_2, \ldots, e_n)$ is an identity of the direct product. Let $a_i \in G_i$ for all i. We have

$$\begin{aligned}(a_1, a_2, \ldots, a_n) \cdot (e_1, e_2, \ldots, e_n) &= (a_1 \star_1 e_1, a_2 \star_2 e_2, \ldots, a_n \star_n e_n) \\ &= (a_1, a_2, \ldots, a_n)\end{aligned}$$

and

$$\begin{aligned}(e_1, e_2, \ldots, e_n) \cdot (a_1, a_2, \ldots, a_n) &= (e_1 \star_1 a_1, e_2 \star_2 a_2, \ldots, e_n \star_n a_n) \\ &= (a_1, a_2, \ldots, a_n)\end{aligned}$$

hence $(e_1, e_2, \ldots, e_n)$ is an identity for the direct product.

We finally check the claim about inverses inverses. Let $a_i \in G_i$ for $1 \leq i \leq n$. For each i, let a_i^{-1} be the inverse of a_i in G_i. We then have

$$\begin{aligned}(a_1, a_2, \ldots, a_n) \cdot (a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}) &= (a_1 \star_1 a_1^{-1}, a_2 \star_2 a_2^{-1}, \ldots, a_n \star_n a_n^{-1}) \\ &= (e_1, e_2, \ldots, e_n)\end{aligned}$$

and

$$\begin{aligned}(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}) \cdot (a_1, a_2, \ldots, a_n) &= (a_1^{-1} \star_1 a_1, a_2^{-1} \star_2 a_2, \ldots, a_n^{-1} \star_n a_n) \\ &= (e_1, e_2, \ldots, e_n)\end{aligned}$$

hence $(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1})$ is an inverse of $(a_1, a_2, \ldots, a_n)$.

Finally, since there are $|G_1|$ elements to put in the first coordinate of the n-tuple, $|G_2|$ elements to put in the second coordinates, etc., it follows that

$$|G_1 \times G_2 \times \cdots \times G_n| = \prod_{i=1}^{n} |G_i|.$$

□

For example, consider the group $G = S_3 \times \mathbb{Z}$ (where we are considering $\mathbb{Z}$ as a group under addition). Elements of G are ordered pairs (σ, n) where $\sigma \in S_3$ and $n \in \mathbb{Z}$. For example, $((1\ 2), 8)$, $(id, -6)$, and $((1\ 3\ 2), 42)$ are all elements of G. The group operation on G is obtained by working in each coordinate separately and performing the corresponding group operation there. For example, we have

$$((1\ 2), 8) \cdot ((1\ 3\ 2), 42) = ((1\ 2) \circ (1\ 3\ 2), 8 + 42) = ((1\ 3), 50).$$

As you can tell, the direct product puts two groups together in a manner that makes them completely ignore each other. Each coordinate goes about doing its business without interacting with the others at all.

Proposition 7.1.3. *The group $G_1 \times G_2 \times \cdots \times G_n$ is abelian if and only if each G_i is abelian.*

Proof. For each i, let $\star_i$ be the group of operation of G_i.

Suppose first that each G_i is abelian. Suppose that $a_i, b_i \in G_i$ for $1 \leq i \leq n$. We then have

$$\begin{aligned}(a_1, a_2, \ldots, a_n) \cdot (b_1, b_2, \ldots, b_n) &= (a_1 \star_1 b_1, a_2 \star_2 b_2, \ldots, a_n \star_n b_n) \\ &= (b_1 \star_1 a_1, b_2 \star_2 a_2, \ldots, b_n \star_n a_n) \\ &= (b_1, b_2, \ldots, b_n) \cdot (a_1, a_2, \ldots, a_n).\end{aligned}$$

Therefore, $G_1 \times G_2 \times \cdots \times G_n$ is abelian.

Suppose conversely that $G_1 \times G_2 \times \cdots \times G_n$ is abelian. Fix i with $1 \leq i \leq n$. Suppose that $a_i, b_i \in G_i$. Consider the elements $(e_1, \ldots, e_{i-1}, a_i, e_i, \ldots, e_n)$ and $(e_1, \ldots, e_{i-1}, b_i, e_i, \ldots, e_n)$ in $G_1 \times G_2 \times \cdots \times G_n$. Using the fact that the direction product is abelian, we see that

$$\begin{aligned}&(e_1, \ldots, e_{i-1}, a_i \star_i b_i, e_{i+1}, \ldots, e_n) \\ &\quad = (e_1 \star_1 e_1, \ldots, e_{i-1} \star_{i-1} e_{i-1}, a_i \star_i b_i, e_{i+1} \star_{i+1} e_{i+1}, \ldots, e_n \star_n e_n) \\ &\quad = (e_1, \ldots, e_{i-1}, a_i, e_{i+1}, \ldots, e_n) \cdot (e_1, \ldots, e_{i-1}, b_i, e_{i+1}, \ldots, e_n) \\ &\quad = (e_1, \ldots, e_{i-1}, b_i, e_{i+1}, \ldots, e_n) \cdot (e_1, \ldots, e_{i-1}, a_i, e_{i+1}, \ldots, e_n) \\ &\quad = (e_1 \star_1 e_1, \ldots, e_{i-1} \star_{i-1} e_{i-1}, b_i \star_i i_i, e_{i+1} \star_{i+1} e_{i+1}, \ldots, e_n \star_n e_n) \\ &\quad = (e_1, \ldots, e_{i-1}, b_i \star_i a_i, e_{i+1}, \ldots, e_n).\end{aligned}$$

Comparing the i^{th} coordinates of the first and last tuple, we conclude that $a_i \star_i b_i = b_i \star_i a_i$. Therefore, G_i is abelian. □

We can build all sorts of groups with this construction. For example $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is an abelian group of order 4 with elements

$$\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} = \{(\overline{0}, \overline{0}), (\overline{0}, \overline{1}), (\overline{1}, \overline{0}), (\overline{1}, \overline{1})\}$$

In this group, the element $(\overline{0}, \overline{0})$) is the identity and all other elements have order 2. We can also use this construction to build nonabelian groups of various orders. For example $S_3 \times \mathbb{Z}/2\mathbb{Z}$ is a nonabelian group of order 12.

Proposition 7.1.4. *Let $G_1, G_2, \ldots, G_n$ be groups. Let $a_i \in G_i$ for $1 \leq i \leq n$. The order of the element $(a_1, a_2, \ldots, a_n) \in G_1 \times G_2 \times \cdots \times G_n$ is the least common multiple of the orders of the a_i's in G_i.*

Proof. For each i, let $m_i = |a_i|$, so m_i is the order of a_i in the group G_i. Now since the group operation in the direct product works in each coordinate separately, a simple induction shows that

$$(a_1, a_2, \ldots, a_n)^k = (a_1^k, a_2^k, \ldots, a_n^k)$$

for all $k \in \mathbb{N}^+$. Now if $m_i \mid k$ for each i, then $a_i^k = e_i$ for each i, so

$$(a_1, a_2, \ldots, a_n)^k = (a_1^k, a_2^k, \ldots, a_n^k) = (e_1, e_2, \ldots, e_n)$$

Conversely, suppose that $k \in \mathbb{N}^+$ is such that there exists i with $m_i \nmid k$. For such an i, we then have that $a_i^k \neq e_i$, so

$$(a_1, a_2, \ldots, a_n)^k = (a_1^k, a_2^k, \ldots, a_n^k) \neq (e_1, e_2, \ldots, e_n)$$

It follows that $(a_1, a_2, \ldots, a_n)^k = (e_1, e_2, \ldots, e_n)$ if and only if $m_i \mid k$ for all i. Since $|(a_1, a_2, \ldots, a_n)|$ is the least $k \in \mathbb{N}^+$ with $(a_1, a_2, \ldots, a_n)^k = (e_1, e_2, \ldots, e_n)$, it follows that $|(a_1, a_2, \ldots, a_n)|$ is the least $k \in \mathbb{N}^+$ satisfying $m_i \mid k$ for all i, which is to say that $|(a_1, a_2, \ldots, a_n)|$ is the least common multiple of the m_i. □

For example, suppose that we are working in the group $S_4 \times \mathbb{Z}/42\mathbb{Z}$ and we consider the element $((1\ 4\ 2\ 3), \overline{7})$. Since $|(1\ 4\ 2\ 3)| = 4$ in S_4 and $|\overline{7}| = 6$ in $\mathbb{Z}/42\mathbb{Z}$, it follows that the order of $((1\ 4\ 2\ 3), \overline{7})$ in $S_4 \times \mathbb{Z}/42\mathbb{Z}$ equals $\text{lcm}(4, 6) = 12$.

For another example, the group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is an abelian group of order 8 in which every nonidentity element has order 2. Generalizing this construction by taking n copies of $\mathbb{Z}/2\mathbb{Z}$, we see how to construct an abelian group of order 2^n in which every nonidentity element has order 2.

7.2 Quotients of Abelian Groups

As mentioned in the introduction of this section, the quotient construction is one of the most subtle but important construction in group theory (and algebra in general). As suggested, the quotient of a group G should somehow be "smaller" than G. Suppose that we have a group G and a subgroup H of G. We want to build a new group which we will denote by G/H. We already know that using H we get an equivalence relation $\sim_H$ which breaks up the group G into cosets. The fundamental idea is to make a new group whose elements are these cosets. However, the first question we need to ask ourselves is whether we will use left cosets or right cosets (i.e. which version of $\sim_H$ will we work with). To avoid this issue, and to focus on the key elements of the construction first, we will begin by assuming that our group is abelian so that there is no worry about which side we work on. As we will see, this abelian assumption helps immensely in other ways as well.

Suppose then that G is an abelian group and that H is a subgroup of G. We let G/H be the set of all (left) cosets of H in G, so elements of G/H are of the form aH for some $a \in G$. In the language established when discussing equivalence relations, we are looking at the set $G/\sim_H$, but we simplify the notation and just write G/H. Now that we have decided what the elements of G/H are, we need to define the operation. How should we define the product of aH and bH? The natural idea is simply to define $(aH) \cdot (bH) = (ab)H$. That is, given the two cosets aH and bH, take the product ab in G and form its coset. Alarm bells should immediately go off in your head and you should be asking: Is this well-defined? After all, a coset has many different representatives. What if $aH = cH$ and $bH = dH$? On the one hand, the product should be $(ab)H$ and on the other the product should be $(cd)H$. Are these necessarily the same? Recall that we are dealing with equivalence classes so $aH = bH$ if and only if $a \sim_H b$. The next proposition says that everything is indeed well-defined in the abelian case we are currently considering.

Proposition 7.2.1. *Let G be an abelian group and let H be a subgroup of G. Suppose that $a \sim_H c$ and $b \sim_H d$. We then have that $ab \sim_H cd$.*

Proof. Fix $h, \ell \in H$ such that $ah = c$ and $b\ell = d$. We then have that

$$cd = ahb\ell = abh\ell.$$

Now $h\ell \in H$ because H is a subgroup of G. Since $cd = (ab)(h\ell)$, it follows that $ab \sim_H cd$. □

Notice how fundamentally we used the fact that G was abelian in this proof to write $hb = bh$. Overcoming this apparent stumbling block will be our primary focus when we get to nonabelian groups.

For example, suppose that $G = \mathbb{R}^2$ and $H = \{(0, b) : b \in \mathbb{R}\}$ is the y-axis. Again, we will write $(a, b) + H$ for the left coset of H In G. The elements of the quotient group G/H are the left cosets of H in G, which we know are the set of vertical lines in the plane. Let's examine how we add two elements of G/H. The definition says that we add left cosets by finding representatives of those cosets, adding those representatives, and then taking the left coset of the result. In other words, to add two vertical lines, we pick points on the lines, add those points, and output the line containing the result. For example, we add the cosets $(3, 2) + H$ and $(4, -7) + H$ by computing $(3, 2) + (4, -7) = (7, -5)$ and outputting the corresponding coset $(7, -5) + H$. In other words, we have

$$((3, 2) + H) + ((4, -7) + H) = (7, -5) + H.$$

Now we could have chosen different representatives of those two cosets. For example, we have $(3, 2) + H = (3, 16) + H$ (after all both are on the line $x = 3$) and $(4, -7) + H = (4, 1) + H$, and if we calculate the sum using these representatives we see that

$$((3, 16) + H) + ((4, 1) + H) = (7, 17) + H.$$

Now although the elements of G given by $(7, -5)$ and $(7, 17)$ are different, the cosets $(7, -5) + H$ and $(7, 17) + H$ are equal because $(7, -5) \sim_H (7, 17)$.

We are now ready to formally define the quotient of an abelian group G by a subgroup H. We verify that the given definition really is a group in the proposition immediately after the definition.

Definition 7.2.2. *Let G be an abelian group and let H be a subgroup of G. We define a new group, called the* quotient *of G by H and denoted G/H, by letting the elements be the left cosets of H in G (i.e. the equivalence classes of G under $\sim_H$), and defining the binary operation $aH \cdot bH = (ab)H$. The identity is eH (where e is the identity of G) and the inverse of aH is $a^{-1}H$.*

Proposition 7.2.3. *Let G be an abelian group and let H be a subgroup of G. The set G/H with the operation just defined is indeed a group with $|G/H| = [G : H]$. Furthermore, it is an abelian group.*

Proof. We verified that the operation $aH \cdot bH = (ab)H$ is well-defined in Proposition 7.2.1. With that in hand, we just need to check the group axioms.

We first check that the $\cdot$ is an associative operation on G/H. For any $a, b, c \in G$ we have

$$\begin{aligned}(aH \cdot bH) \cdot cH &= (ab)H \cdot cH \\ &= ((ab)c)H \\ &= (a(bc))H && \text{(since } \cdot \text{ is associative on } G) \\ &= aH \cdot (bc)H \\ &= aH \cdot (bH \cdot cH).\end{aligned}$$

We next check that eH is an identity. For any $a \in G$ we have

$$aH \cdot eH = (ae)H = aH$$

and

$$eH \cdot aH = (ea)H = aH.$$

For inverses, notice that given any $a \in G$, we have

$$aH \cdot a^{-1}H = (aa^{-1})H = eH$$

and

$$a^{-1}H \cdot aH = (a^{-1}a)H = eH.$$

Thus, G/H is indeed a group, and it has order $[G : H]$ because the elements are the left cosets of H in G. Finally, we verify that G/H is abelian by noting that for any $a, b \in G$, we have

$$\begin{aligned}aH \cdot bH &= (ab)H \\ &= (ba)H && \text{(since } G \text{ is abelian)} \\ &= bH \cdot aH.\end{aligned}$$

□

Here is another example. Suppose that $G = U(\mathbb{Z}/18\mathbb{Z})$ and let $H = \langle 17 \rangle = \{\overline{1}, \overline{17}\}$. We then have that the left cosets of H in G are:

- $\overline{1}H = \overline{17}H = \{\overline{1}, \overline{17}\}$
- $\overline{5}H = \overline{13}H = \{\overline{5}, \overline{13}\}$
- $\overline{7}H = \overline{11}H = \{\overline{7}, \overline{11}\}$

Therefore, $|G/H| = 3$. To multiply two cosets, we choose representatives and multiply. For example, we could calculate

$$\overline{5}H \cdot \overline{7}H = (\overline{5} \cdot \overline{7})H = \overline{17}H.$$

We can multiply the exact same two cosets using different representatives. For example, we have $\overline{7}H = \overline{11}H$, so we could calculate

$$\overline{5}H \cdot \overline{11}H = (\overline{5} \cdot \overline{11})H = \overline{1}H.$$

Notice that we obtained the same answer since $\overline{1}H = \overline{17}H$. Now there is no canonical choice of representatives for the various cosets, so if you want to give each element of G/H a unique "name", then you simply have to pick which representative of each coset you will use. We will choose (somewhat arbitrarily) to view G/H as the following:

$$G/H = \{\overline{1}H, \overline{5}H, \overline{7}H\}.$$

Here is the Cayley table of G/H using these choices of representatives.

$\cdot$	$\overline{1}H$	$\overline{5}H$	$\overline{7}H$
$\overline{1}H$	$\overline{1}H$	$\overline{5}H$	$\overline{7}H$
$\overline{5}H$	$\overline{5}H$	$\overline{7}H$	$\overline{1}H$
$\overline{7}H$	$\overline{7}H$	$\overline{1}H$	$\overline{5}H$

Again, notice that using the definition we have $\overline{5}H \cdot \overline{7}H = \overline{17}H$, but since $\overline{17}$ was not one of our chosen representatives and $\overline{17}H = \overline{1}H$ where $\overline{1}$ is one of our chosen representatives, we used $\overline{5}H \cdot \overline{7}H = \overline{1}H$ in the above table.

Finally, let us take a moment to realize that we have been dealing with quotient groups all along when working with $\mathbb{Z}/n\mathbb{Z}$. This group is exactly the quotient of the group $G = \mathbb{Z}$ under addition by the subgroup $H = n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$. Recall from Section 6.1.1 that given $a, b \in \mathbb{Z}$, we have

$$a \sim_{n\mathbb{Z}} b \iff a \equiv_n b$$

so the left cosets of $n\mathbb{Z}$ are precisely the equivalence classes of $\equiv_n$. Stated in symbols, if $\overline{a}$ is the equivalence class of a under $\equiv_n$, then $\overline{a} = a + n\mathbb{Z}$ (we are again using $+$ in left cosets because that is the operation in $\mathbb{Z}$). Furthermore, our definition of the operation in $\mathbb{Z}/n\mathbb{Z}$ was given by

$$\overline{a} + \overline{b} = \overline{a + b}.$$

However, in terms of cosets, this simply says

$$(a + n\mathbb{Z}) + (b + n\mathbb{Z}) = (a + b) + n\mathbb{Z}$$

which is exactly our definition of the operation in the quotient $\mathbb{Z}/n\mathbb{Z}$. Therefore, the group $\mathbb{Z}/n\mathbb{Z}$ is precisely the quotient of the group $\mathbb{Z}$ by the subgroup $n\mathbb{Z}$, which is the reason why we have used that notation all along.

7.3 Normal Subgroups

In discussing quotients of abelian groups, we used the abelian property in two places. The first place was to avoid choosing whether we were dealing with left cosets or right cosets (since they will be the same whenever G is abelian). The second place we used commutativity was in checking that the group operation on the quotient was well-defined. Now the first choice of left/right cosets doesn't seem so hard to overcome because we can simply choose one. For the moment, say we choose to work with left cosets. However, the well-defined issue looks like it might be hard to overcome because we used commutativity in what appears to be a very

central manner. Let us recall our proof. We had a subgroup H of an abelian group G, and we were assuming $a \sim_H c$ and $b \sim_H d$. We then fixed $h, \ell \in H$ with $ah = c$ and $b\ell = d$ and observed that

$$cd = ahb\ell = abh\ell$$

hence $ab \sim_H cd$ because $h\ell \in H$. Notice the key use of commutativity to write $hb = bh$. In fact, we can easily see that the operation is not always well-defined if G is nonabeilan. Consider our usual example of $G = S_3$ with $H = \langle(1\ 2)\rangle = \{id, (1\ 2)\}$. We computed the left cosets in Section 6.1:

1. $idH = (1\ 2)H = \{id, (1\ 2)\}$
2. $(1\ 3)H = (1\ 2\ 3)H = \{(1\ 3), (1\ 2\ 3)\}$
3. $(2\ 3)H = (1\ 3\ 2)H = \{(2\ 3), (1\ 3\ 2)\}$

Thus, we have $id \sim_H (1\ 2)$ and $(1\ 3) \sim_H (1\ 2\ 3)$. Now $id(1\ 3) = (1\ 3)$ and $(1\ 2)(1\ 2\ 3) = (2\ 3)$, but a quick look at the left cosets shows that $(1\ 3) \not\sim_H (2\ 3)$. Hence

$$id(1\ 3) \not\sim_H (1\ 2)(1\ 2\ 3).$$

In other words, the operation $\sigma H \cdot \tau H = (\sigma\tau)H$ is not well-defined because on the one hand we would have

$$idH \cdot (1\ 3)H = (1\ 3)H$$

and on the other hand

$$(1\ 2)H \cdot (1\ 2\ 3)H = (2\ 3)H$$

which contradicts the definition of a function since $(1\ 3)H \neq (2\ 3)H$.

With the use of commutativity so essential in our well-defined proof and a general counterexample in hand, it might appear that we have little hope in dealing with nonabelian groups. We just showed that we have no hope for an arbitrary subgroup H of G, but maybe we get by with some special subgroups. Recall again our proof that used

$$cd = ahb\ell = abh\ell$$

We do really need to get b next to a and with a little thought we can see some wiggle room. For example, one tiny way we could get by is if $H \subseteq Z(G)$. Under this assumption, the elements elements of H commute with everything in G, so the above argument still works. This suffices, but it is quite restrictive. Notice that we could make things work with even less. We would be able to get by if we weakened the assumption that H commutes with all elements of G to the following:

$$\text{For all } g \in G \text{ and all } h \in H, \text{ there exists } k \in H \text{ with } hg = gk.$$

In other words, we do not need to be able to move b past h in a manner which does not disturb h at all, but instead we could get by with moving b past h in a manner which changes h to perhaps a different element of H.

It turns out that the above condition on a subgroup H of a group G is equivalent to many other fundamental concepts. First, we introduce the following definition.

Definition 7.3.1. *Let G be a group and let H be a subgroup of G. Given $g \in G$, we define*

$$gHg^{-1} = \{ghg^{-1} : g \in G, h \in H\}$$

Proposition 7.3.2. *Let G be a group and let H be a subgroup of G. The following are equivalent.*

1. *For all $g \in G$ and all $h \in H$, there exists $k \in H$ with $hg = gk$.*

2. *For all $g \in G$ and all $h \in H$, there exists $k \in H$ with $gh = kg$.*

3. *$g^{-1}hg \in H$ for all $g \in G$ and all $h \in H$.*

4. *$ghg^{-1} \in H$ for all $g \in G$ and all $h \in H$.*

5. *$gHg^{-1} \subseteq H$ for all $g \in G$.*

6. *$gHg^{-1} = H$ for all $g \in G$.*

7. *$Hg \subseteq gH$ for all $g \in G$.*

8. *$gH \subseteq Hg$ for all $g \in G$.*

9. *$gH = Hg$ for all $g \in G$.*

Proof. $1 \Rightarrow 2$: Suppose that we know 1. Let $g \in G$ and let $h \in H$. Applying 1 with $g^{-1} \in G$ and $h \in H$, we may fix $k \in H$ with $hg^{-1} = g^{-1}k$. Multiplying on the left by g we see that $ghg^{-1} = k$, and then multiplying on the right by g we conclude that $gh = kg$.

$2 \Rightarrow 1$: Suppose that we know 2. Let $g \in G$ and let $h \in H$. Applying 2 with $g^{-1} \in G$ and $h \in H$, we may fix $k \in H$ with $g^{-1}h = kg^{-1}$. Multiplying on the right by g we see that $g^{-1}hg = k$, and then multiplying on the left by g we conclude that $hg = gk$.

$1 \Rightarrow 3$: Suppose that we know 1. Let $g \in G$ and let $h \in H$. By 1, we may fix $k \in H$ with $hg = gk$. Multiplying on the left by g^{-1}, we see that $g^{-1}hg = k \in H$. Since $g \in G$ and $h \in H$ were arbitrary, we conclude that $g^{-1}hg \in H$ for all $g \in G$ and all $h \in H$.

$3 \Rightarrow 1$: Suppose that we know 3. Let $g \in G$ and let $h \in H$. Now

$$hg = ehg = (gg^{-1})hg = g(g^{-1}hg)$$

and by 3 we know that $g^{-1}hg \in H$. Part 1 follows.

$2 \Leftrightarrow 4$: This follows in exactly that same manner as $1 \Leftrightarrow 3$.

$4 \Leftrightarrow 5$: These two are simply restatements of each other.

$5 \Rightarrow 6$: Suppose we know 5. To prove 6, we need only prove that $H \subseteq gHg^{-1}$ for all $g \in G$ (since the reverse containment is given to us). Let $g \in G$ and let $h \in H$. By 5 applied to g^{-1}, we see that $g^{-1}H(g^{-1})^{-1} \subseteq H$, hence $g^{-1}Hg \subseteq H$ and in particular we conclude that $g^{-1}hg \in H$. Since

$$h = ehe = (gg^{-1})h(gg^{-1}) = g(g^{-1}hg)g^{-1}$$

and $g^{-1}hg \in H$, we see that $h \in gHg^{-1}$. Since $g \in G$ and $h \in H$ were arbitrary, it follows that $H \subseteq gHg^{-1}$ for all $g \in G$.

$6 \Rightarrow 5$: This is trivial.

$1 \Leftrightarrow 7$: These two are simply restatements of each other.

$2 \Leftrightarrow 8$: These two are simply restatements of each other.

$9 \Rightarrow 8$: This is trivial.

$7 \Rightarrow 9$: Suppose that we know 7. Since $7 \Rightarrow 1 \Rightarrow 2 \Rightarrow 8$, it follows that we know 8 as well. Putting 7 and 8 together we conclude 9. □

As the Proposition shows, the condition we are seeking to ensure that multiplication of left cosets is well-defined is equivalent to the condition that the left cosets of H in G are equal to the right cosets of H in G. Thus, by adopting that condition we automatically get rid of the other problematic question of which side to work on. This condition is shaping up to be so useful that we given the subgroups which satisfy it a special name.

Definition 7.3.3. *Let G be a group and let H be a subgroup of G. We say that H is a* normal subgroup *of G if $gHg^{-1} \subseteq H$ for all $g \in G$ (or equivalently any of properties in the previous proposition hold).*

Our entire goal in defining and exploring the concept of a normal subgroup H of a group G was to allow us to prove that multiplication of left cosets via representatives is well-defined. It turns out that this condition is precisely equivalent to this operation being well-defined.

Proposition 7.3.4. *Let G be an group and let H be a subgroup of G. The following are equivalenct.*

1. *H is a normal subgroup of G.*
2. *Whenever $a, b, c, d \in G$ with $a \sim_H c$ and $b \sim_H d$, we have $ab \sim_H cd$. (Here, $\sim_H$ is the equivalence relation corresponding to left cosets)*

Proof. We first prove that 1 implies 2. Suppose that $a, b, c, d \in G$ with $a \sim_H c$ and $b \sim_H d$. Fix $h, \ell \in H$ such that $ah = c$ and $b\ell = d$. Since H is a normal subgroup of G, we may fix $k \in H$ with $hb = bk$. We then have that

$$cd = ahb\ell = abk\ell.$$

Now $k\ell \in H$ because H is a subgroup of G. Since $cd = (ab)(k\ell)$, it follows that $ab \sim_H cd$.

We now prove that 2 implies 1. We prove that H is a normal subgroup of G by showing that $g^{-1}hg \in H$ for all $g \in G$ and $h \in H$. Let $g \in G$ and let $h \in H$. Notice that we have $e \sim_H h$ because $eh = h$ and $g \sim_H g$ because $ge = g$. Since we are assuming 2 it follows that $eg \sim_H hg$ and hence $g \sim_H hg$. Fix $k \in H$ with $gk = hg$. Multiply on the left by g^{-1} we get $k = g^{-1}hg$ so $g^{-1}hg \in H$. The result follows. □

Proposition 7.3.5. *Let G be a group and let H be a subgroup of G. If $H \subseteq Z(G)$, then H is a normal subgroup of G.*

Proof. For any $g \in G$ and $h \in H$, we have $hg = gh$ because $h \in Z(G)$. Therefore, H is a normal subgroup of G by Condition 1 above. □

Example 7.3.6. *Suppose that $n \geq 4$ is even and write $n = 2k$ for $k \in \mathbb{N}^+$. Since $Z(D_n) = \{id, r^k\}$, the subgroup $\{id, r^k\}$ is a normal subgroup of D_n.*

Example 7.3.7. *The subgroup $A_3 = \langle (1\ 2\ 3) \rangle$ is a normal subgroup of S_3.*

Proof. We have

$$A_3 = \langle (1\ 2\ 3) \rangle = \{id, (1\ 2\ 3), (1\ 3\ 2)\}.$$

One proof is to notice that

$$[S_3 : A_3] = \frac{|S_3|}{|A_3|} = \frac{6}{3} = 2$$

so A_3 is a normal subgroup of S_3 by the homework. In fact, using that homework problem, it follows that A_n is a normal subgroup of S_n for all n. However, you can also do a direct computation of the cosets. The left cosets of A_3 in S_3 are

- $idA_3 = (1\ 2\ 3)A_3 = (1\ 3\ 2)A_3 = \{id, (1\ 2\ 3), (1\ 3\ 2)\}$
- $(1\ 2)A_3 = (1\ 3)A_3 = (2\ 3)A_3 = \{(1\ 2), (1\ 3), (2\ 3)\}$

The right cosets of A_3 in S_3 are

- $A_3id = A_3(1\ 2\ 3) = A_3(1\ 3\ 2) = \{id, (1\ 2\ 3), (1\ 3\ 2)\}$
- $A_3(1\ 2) = A_3(1\ 3) = A_3(2\ 3) = \{(1\ 2), (1\ 3), (2\ 3)\}$

Thus, $\sigma A_3 = A_3\sigma$ for all $\sigma \in S_3$ and hence A_3 is a normal subgroup of S_3. □

Example 7.3.8. *For any $n \geq 3$, the subgroup $H = \langle r \rangle$ is a normal subgroup of D_n.*

Proof. Again, one quick proof is to notice that $[D_n : H] = 2$ and use the homework. However, on a previous homework, you already calculated the left cosets and the right cosets. We have that the left cosets of H in D_n are:

- $idH = rH = r^2H = \cdots = r^{n-1}H = \{id, r, r^2, \ldots, r^{n-1}\}$
- $sH = rsH = r^2sH = \cdots = r^{n-1}sH = \{s, rs, r^2s, \ldots, r^{n-1}s\}$

and the right cosets of H in D_n are:

- $Hid = Hr = Hr^2 = \cdots = Hr^{n-1} = \{id, r, r^2, \ldots, r^{n-1}\}$
- $Hs = Hrs = Hr^2s = \cdots = Hr^{n-1}s = \{s, rs, r^2s, \ldots, r^{n-1}s\}$

Thus, $aH = Ha$ for all $a \in D_n$ and hence H is a normal subgroup of D_n. □

7.4 Quotient Groups

Definition 7.4.1. *Let G be an group and let H be a normal subgroup of G. We define a new group, called the* quotient *of G by H and denoted G/H, by letting the elements be the left cosets of H in G (i.e. the equivalence classes of G under $\sim_H$), and defining the binary operation $aH \cdot bH = (ab)H$. The identity is eH (where e is the identity of G) and the inverse of aH is $a^{-1}H$.*

Proposition 7.4.2. *Let G be a group and let H be a normal subgroup of G. The set G/H with the operation just defined is indeed a group with $|G/H| = [G : H]$.*

Proof. We verified that the operation $aH \cdot bH = (ab)H$ is well-defined in Proposition 7.3.4. With that in hand, we just need to check the group axioms.

We first check that $\cdot$ is an associative operation on G/H. For any $a, b, c \in G$ we have

$$\begin{aligned}(aH \cdot bH) \cdot cH &= (ab)H \cdot cH \\ &= ((ab)c)H \\ &= (a(bc))H && \text{(since } \cdot \text{ is associative on } G) \\ &= aH \cdot (bc)H \\ &= aH \cdot (bH \cdot cH).\end{aligned}$$

We next check that eH is an identity. For any $a \in G$ we have

$$aH \cdot eH = (ae)H = aH$$

and

$$eH \cdot aH = (ea)H = aH.$$

For inverses, notice that given any $a \in G$, we have

$$aH \cdot a^{-1}H = (aa^{-1})H = eH$$

and

$$a^{-1}H \cdot aH = (a^{-1}a)H = eH.$$

Thus, G/H is indeed a group, and it has order $[G : H]$ because the elements are the left cosets of H in G. □

For example, suppose that $G = D_4$ and $H = Z(G) = \{id, r^2\}$. We know from above that H is a normal subgroup of G. The left cosets (and hence right cosets because H is normal in G) of H in G are:

- $idH = r^2H = \{id, r^2\}$
- $rH = r^3H = \{r, r^3\}$
- $sH = r^2sH = \{s, r^2s\}$
- $rsH = r^3sH = \{rs, r^3s\}$

As usual, there are no "best" choices of representatives for these cosets when we consider G/H. We choose to take

$$G/H = \{idH, rH, sH, rsH\}.$$

The Cayley table of G/H using these representatives is:

$\cdot$	idH	rH	sH	rsH
idH	idH	rH	sH	rsH
rH	rH	idH	rsH	sH
sH	sH	rsH	idH	rH
rsH	rsH	sH	rH	idH

Notice that we had to switch to our "chosen" representatives several times when constructing this table. For example, we have

$$rH \cdot rH = r^2H = idH$$

and

$$sH \cdot rH = srH = r^{-1}sH = r^3sH = rsH.$$

Examining the table, we see a few interesting facts. The group G/H is abelian even though G is not abelian. Furthermore, every nonidentity element of G/H has order 2 even though r itself has order 4. We next see how the order of an element in the quotient relates to the order of the representative in the original group.

Proposition 7.4.3. *Let G be a group and let H be a normal subgroup of G. Suppose that $a \in G$ has finite order. The order of aH (in the group G/H) is finite and divides the order of a (in the group G).*

Proof. Let $n = |a|$ (in the group G). We have $a^n = e$, so

$$(aH)^n = a^nH = eH.$$

Now eH is the identity of G/H, so we have found some power of the element aH which gives the identity in G/H. Thus, aH has finite order. Let $m = |aH|$ (in the group G/H). Since we checked that n is a power of aH giving the identity, it follow from Proposition 4.2.5 that $m \mid n$. □

It is possible that $|aH|$ is strictly smaller than $|a|$. In the above example of D_4, we have $|r| = 4$ but $|rH| = 2$. Notice however that $2 \mid 4$ as the previous proposition proves must be the case.

We now show how it is possible to prove theorems about groups using quotients and induction. The general idea is as follows. Given a group G, proper subgroups of G and nontrivial quotients of G (that is by normal subgroups other than G) are "smaller" than G. So the idea is to prove a result about finite groups by using induction on the order of the group. By the inductive hypothesis, we know information about the proper subgroups and nontrivial quotients, so the hope is to piece together that information to prove the result about G. We give an example of this technique by proving the following theorem.

Theorem 7.4.4. *Let $p \in \mathbb{N}^+$ be prime. If G is a finite abelian group with $p \mid |G|$, then G has an element of order p.*

Before jumping into the proof, we first establish the following useful lemma.

Lemma 7.4.5. *If G has an element of order $n \in \mathbb{N}^+$, then G has an element of order d for every positive $d \mid n$.*

Proof. Suppose that G has an element of order n, and fix $a \in G$ with $|a| = n$. Let $d \in \mathbb{N}^+$ with $d \mid n$. Fix $k \in \mathbb{N}^+$ with $kd = n$. Using Proposition 4.2.6 and the fact that $k \mid n$, we have

$$|a^k| = \frac{n}{\gcd(n,k)} = \frac{n}{k} = d$$

Thus a^k is an element of G with order d. □

We are now ready to prove the theorem.

Proof of Theorem 7.4.4. The proof is by induction on $|G|$. If $|G| = 1$, then the result is trivial because $p \nmid 1$ (if you don't like this vacuous base case, simply note the if $|G| = p$, then every nonidentity element of G has order p by Lagrange's Theorem). Suppose then that G is a finite group with $p \mid |G|$, and suppose that the result is true for all groups K satisfying $p \mid |K|$ and $|K| < |G|$. Fix $a \in G$ with $a \neq e$, and let $H = \langle a \rangle$. We then have that H is a normal subgroup of G because G is abelian. We now have two cases.

Case 1: Suppose that $p \mid |a|$. By the lemma, G has an element of order p.

Case 2: Suppose that $p \nmid |a|$. We have

$$|G/H| = [G : H] = \frac{|G|}{|H|}$$

so $|G| = |H| \cdot |G/H|$. Since $p \mid |G|$ and $p \nmid |H|$ (because $|H| = |a|$), it follows that $p \mid |G/H|$. Now $|G/H| < |G|$ because $|H| > 1$, so by induction there exists an element $bH \in G/H$ with $|bH| = p$. By Proposition 7.4.3, we may conclude that $p \mid |b|$. Therefore, by the lemma, G has an element of order p.

In either case, we have concluded that G has an element of order p. The result follows by induction. □

Notice where we used the two fundamental assumptions that p is prime and that G is abelian. For the prime assumption, we used the key fact that if $p \mid ab$, then either $p \mid a$ or $p \mid b$. In fact, the result is not true if you leave out the assumption that p is prime. The abelian group $U(\mathbb{Z}/8\mathbb{Z})$ has order 4 but it has no element of order 4.

We made use of the abelian assumption to get that H was a normal subgroup of G without any work. In general, with a little massaging, we could slightly alter the above proof as long as we could assume that every group has some normal subgroup other than the two trivial normal subgroups $\{e\}$ and G (because this would allow us to either use induction on H or on G/H). Unfortunately, it is not in general true that every group always has such a normal subgroup. Those that do not are given a name.

Definition 7.4.6. *A group G is* simple *if $|G| > 1$ and the only normal subgroups of G are $\{e\}$ and G.*

A simple group is a group that we are unable to "break up" into a smaller normal subgroup H and corresponding smaller quotient G/H. They are the "atoms" of the groups and are analogous to the primes. In fact, for every prime p, the abelian group $\mathbb{Z}/p\mathbb{Z}$ is simple for the trivial reason that its only subgroups at all are $\{\overline{0}\}$ and all of $\mathbb{Z}/p\mathbb{Z}$ by Lagrange's Theorem. It turns out that these are the only simple abelian groups (see homework). Now if these were the only simple groups at all, there would not be much of problem because they are quite easy to get a handle on. However, there are infinitely many finite simple nonabelian groups. For example, we will see later that A_n is a simple group for all $n \geq 5$. The existence of these groups is a serious obstruction to any inductive proof for all groups in the above style. Nonetheless, the above theorem is true for all groups (including nonabelian ones), and the result is known as Cauchy's Theorem. We will prove it later using more advanced tools, but we will start that proof knowing that we have already handled the abelian case.

Chapter 8

Isomorphisms and Homomorphisms

8.1 Isomorphisms

Definitions and Examples

We have developed several ways to construct groups. We started with well known groups like $\mathbb{Z}$, $\mathbb{Q}$ and $GL_n(\mathbb{R})$. For there, we introduced the groups $\mathbb{Z}/n\mathbb{Z}$ and $U(\mathbb{Z}/n\mathbb{Z})$. After that, we developed our first family of nonabelian groups in the symmetric groups S_n. With those in hand, we obtained many other groups as subgroups of these, such as $SL_n(\mathbb{R})$, $O(n,\mathbb{R})$, A_n, and D_n. Finally, we build new groups from all of these using direct products and quotients.

With such a rich supply on groups now, it is time to realize that some of these groups are essentially the "same". For example, let $G = \mathbb{Z}/2\mathbb{Z}$, let $H = S_2$, and let K be the group $(\{T, F\}, \oplus)$ where $\oplus$ is "exclusive or" discussed in the first section. Here are the Cayley tables of this groups.

$+$	$\overline{0}$	$\overline{1}$
$\overline{0}$	$\overline{0}$	$\overline{1}$
$\overline{1}$	$\overline{1}$	$\overline{0}$

$\circ$	id	$(1\ 2)$
id	id	$(1\ 2)$
$(1\ 2)$	$(1\ 2)$	id

$\oplus$	F	T
F	F	T
T	T	F

Now of course these groups are different because the sets are completely different. The elements of G are equivalence classes and thus subsets of $\mathbb{Z}$, the elements of H are permutations of the set $\{1, 2\}$, and the elements of K are T and F. Furthermore the operations themselves have little in common since in G we have addition of cosets via representatives, in H we have function composition, and in K we have this funny logic operation. However, despite all these differences, a glance at the above tables tells us that there is a deeper "sameness" to them. For G and H, if we pair off $\overline{0}$ with id and pair off $\overline{1}$ and $(1\ 2)$, then we have provided a kind of "rosetta stone" for translating between the groups. This is formalized with the following definition.

Definition 8.1.1. *Let $(G, \cdot)$ and $(H, \star)$ be groups. An* isomorphism *from G to H is a function $\varphi\colon G \to H$ such that*

1. *φ is a bijection.*
2. *$\varphi(a \cdot b) = \varphi(a) \star \varphi(b)$ for all $a, b \in G$. In shorthand, φ preserves the group operation.*

Thus, an isomorphism $\varphi\colon G \to H$ is a pairing of elements of G with elements of H (that is the bijection part) in such a way that we can either operate on the G side first and then walk over to H, or equivalently can walk over to H with our elements first and then operate on that side. In our case of $G = \mathbb{Z}/2\mathbb{Z}$ and $H = S_2$, we define $\varphi\colon G \to H$ by letting $\varphi(\overline{0}) = id$ and $\varphi(\overline{1}) = (1\ 2)$. Clearly, φ is a bijection. Now we have to check four pairs to check the second property. We have

- $\varphi(\overline{0}+\overline{0}) = \varphi(\overline{0}) = id$ and $\varphi(\overline{0}) + \varphi(\overline{0}) = id \circ id = id$.
- $\varphi(\overline{0}+\overline{1}) = \varphi(\overline{1}) = (1\ 2)$ and $\varphi(\overline{0}) + \varphi(\overline{1}) = id \circ (1\ 2) = (1\ 2)$.
- $\varphi(\overline{1}+\overline{0}) = \varphi(\overline{1}) = (1\ 2)$ and $\varphi(\overline{1}) + \varphi(\overline{0}) = (1\ 2) \circ id = (1\ 2)$.
- $\varphi(\overline{1}+\overline{1}) = \varphi(\overline{0}) = id$ and $\varphi(\overline{1}) + \varphi(\overline{1}) = (1\ 2) \circ (1\ 2) = id$.

Therefore, $\varphi(a+b) = \varphi(a) \circ \varphi(b)$ for all $a, b \in \mathbb{Z}/2\mathbb{Z}$. Since we already noted that φ is a bijection, it follows that φ is an isomorphism. All of these checks are just really implicit in the above table. The bijection $\varphi\colon G \to H$ we defined pairs off $\overline{0}$ with id and pairs off $\overline{1}$ with $(1\ 2)$. Since this aligning of elements carries the table of G to the table of H as seen in the above tables, φ is an isomorphism.

Notice if instead we define $\psi\colon G \to H$ by letting $\psi(\overline{0}) = (1\ 2)$ and $\psi(\overline{1}) = id$, then ψ is not an isomorphism. To see this, we just need to find a counterexample to the second property. We have

$$\psi(\overline{0}+\overline{0}) = \psi(\overline{0}) = (1\ 2)$$

and

$$\psi(\overline{0}) + \psi(\overline{0}) = (1\ 2) \circ (1\ 2) = id$$

so

$$\psi(\overline{0}+\overline{0}) \neq \psi(\overline{0}) + \psi(\overline{0})$$

Essentially we are writing the Cayley tables as:

$+$	$\overline{0}$	$\overline{1}$
$\overline{0}$	$\overline{0}$	$\overline{1}$
$\overline{1}$	$\overline{1}$	$\overline{0}$

$\circ$	$(1\ 2)$	id
$(1\ 2)$	id	$(1\ 2)$
id	$(1\ 2)$	id

and noting that ψ does not carry the table of G to the table of H (as can be seen in the $(1,1)$ entry). Thus, even if one bijection is an isomorphism, there may be other bijections which are not. We make the following definition.

Definition 8.1.2. *Let G and H be groups. We say that G and H are* isomorphic*, and write $G \cong H$, if there exists an isomorphism $\varphi\colon G \to H$.*

In colloquial language, two groups G and H are isomorphic exactly when there is *some* way to pair off elements of G with H which maps the Cayley table of G onto the Cayley table of H. For example, we have $U(\mathbb{Z}/8\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ via the bijection

$$\varphi(\overline{1}) = (\overline{0},\overline{0}) \qquad \varphi(\overline{3}) = (\overline{0},\overline{1}) \qquad \varphi(\overline{5}) = (\overline{1},\overline{0}) \qquad \varphi(\overline{7}) = (\overline{1},\overline{1})$$

as shown by the following tables:

$\cdot$	$\overline{1}$	$\overline{3}$	$\overline{5}$	$\overline{7}$
$\overline{1}$	$\overline{1}$	$\overline{3}$	$\overline{5}$	$\overline{7}$
$\overline{3}$	$\overline{3}$	$\overline{1}$	$\overline{7}$	$\overline{5}$
$\overline{5}$	$\overline{5}$	$\overline{7}$	$\overline{1}$	$\overline{3}$
$\overline{7}$	$\overline{7}$	$\overline{5}$	$\overline{3}$	$\overline{1}$

$+$	$(\overline{0},\overline{0})$	$(\overline{0},\overline{1})$	$(\overline{1},\overline{0})$	$(\overline{1},\overline{1})$
$(\overline{0},\overline{0})$	$(\overline{0},\overline{0})$	$(\overline{0},\overline{1})$	$(\overline{1},\overline{0})$	$(\overline{1},\overline{1})$
$(\overline{0},\overline{1})$	$(\overline{0},\overline{1})$	$(\overline{0},\overline{0})$	$(\overline{1},\overline{1})$	$(\overline{1},\overline{0})$
$(\overline{1},\overline{0})$	$(\overline{1},\overline{0})$	$(\overline{1},\overline{1})$	$(\overline{0},\overline{0})$	$(\overline{0},\overline{1})$
$(\overline{1},\overline{1})$	$(\overline{1},\overline{1})$	$(\overline{1},\overline{0})$	$(\overline{0},\overline{1})$	$(\overline{0},\overline{0})$

Furthermore, looking back at the introduction we see that the crazy group $G = \{3, \aleph, @\}$ is isomorphic to $\mathbb{Z}/3\mathbb{Z}$ as the following ordering of elements of shows.

$\cdot$	3	$\aleph$	@
3	@	3	$\aleph$
$\aleph$	3	$\aleph$	@
@	$\aleph$	@	3

$+$	$\overline{1}$	$\overline{0}$	$\overline{2}$
$\overline{1}$	$\overline{2}$	$\overline{1}$	$\overline{0}$
$\overline{0}$	$\overline{1}$	$\overline{0}$	$\overline{2}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{1}$

Also, the 6 element group at the very end of Section 1.2 is isomorphic to S_3:

$*$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	1	6	5	4	3
3	3	5	1	6	2	4
4	4	6	5	1	3	2
5	5	3	4	2	6	1
6	6	4	2	3	1	5

$\circ$	id	(1 2)	(1 3)	(2 3)	(1 2 3)	(1 3 2)
id	id	(1 2)	(1 3)	(2 3)	(1 2 3)	(1 3 2)
(1 2)	(1 2)	id	(1 3 2)	(1 2 3)	(2 3)	(1 3)
(1 3)	(1 3)	(1 2 3)	id	(1 3 2)	(1 2)	(2 3)
(2 3)	(2 3)	(1 3 2)	(1 2 3)	id	(1 3)	(1 2)
(1 2 3)	(1 2 3)	(1 3)	(2 3)	(1 2)	(1 3 2)	id
(1 3 2)	(1 3 2)	(2 3)	(1 2)	(1 3)	id	(1 2 3)

The property of being isomorphic has the following basic properties. Roughly, we are saying that isomorphism is an equivalence relation on the set of all groups. Formally, there are technical problems talking about "the set of all groups" (some collections are simply too big to be sets), but let's not dwell on those details here.

Proposition 8.1.3.

1. *For any group G, the function $id_G\colon G \cong G$ is an isomorphism, so $G \cong G$.*
2. *If $\varphi\colon G \to H$ is an isomorphism, then $\varphi^{-1}\colon H \to G$ is an isomorphism. In particular, if $G \cong H$, then $H \cong G$.*
3. *If $\varphi\colon G \to H$ and $\psi\colon H \to K$ are isomorphisms, then $\psi \circ \varphi\colon G \to K$ is an isomorphism. In particular, if $G \cong H$ and $H \cong K$, then $G \cong K$.*

Proof.

1. Let $(G, \cdot)$ be a group. The function $id_G\colon G \to G$ is a bijection, and for any $a, b \in G$ we have
$$id_G(a \cdot b) = a \cdot b = id_G(a) \cdot id_G(b)$$
so $id_G\colon G \to G$ is an isomorphism.
2. Let $\cdot$ be the group operation in G and let $\star$ be the group operation in H. Since $\varphi\colon G \to H$ is a bijection, we know that it has an inverse $\varphi^{-1}\colon H \to G$ which is also a bijection (because φ^{-1} has an inverse, namely φ). We need only check the second property. Let $c, d \in H$. Since φ is a bijection, it is a surjection, so we may fix $a, b \in G$ with $\varphi(a) = c$ and $\varphi(b) = d$. By definition of φ^{-1}, we then have $\varphi^{-1}(c) = a$ and $\varphi^{-1}(d) = b$. Now
$$\varphi(a \cdot b) = \varphi(a) \star \varphi(b) = c \star d$$
so by definition of φ^{-1} we also have $\varphi^{-1}(c \star d) = a \cdot b$. Therefore,
$$\varphi^{-1}(c \star d) = a \cdot b$$
Putting this information together we see that
$$\varphi^{-1}(c \star d) = a \cdot b = \varphi^{-1}(c) \cdot \varphi^{-1}(d)$$
Since $c, d \in H$ were arbitrary, the second property holds. Therefore, $\varphi^{-1}\colon H \to G$ is an isomorphism.

3. Let $\cdot$ be the group operation in G, let $\star$ be the group operation in H, and let $*$ be the group operation in K. Since the composition of bijections is a bijection, it follows that $\psi \circ \varphi \colon G \to K$ is a bijection. For any $a, b \in G$, we have

$$\begin{aligned}(\psi \circ \varphi)(a \cdot b) &= \psi(\varphi(a \cdot b)) \\ &= \psi(\varphi(a) \star \varphi(b)) && \text{(because } \varphi \text{ is an isomorphism)} \\ &= \psi(\varphi(a)) * \psi(\varphi(b)) && \text{(because } \psi \text{ is an isomorphism)} \\ &= (\psi \circ \varphi)(a) * (\psi \circ \varphi)(b)\end{aligned}$$

Therefore, $\psi \circ \phi \colon G \to K$ is an isomorphism.

□

It gets tiresome consistently using different notation for the operation in G and the operation in H, so we will stop doing it unless absolutely necessary. Thus, we will write

$$\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b)$$

where you need to keep in mind that the $\cdot$ on the left is the group operation in G and the $\cdot$ on the right is the group operation in H.

Proposition 8.1.4. *Let $\varphi \colon G \to H$ be an isomorphism. We have the following.*

1. $\varphi(e_G) = e_H$.
2. $\varphi(a^{-1}) = \varphi(a)^{-1}$ *for all* $a \in G$.
3. $\varphi(a^n) = \varphi(a)^n$ *for all* $a \in G$ *and all* $n \in \mathbb{Z}$.

Proof.

1. We have
$$\varphi(e_G) = \varphi(e_G \cdot e_G) = \varphi(e_G) \cdot \varphi(e_G)$$
hence
$$e_H \cdot \varphi(e_G) = \varphi(e_G) \cdot \varphi(e_G)$$
Using the cancellation law, it follows that $\varphi(e_G) = e_H$.

2. Let $a \in G$. We have
$$\varphi(a) \cdot \varphi(a^{-1}) = \varphi(a \cdot a^{-1}) = \varphi(e_G) = e_H$$
and
$$\varphi(a^{-1}) \cdot \varphi(a) = \varphi(a^{-1} \cdot a) = \varphi(e_G) = e_H$$
Therefore, $\varphi(a^{-1}) = \varphi(a)^{-1}$.

3. For $n = 0$, this says that $\varphi(e_G) = e_H$, which is true by part 1. The case $n = 1$ is trivial, and the case $n = -1$ is part 2. We first prove the result for all $n \in \mathbb{N}^+$ by induction. We already noticed that $n = 1$ is trivial. Suppose that $n \in \mathbb{N}^+$ is such that $\varphi(a^n) = \varphi(a)^n$ for all $a \in G$. For any $a \in G$ we have
$$\begin{aligned}\varphi(a^{n+1}) &= \varphi(a^n \cdot a) \\ &= \varphi(a^n) \cdot \varphi(a) (\text{since } \varphi \text{ is an isomorphism}) \\ &= \varphi(a)^n \cdot \varphi(a) (\text{by induction}) \\ &= \varphi(a)^{n+1}\end{aligned}$$

Thus, the result holds for $n+1$. Therefore, the result is true for all $n \in \mathbb{N}^+$ by induction. We finally handle $n \in \mathbb{Z}$ with $n < 0$. For any $a \in G$ we have

$$\begin{aligned} \varphi(a^n) &= \varphi((a^{-1})^{-n}) \\ &= \varphi(a^{-1})^{-n} && (\text{since } -n > 0) \\ &= (\varphi(a)^{-1})^{-n} && (\text{by part 2}) \\ &= \varphi(a)^n \end{aligned}$$

Thus, the result is true for all $n \in \mathbb{Z}$.

□

Theorem 8.1.5. *Let G be a cyclic group.*

1. *If $|G| = \infty$, then $G \cong \mathbb{Z}$.*
2. *If $|G| = n$, then $G \cong \mathbb{Z}/n\mathbb{Z}$.*

Proof.

1. Suppose that $|G| = \infty$. Since G is cyclic, we may fix $c \in G$ with $G = \langle c \rangle$. Since $|G| = \infty$, it follows that $|c| = \infty$. Define $\varphi\colon \mathbb{Z} \to G$ by letting $\varphi(n) = c^n$. Notice that φ is surjective because $G = \langle c \rangle$. Also, if $\varphi(m) = \varphi(n)$, then $c^m = c^n$, and hence $m = n$ by Proposition 4.3.9, hence φ is injective. Putting this together, we conclude that φ is a bijection. For any $k, \ell \in \mathbb{Z}$, we have

   $$\varphi(k + \ell) = c^{k+\ell} = c^k \cdot c^\ell = \varphi(k) \cdot \varphi(\ell)$$

   Therefore, φ is an isomorphism. It follows that $\mathbb{Z} \cong G$ and hence $G \cong \mathbb{Z}$.

2. Suppose that $|G| = n$. Since G is cyclic, we may fix $c \in G$ with $G = \langle c \rangle$. Since $|G| = n$, it follows from Proposition 4.3.9 that $|c| = n$. Define $\varphi\colon \mathbb{Z}/n\mathbb{Z} \to G$ by letting $\varphi(\overline{k}) = c^k$. Since we are defining a function on equivalence classes, we first need to check that φ is well-defined.

   Suppose that $k, \ell \in \mathbb{Z}$ with $\overline{k} = \overline{\ell}$. We then have that $k \equiv_n \ell$, so $n \mid (k - \ell)$. Fix $d \in \mathbb{Z}$ with $nd = k - \ell$. We then have $k = \ell + nd$, hence

   $$\varphi(\overline{k}) = c^k = c^{\ell+nd} = c^\ell (c^n)^d = c^\ell e^d = c^\ell = \varphi(\overline{\ell})$$

   Thus, φ is well-defined.

   We next check that φ is a bijection. Suppose that $k, \ell \in \mathbb{Z}$ with $\varphi(\overline{k}) = \varphi(\overline{\ell})$. We then have that $c^k = c^\ell$, hence $c^{k-\ell} = e$. Since $|c| = n$, we know from Proposition 4.2.5 that $n \mid (k - \ell)$. Therefore $k \equiv_n \ell$, and hence $\overline{k} = \overline{\ell}$. It follows that φ is injective. To see this φ is surjective, if $a \in G$, then $a \in \langle c \rangle$, so we may fix $k \in \mathbb{Z}$ with $a = c^k$ and note that $\varphi(\overline{k}) = c^k = a$. Hence, φ is a bijection.

   Finally notice that for any $k, \ell \in \mathbb{Z}$, we have

   $$\varphi(\overline{k} + \overline{\ell}) = \varphi(\overline{k + \ell}) = c^{k+\ell} = c^k \cdot c^\ell = \varphi(k) \cdot \varphi(\ell)$$

   Therefore, φ is an isomorphism. It follows that $\mathbb{Z}/n\mathbb{Z} \cong G$ and hence $G \cong \mathbb{Z}/n\mathbb{Z}$.

□

Corollary 8.1.6. *Let $p \in \mathbb{N}^+$ be prime. Any two groups of order p are isomorphic.*

Proof. Suppose that G and H have order p. By the previous theorem, we have $G \cong \mathbb{Z}/p\mathbb{Z}$ and $H \cong \mathbb{Z}/p\mathbb{Z}$. Using symmetry and transitivity of $\cong$, it follows that $G \cong H$. □

Properties Preserved by Isomorphisms

We have spent a lot of time establishing the existence of isomorphisms between various groups. How do we show that two groups are *not* isomorphic? The first thing to note is that if $G \cong H$, the we must have $|G| = |H|$ simply because if there is a bijection between two sets, then they have the same size. But what can we do if $|G| = |H|$? Looking at all possible bijections and ruling them out is not particularly feasible. The fundamental idea is that isomorphic groups are really just "renamings" of each other, so they should have all of the same fundamental properties. The next proposition is an example of this idea.

Proposition 8.1.7. *Suppose that G and H are groups with $G \cong H$. If G is abelian, then H is abelian.*

Proof. Since $G \cong H$, we may fix an isomorphism $\varphi\colon G \to H$. Let $h_1, h_2 \in H$. Since φ is an isomorphism, it is in particular surjective, so we may fix $g_1, g_2 \in G$ with $\varphi(g_1) = h_1$ and $\varphi(g_2) = h_2$. We then have

$$\begin{aligned} h_1 \cdot h_2 &= \varphi(g_1) \cdot \varphi(g_2) \\ &= \varphi(g_1 \cdot g_2) && \text{(since } \varphi \text{ is an isomorphism)} \\ &= \varphi(g_2 \cdot g_1) && \text{(since } G \text{ is abelian)} \\ &= \varphi(g_2) \cdot \varphi(g_1) && \text{(since } \varphi \text{ is an isomorphism)} \\ &= h_2 \cdot h_1 \end{aligned}$$

Since $h_1, h_2 \in H$ were arbitrary, it follows that H is abelian. □

As an example, even though $|S_3| = 6 = |\mathbb{Z}/6\mathbb{Z}|$ we have $S_3 \ncong \mathbb{Z}/6\mathbb{Z}$ because S_3 is nonabelian while $\mathbb{Z}/6\mathbb{Z}$ is abelian.

Proposition 8.1.8. *Suppose that G and H are groups with $G \cong H$. If G is cyclic, then H is cyclic.*

Proof. Since $G \cong H$, we may fix an isomorphism $\varphi\colon G \to H$. Since G is cyclic, we may fix $c \in G$ with $G = \langle c \rangle$. We claim that $H = \langle \varphi(c) \rangle$. Let $h \in H$. Since φ is in particular a surjection, we may fix $g \in G$ with $\varphi(g) = h$. Since $g \in G = \langle c \rangle$, we may fix $n \in \mathbb{Z}$ with $g = c^n$. We then have

$$h = \varphi(g) = \varphi(c^n) = \varphi(c)^n$$

so $h \in \langle \varphi(c) \rangle$. Since $h \in H$ was arbitrary, it follows that $H = \langle \varphi(c) \rangle$ and hence H is cyclic. □

As an example, consider the groups $\mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Each of these groups are abelian or order 4. However, $\mathbb{Z}/4\mathbb{Z}$ is cyclic while $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is not. It follows that $\mathbb{Z}/4\mathbb{Z} \ncong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

Proposition 8.1.9. *Suppose that $\varphi\colon G \to H$ is an isomorphism. For any $a \in G$, we have $|a| = |\varphi(a)|$ (in other words, the order of a in G equals the order of $\varphi(a)$ in H).*

Proof. Let $a \in G$. Suppose that $n \in \mathbb{N}^+$ and $a^n = e_G$. Using the fact that $\varphi(e_G) = e_H$ we have

$$\varphi(a)^n = \varphi(a^n) = \varphi(e_G) = e_H$$

so $\varphi(a)^n = e_H$. Conversely suppose that $n \in \mathbb{N}^+$ and $\varphi(a)^n = e_H$. We then have that

$$\varphi(a^n) = \varphi(a^n) = e_H$$

so $\varphi(a^n) = e_H = \varphi(e_G)$, and hence $a^n = e_G$ because φ is injective. Combining both of these, we have shown that

$$\{n \in \mathbb{N}^+ : a^n = e_G\} = \{n \in \mathbb{N}^+ : \varphi(a)^n = e_H\}$$

It follows that both of these sets are either empty (and so $|a| = \infty = |\varphi(a)|$) or both have the same least element (equal to the common order of $|a|$ and $|\varphi(a)|$). □

Thus, for example, we have $\mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z} \ncong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$ because the latter group has an element of order 8 while $a^4 = e$ for all $a \in \mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$.

Cayley's Theorem

Historically, the concept of a group was originally studied in the context of permuting the roots of a given polynomial. In particular, group theory in the 18th and early 19th centuries really consisted of the study of subgroups of S_n and thus had a much more "concrete" feel to it. The more general abstract definition of a group (as *any* set with *any* operation satisfying the axioms) didn't arise until later. In particular, Lagrange's Theorem was originally proved only for subgroups of symmetric groups. However, once the abstract definition of a group as we now know it came about, it was quickly proved that *every* finite group was isomorphic to a subgroup of a symmetric group. Thus, up to isomorphism, the older more "concrete" study of groups is equivalent to the more abstract study we are conducting. This result is known as Cayley's Theorem, and we now go about proving it. Before jumping into the proof, we first handle some preliminaries about the symmetric groups.

Proposition 8.1.10. *Suppose that X and Y are sets and that $f\colon X \to Y$ is a bijection. For each $\sigma \in S_X$, the function $f \circ \sigma \circ f^{-1}$ is a permutation of Y. Furthermore, the function $\varphi\colon S_X \to S_Y$ given by $\varphi(\sigma) = f \circ \sigma \circ f^{-1}$ is an isomorphism. In particular, if $|X| = |Y|$, then $S_X \cong S_Y$.*

Proof. Notice that $f \circ \sigma \circ f^{-1}$ is a function from Y to Y, and is a composition of bijections, so is itself a bijection. Thus, $f \circ \sigma \circ f^{-1}$ is a permutation of Y. Define $\varphi\colon S_X \to S_Y$ by letting $\varphi(\sigma) = f \circ \sigma \circ f^{-1}$. We check the following.

- φ is bijective. To see this, we show that φ has an inverse. Define $\psi\colon S_Y \to S_X$ by letting $\psi(\tau) = f^{-1} \circ \tau \circ f$ (notice that $f^{-1} \circ \tau \circ f$ is indeed a permutation of X by a similar argument as above). For any $\sigma \in S_X$, we have

$$\begin{aligned}(\psi \circ \varphi)(\sigma) &= \psi(\varphi(\sigma)) \\ &= \psi(f \circ \sigma \circ f^{-1}) \\ &= f^{-1} \circ (f \circ \sigma \circ f^{-1}) \circ f \\ &= (f^{-1} \circ f) \circ \sigma \circ (f^{-1} \circ f) \\ &= id_X \circ \sigma \circ id_X \\ &= \sigma\end{aligned}$$

- φ preserves the group operation. Let $\sigma_1, \sigma_2 \in S_X$. We then have

$$\begin{aligned}\varphi(\sigma_1 \circ \sigma_2) &= f \circ (\sigma_1 \circ \sigma_2) \circ f^{-1} \\ &= f \circ \sigma_1 \circ (f^{-1} \circ f) \circ \sigma_2 \circ f^{-1} \\ &= (f \circ \sigma_1 \circ f^{-1}) \circ (f \circ \sigma_2 \circ f^{-1}) \\ &= \varphi(\sigma_1) \circ \varphi(\sigma_2)\end{aligned}$$

Therefore, $\varphi\colon S_X \to S_Y$ is an isomorphism. □

Corollary 8.1.11. *If X is a finite set with $|X| = n$, then $S_X \cong S_n$.*

Proof. If X is finite with $|X| = n$, we may fix a bijection $f\colon X \to \{1, 2, 3, \ldots, n\}$ and apply the previous proposition. □

Proposition 8.1.12. *Let G and H be groups. Suppose that $\varphi\colon G \to H$ is an injective function with the property that $\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b)$ for all $a, b \in G$. Letting $K = range(\varphi)$, we then have that K is a subgroup of H and that $\varphi\colon G \to K$ is an isomorphism.*

Proof. We need to check that K is a subgroup of H. We know that $\varphi(e_G) = e_H$, so $e_H \in K$. Suppose that $c, d \in K$. Fix $a, b \in G$ with $\varphi(a) = c$ and $\varphi(b) = d$. We then have

$$\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b) = c \cdot d$$

so $c \cdot d \in K$. Finally, suppose that $c \in K$. Fix $a \in G$ with $\varphi(a) = c$. We then have

$$\varphi(a^{-1}) = \varphi(a)^{-1} = c^{-1}$$

so $c^{-1} \in K$. Therefore, K is a subgroup of G. Since $\varphi\colon G \to H$ is an injective function and we have defined $K = \text{range}(\varphi)$, if we view φ as a function $\varphi\colon G \to K$, then φ is now also surjective (for the simple reason that we have "cut down" only the set H to range(φ)), so φ is bijective. We already know that φ preserves the group operation by assumption, so φ is an isomorphism from G to K. □

Theorem 8.1.13 (Cayley's Theorem)**.** *Let G be a group. There exists a subgroup of H of S_G such that $G \cong H$. Therefore, if $|G| = n$, then G is isomorphic to a subgroup of S_n.*

Proof. For each $a \in G$, define a function $\lambda_a\colon G \to G$ by letting $\lambda_a(g) = ag$. We claim that λ_a is a permutation of G for each $a \in G$. Let $a \in G$.

- λ_a is injective: If $g_1, g_2 \in G$ with $\lambda_a(g_1) = \lambda_a(g_2)$, then $ag_1 = ag_2$, so $g_1 = g_2$ by cancellation.
- λ_a is surjective: For any $g \in G$, we have $\lambda_a(a^{-1}g) = a(a^{-1}g) = (aa^{-1})g = eg = g$, so $g \in \text{range}(\lambda_a)$.

Thus, λ_a is indeed a permutation of G for each $a \in G$.

We next claim that $\lambda_a \circ \lambda_b = \lambda_{ab}$ for all $a, b \in G$. Suppose that $a, b \in G$. For every $g \in G$, we have

$$\begin{aligned}(\lambda_a \circ \lambda_b)(g) &= \lambda_a(\lambda_b(g)) \\ &= \lambda_a(bg) \\ &= a(bg) \\ &= (ab)g \\ &= \lambda_{ab}(g)\end{aligned}$$

Since $g \in G$ was arbitrary, it follows that $\lambda_a \circ \lambda_b = \lambda_{ab}$.

Define $\varphi\colon G \to S_G$ by letting $\varphi(a) = \lambda_a$. For any $a, b \in G$, we have

$$\begin{aligned}\varphi(ab) &= \lambda_{ab} \\ &= \lambda_a \circ \lambda_b && \text{(from above)} \\ &= \varphi(a) \circ \varphi(b)\end{aligned}$$

We now check that φ is injective. Suppose that $a, b \in G$ with $\varphi(a) = \varphi(b)$. We then have that $\lambda_a = \lambda_b$, so in particular we have $\lambda_a(e) = \lambda_b(e)$. Thus

$$a = ae = \lambda_a(e) = \lambda_b(e) = be = b$$

It follows that φ is injective. Letting $H = \text{range}(\varphi)$, we know by the previous proposition that H is a subgroup of S_G and that $\varphi\colon G \to H$ is an isomorphism.

Suppose finally that $|G| = n$. We know from above that $S_G \cong S_n$, we we may fix an isomorphism $\psi\colon S_G \to S_n$. We then have that $\psi \circ \varphi\colon G \to S_n$ is injective (because the composition of injective functions is injective) and that $\psi \circ \varphi$ preserves the group operation (as in the proof that the composition of isomorphisms is an isomorphism), so G is isomorphic to the subgroup range$(\psi \circ \varphi)$ of S_n by the previous proposition. □

Internal Direct Products

We know how to take two groups and "put them together" by taking the direct product. We now want to think about how to reverse this process. Given a group G which is not obviously a direct product, how can we recognize it as being naturally isomorphic to the direct product of two of its subgroups? Before jumping in, let's look at the direct product again.

Suppose then that H and K are groups and consider their direct product $H \times K$. Define

$$H' = \{(h, e_K) : h \in H\}$$

and

$$K' = \{e_H, k) : k \in K\}$$

We claim that H' is a normal subgroup of $H \times K$. To see that it is a subgroup, simply note that $(e_H, e_K) \in H'$, that

$$(h_1, e_K) \cdot (h_2, e_K) = (h_1 h_2, e_K e_K) = (h_1 h_2, e_K)$$

and that

$$(h, e_K)^{-1} = (h^{-1}, e_K^{-1}) = (h^{-1}, e_K)$$

To see that H' is a normal subgroup of $H \times K$, notice that if $(h, e_K) \in H'$ and $(a, b) \in H \times K$, then

$$\begin{aligned}(a, b) \cdot (h, e_K) \cdot (a, b)^{-1} &= (a, b) \cdot (h, e_K) \cdot (a^{-1}, b^{-1}) \\ &= (ah, be_K) \cdot (a^{-1}, b^{-1}) \\ &= (aha^{-1}, be_K b^{-1}) \\ &= (aha^{-1}, e_K)\end{aligned}$$

which is an element of H'. Similarly, K' is a normal subgroup of $H \times K$. Notice further that $H \cong H'$ via the function $\varphi(h) = (h, e_K)$ and similarly $K \cong K'$ via the function $\psi(k) = (e_H, k)$.

We can say some more about these relationship between the subgroups H' and K' of $H \times K$. Notice that

$$H' \cap K' = \{(e_H, e_K)\} = \{e_{H \times K}\}$$

In other words, the only thing that H' and K' have in common is the identity element of $H \times K$. Finally, note that any element of $H \times K$ can be written as a product of an element of H' and an element of K': If $(h, k) \in H \times K$, then $(h, k) = (h, e_K) \cdot (e_H, k)$.

It turns out that these various facts characterize when a given group G is naturally isomorphic to the direct product of two of its subgroups.

Theorem 8.1.14. *Suppose that H and K are subgroups of G with the following properties*

1. *H and K are both normal subgroups of G.*
2. *$H \cap K = \{e\}$.*
3. *$HK = G$.*

In this case, the function $\varphi\colon H \times K \to G$ defined by $\varphi((h, k)) = h \cdot k$ is an isomorphism. Thus, we have $G \cong H \times K$.

Begin jumping into the proof, we first prove the following important lemma.

Lemma 8.1.15. *Suppose that H and K are both normal subgroups of a group G and that $H \cap K = \{e\}$. We then have that $hk = kh$ for all $h \in H$ and $k \in K$.*

Proof. Let $h \in H$ and $k \in K$. Consider the element $hkh^{-1}k^{-1}$. Since K is normal in G, we have $hkh^{-1} \in K$ and since $k^{-1} \in K$ it follows that $hkh^{-1}k^{-1} \in K$. Similarly, since H is normal in G, we have $kh^{-1}k^{-1} \in H$ and since sine $h \in H$ it follows that $hkh^{-1}k^{-1} \in H$. Therefore, $hkh^{-1}k^{-1} \in H \cap K$ and hence $hkh^{-1}k^{-1} = e$. Multiplying on the right by kh, we conclude that $hk = kh$. □

We now have all we need to carry out the proof of our theorem.

Proof of Theorem 8.1.14. Define $\varphi\colon H \times K \to G$ by letting $\varphi((h,k)) = h \cdot k$.

- φ is injective: Suppose that $\varphi((h_1,k_1)) = \varphi((h_2,k_2))$. We then have $h_1k_1 = h_2k_2$, so multiplying on the left by h_2^{-1} and on the right by k_1^{-1}, we see that $h_2^{-1}h_1 = k_2k_1^{-1}$. Now $h_2^{-1}h_1 \in H$ because H is a subgroup of G, and $k_2k_1^{-1} \in K$ because K is a subgroup of G. Therefore, this common value is an element of $H \cap K = \{e\}$. Thus, we have $h_2^{-1}h_1 = e$ and hence $h_1 = h_2$, and similarly we have $k_2k_1^{-1} = e$ and hence $k_1 = k_2$. It follows that $(h_1,k_1) = (h_2,k_2)$.
- φ is surjective: This is immediate from the assumption that $G = HK$.
- φ preserves the group operation: Suppose that $(h_1,k_1) \in H \times K$ and $(h_2,k_2) \in H \times K$, we have

$$\begin{aligned}\varphi((h_1,k_1)\cdot(h_2,k_2)) &= \varphi((h_1h_2,k_1k_2)) \\ &= h_1h_2k_1k_2 \\ &= h_1k_1h_2k_2 && \text{(by the lemma)} \\ &= \varphi((h_1,k_1))\cdot\varphi((h_2,k_2))\end{aligned}$$

□

Definition 8.1.16. *Let G a group. Let H and K be subgroups of G such that:*

1. *H and K are both normal subgroups of G.*
2. *$H \cap K = \{e\}$.*
3. *$HK = G$.*

We then say that G is the internal direct product *of H and K.*

As an example, consider the group $G = U(\mathbb{Z}/8\mathbb{Z})$. Let $H = \langle\overline{3}\rangle = \{\overline{1},\overline{3}\}$ and let $K = \langle\overline{5}\rangle = \{\overline{1},\overline{5}\}$. We then have that H and K are both normal subgroups of G (because G is abelian), that $H \cap K = \{\overline{1}\}$, and that

$$HK = \{\overline{1}\cdot\overline{1}, \overline{3}\cdot\overline{1}, \overline{1}\cdot\overline{5}, \overline{3}\cdot\overline{5}\} = \{\overline{1},\overline{3},\overline{5},\overline{7}\} = G$$

Therefore, $U(\mathbb{Z}/8\mathbb{Z})$ is the internal direct product of H and K. I want to emphasize that $U(\mathbb{Z}/8\mathbb{Z})$ does not *equal* $H \times K$. After all, as sets we have

$$U(\mathbb{Z}/8\mathbb{Z}) = \{\overline{1},\overline{3},\overline{5},\overline{7}\}$$

while

$$H \times K = \{(\overline{1},\overline{1}),(\overline{3},\overline{1}),(\overline{1},\overline{5}),(\overline{3},\overline{5})\}$$

However, the above result shows that these two groups are isomorphic via the function $\varphi\colon H \times K \to U(\mathbb{Z}/8\mathbb{Z})$ given by $\varphi((h,k)) = h \cdot k$. Since H and K are each cyclic of order 2, it follows that each of H and K are isomorphic to $\mathbb{Z}/2\mathbb{Z}$. Therefore, by Problem 1 on the Homework 18, it follows that $U(\mathbb{Z}/8\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

Corollary 8.1.17. *Let G a finite group. Let H and K be subgroups of G such that:*

1. *H and K are both normal subgroups of G.*
2. *$H \cap K = \{e\}$.*
3. *$|H| \cdot |K| = |G|$.*

We then have that G is the internal direct product of H and K.

Proof. We need only prove that $HK = G$. The in the above proof where showed that $h_1k_1 = h_2k_2$ implies $h_1 = h_2$ and $k_1 = k_2$, we only used the fact that $H \cap K = \{e\}$. Therefore, since we are assuming that $H \cap K = \{e\}$, it follows that $|HK| = |H| \cdot |K|$. Since we are assuming that $|H| \cdot |K| = |G|$, it follows that $|HK| = |G|$, and since G is finite we conclude that $HK = G$. □

Suppose that $G = D_6$. Let

$$H = \{id, r^2, r^4, s, r^2s, r^4s\}$$

and let

$$K = \{id, r^3\}$$

We have that $K = Z(G)$ from the homework, so K is a normal subgroup of G. It is straightforward to check that H is subgroup of G and

$$[G : H] = \frac{|G|}{|H|} = \frac{12}{6} = 2$$

By the homework, we conclude that H is a normal subgroup of G. Now $H \cap K = \{id\}$ and $|H| \cdot |K| = 6 \cdot 2 = 12 = |G|$ (you can also check directly that $HK = G$). It follows that G is the internal direct product of H and K, so in particular we have $G \cong H \times K$. Notice that K is cyclic of order 2, so $K \cong \mathbb{Z}/2\mathbb{Z}$. Furthermore, H is a group of order 6, and it is not hard to convince yourself that $H \cong D_3$: Roughly, you can map r^2 (where r is rotation in D_6) to r (where r is rotation in D_3). Geometrically, when working with H, you are essentially looking at the regular hexagon and focusing only on the rotations by 120^o (corresponding to r^2), 240^o (corresponding to r^4) and the identity, along with the standard flip. This really just corresponds exactly to the rigid motions of the triangle. I hope that convinces you that $H \cong D_3$, but you can check formally by looking at the corresponding Cayley tables. Now $D_3 = S_3$, so it follows that $H \cong S_3$ and hence

$$D_6 \cong H \times K \cong S_3 \times \mathbb{Z}/2\mathbb{Z}$$

where the last line uses Problem 1 on Homework 18.

8.2 Homomorphisms

Our definition of isomorphism had two requirements: that the function was a bijection and that it preserves the operation. We next investigate what happens if we drop the former and just require that the function preserves the operation.

Definition 8.2.1. *Let G and H be groups. A* homomorphism *from G to H is a function $\varphi\colon G \to H$ such that $\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b)$ for all $a, b \in G$.*

The following functions are homomorphisms:

- The determinant function $det\colon GL_n(\mathbb{R}) \to \mathbb{R}\backslash\{0\}$ where the operation on $\mathbb{R}\backslash\{0\}$ is multiplication. This is because $det(AB) = det(A) \cdot det(B)$ for any matrices A and B. Notice that det is certainly not an injective function.
- The sign function $\varepsilon\colon S_n \to \{\pm 1\}$ where the operation on $\{\pm 1\}$ is multiplication. Notice that ε is homomorphism by Proposition 5.3.7. Again, for $n \geq 3$, the function ε is very far from being injective.

- For any groups G and H, the function $\pi_1 \colon G \times H \to G$ given by $\pi_1((g,h)) = g$ and the function $\pi_2 \colon G \times H \to H$ given by $\pi_2((g,h)) = h$. For any $g_1, g_2 \in G$ and $h_1, h_2 \in H$ we have

$$\pi((g_1,h_1)\cdot(g_2,h_2)) = \pi_1((g_1g_2, h_1h_2)) = g_1g_2 = \pi_1((g_1,h_1))\cdot\pi_2((g_2,h_2))$$

  and similarly for π_2.

- For a given group G with normal subgroup N, the function $\pi \colon G \to G/N$ by letting $\pi(a) = aN$ for all $a \in G$. For any $a, b \in G$, we have

$$\pi(ab) = abH = aH \cdot bH = \pi(a)\cdot\pi(b)$$

Proposition 8.2.2. *Let $\varphi \colon G \to H$ be an homomorphism. We have the following.*

1. *$\varphi(e_G) = e_H$.*
2. *$\varphi(a^{-1}) = \varphi(a)^{-1}$ for all $a \in G$.*
3. *$\varphi(a^n) = \varphi(a)^n$ for all $a \in G$ and all $n \in \mathbb{Z}$.*

Proof. In the corresponding proof for isomorphisms, we only used the fact that φ preserves the group operation, not that it is a bijection. □

Definition 8.2.3. *Let $\varphi \colon G \to H$ be a homomorphism. We define $ker(\varphi) = \{g \in G : \varphi(g) = e_H\}$. The set $ker(\varphi)$ is called the* kernel *of φ.*

For example, for $det \colon GL_n(\mathbb{R}) \to \mathbb{R}\backslash\{0\}$, we have

$$ker(det) = \{A \in GL_n(\mathbb{R}) : det(A) = 1\} = SL_n(\mathbb{R})$$

and for $\varepsilon \colon S_n \to \{\pm 1\}$, we have

$$ker(\varepsilon) = \{\sigma \in S_n : \varepsilon(\sigma) = 1\} = A_n$$

Given a homomorphism $\varphi \colon G \to H$, the kernel of φ is always a subgroup of G and in fact it is always a normal subgroup of G.

Proposition 8.2.4. *Suppose that $\varphi \colon G \to H$ is a homomorphism. Letting $K = ker(\varphi)$, we then have that K is a normal subgroup of G.*

Proof. We first check that K is indeed a subgroup of G. Since $\varphi(e_G) = e_H$ by the proposition, we see that $e_G \in K$. Given $a, b \in K$, we have $\varphi(a) = e_H = \varphi(b)$, so

$$\varphi(ab) = \varphi(a)\cdot\varphi(b) = e_H \cdot e_H = e_H$$

and hence $ab \in K$. Given $a \in K$, we have $\varphi(a) = e_H$, so

$$\varphi(a^{-1}) = \varphi(a)^{-1} = e_H^{-1} = e_H$$

so $a^{-1} \in K$. Putting this all together, we see that K is a subgroup of G. Suppose now that $a \in K$ and $g \in G$. Since $a \in K$ we have $\varphi(a) = e_H$, so

$$\begin{aligned}\varphi(gag^{-1}) &= \varphi(g)\cdot\varphi(a)\cdot\varphi(g^{-1}) \\ &= \varphi(g)\cdot e_H \cdot \varphi(g)^{-1} \\ &= \varphi(g)\cdot\varphi(g)^{-1} \\ &= e_H\end{aligned}$$

and hence $gag^{-1} \in K$. Therefore, K is a normal subgroup of G. □

We've just seen that the kernel of a homomorphism is always a normal subgroup of G. It's a nice fact that every normal subgroup of a group arises in this way, so we get another equivalent characterization of a normal subgroup.

Theorem 8.2.5. *Let G be a group and let K be a normal subgroup of G. There exists a group H and a homomorphism $\varphi\colon G \to H$ with $K = ker(\varphi)$.*

Proof. Suppose that K is a normal subgroup of G. We can then form the quotient group $H = G/K$. Define $\varphi\colon G \to H$ by letting $\varphi(a) = aK$ for all $a \in G$. As discussed above, φ is a homomorphism. Notice that for any $a \in G$, we have

$$\begin{aligned} a \in \ker(\varphi) &\iff \varphi(a) = eK \\ &\iff aK = eK \\ &\iff eK = aK \\ &\iff e^{-1}a \in K \\ &\iff a \in K \end{aligned}$$

Therefore φ is a homomorphism with $\ker(\varphi) = K$. □

Proposition 8.2.6. *Let $\varphi\colon G \to H$ be a homomorphism. φ is injective if and only if $ker(\varphi) = \{e_G\}$.*

Proof. Suppose first that φ is injective. We know that $\varphi(e_G) = e_H$, so $e_G \in \ker(\varphi)$. Suppose now that $a \in \ker(\varphi)$. We then have $\varphi(a) = e_H = \varphi(e_G)$, so since φ is injective we conclude that $a = e_G$. Therefore, $\ker(\varphi) = \{e_G\}$.

Suppose conversely that $\ker(\varphi) = \{e_G\}$. Let $a, b \in G$ with $\varphi(a) = \varphi(b)$. We then have

$$\begin{aligned} \varphi(a^{-1}b) &= \varphi(a^{-1}) \cdot \varphi(b) \\ &= \varphi(a)^{-1} \cdot \varphi(b) \\ &= \varphi(a)^{-1} \cdot \varphi(a) \\ &= e_H \end{aligned}$$

so $a^{-1}b \in \ker(\varphi)$. Now we are assuming that $\ker(\varphi) = \{e_G\}$, so we conclude that $a^{-1}b = e_G$. Multiplying on the left by a, we see that $a = b$. Therefore, φ is injective. □

Proposition 8.2.7. *Suppose that $\varphi\colon G \to H$ is a homomorphism. If $a \in G$ has finite order, then $|\varphi(a)|$ divides $|a|$.*

Proof. Suppose that $a \in G$ has finite order and let $n = |a|$. We have $a^n = e_G$, so

$$\varphi(a)^n = \varphi(a^n) = \varphi(e_G) = e_H$$

Thus, $\varphi(a)$ has finite order, and furthermore, if we let $m = |\varphi(a)|$, then $m \mid n$ by Proposition 4.2.5. □

Proposition 8.2.8. *Let $\varphi\colon G_1 \to G_2$ be a homomorphism. We have the following*

1. *For all subgroups H_1 of G_1, we have $\varphi(H_1) = \{\varphi(a) : a \in H_1\}$ is a subgroup of G_2. In particular, range$(\varphi) = \varphi(G_1)$ is a subgroup of G_2.*
2. *For all subgroups H_2 of G_2, we have $\varphi^{-1}(H_2) = \{a \in G_1 : \varphi(a) \in H_2\}$ is a subgroup of G_1.*

Proof. Let e_1 be the identity of G_1 and let e_2 be the identity of G_2.

1. Let H_1 be a subgroup of G_1. Notice that we have $e_1 \in H_1$ because H_1 is a subgroup of G_2, so $e_2 = \varphi(e_1) \in \varphi(H_1)$. Suppose now that $c, d \in \varphi(H_1)$ and fix $a, b \in H_1$ with $\varphi(a) = c$ and $\varphi(b) = d$. Since $a, b \in H_1$ and H_1 is a subgroup of G_1, it follows that $ab \in H_1$. Now

$$\varphi(ab) = \varphi(a) \cdot \varphi(b) = cd$$

so $cd \in \varphi(H_1)$. Finally, suppose that $c \in \varphi(H_1)$ and fix $a \in H_1$ with $\varphi(a) = c$. Since $a \in H_1$ and H_1 is a subgroup of G_1, it follows that $a^{-1} \in H_1$. Now

$$\varphi(a^{-1}) = \varphi(a)^{-1} = c^{-1}$$

so $c^{-1} \in \varphi(H_1)$. Putting it all together, we conclude that $\varphi(H_1)$ is a subgroup of G_2.

2. Let H_2 be a subgroup of G_2. Notice that we have $e_2 \in H_2$ because H_2 is a subgroup of H_1. Since $\varphi(e_1) = e_2 \in H_2$, it follows that $e_1 \in \varphi^{-1}(H_2)$. Suppose now that $a, b \in \varphi^{-1}(H_2)$ so $\varphi(a) \in H_2$ and $\varphi(b) \in H_2$. We then have

$$\varphi(ab) = \varphi(a) \cdot \varphi(b) \in H_2$$

because H_2 is a subgroup of G_2, so $ab \in \varphi^{-1}(H_2)$. Finally, suppose that $a \in \varphi^{-1}(H_2)$ so that $\varphi(a) \in H_2$. We then have

$$\varphi(a^{-1}) = \varphi(a)^{-1} \in H_2$$

because H_2 is a subgroup of G_2, so $a^{-1} \in \varphi^{-1}(H_2)$ Putting it all together, we conclude that $\varphi^{-1}(H_2)$ is a subgroup of G_1.

□

8.3 The Isomorphism and Correspondence Theorems

Suppose that $\varphi\colon G \to H$ is a homomorphism. There are two ways in which φ can fail to be an isomorphism; namely φ can fail to be injective and φ can fail to be surjective. The latter is not really a problem at all, because $\varphi(G)$ is subgroup of H and if we restrict H down to this subgroup, then φ becomes surjective. The real issue is how to deal with a lack of injectivity. Let $K = \ker(\varphi)$. As we've seen, if $K = \{e_G\}$, then φ is injective. But what if K is a nontrivial subgroup?

The general idea is as follows. All elements of K get sent to the same element of H via φ, namely to e_H. Now viewing K as a subgroup of G, the cosets of K break up G into pieces. The key insight is that two elements which belong to the same coset of K must get sent via φ to the same element of H, and conversely if two elements below to different cosets of K then they must be sent via φ to distinct values of H. This is the content of the following lemma.

Lemma 8.3.1. *Suppose that $\varphi\colon G \to H$ is a homomorphism and let $K = ker(\varphi)$. For any $a, b \in G$, we have that $a \sim_K b$ if and only if $\varphi(a) = \varphi(b)$.*

Proof. Suppose first that $a, b \in G$ satisfy $a \sim_K b$. Fix $k \in K$ with $ak = b$ and notice that

$$\begin{aligned}\varphi(b) &= \varphi(ak)\\ &= \varphi(a) \cdot \varphi(k)\\ &= \varphi(a) \cdot e_H\\ &= \varphi(a)\end{aligned}$$

Suppose conversely that $a, b \in G$ satisfy $\varphi(a) = \varphi(b)$. We then have

$$\begin{aligned}\varphi(a^{-1}b) &= \varphi(a^{-1}) \cdot \varphi(b) \\ &= \varphi(a)^{-1} \cdot \varphi(b) \\ &= \varphi(a)^{-1} \cdot \varphi(a) \\ &= e_H\end{aligned}$$

so $a^{-1}b \in K$. It follows that $a \sim_K b$. □

In less formal terms, the lemma says that φ is constant on each coset of K, and assigns distinct values to distinct cosets. This sets up a well-defned injective function from the quotient group G/K to the group H. Restricting down to range(φ) on the right, this function is a bijection. Furthermore, this function is an isomorphism as we now prove.

Theorem 8.3.2 (First Isomorphism Theorem)**.** *Let $\varphi\colon G \to H$ be a homomorphism and let $K = ker(\varphi)$. Define a function $\psi\colon G/K \to H$ by letting $\psi(aK) = \varphi(a)$. We then have that ψ is a well-defined function which is an isomorphism onto the subgroup range(φ) of H. Therefore*

$$G/ker(\varphi) \cong range(\varphi)$$

Proof. We check the following.

- ψ is well-defined: Suppose that $a, b \in G$ with $aK = bK$. We then have $a \sim_K b$, so by the lemma we have $\varphi(a) = \varphi(b)$, and hence $\psi(aK) = \psi(bK)$.
- ψ is injective: Suppose that $a, b \in G$ with $\psi(aK) = \psi(bK)$. We then have that $\varphi(a) = \varphi(b)$, so $a \sim_K b$ by the lemma (in the other direction), and hence $aK = bK$.
- ψ is surjective onto range(φ): Since $\varphi\colon G \to H$ is a homomorphism, we know from Proposition 8.2.8 that range(φ) is a subgroup of H. Now since $\psi(aK) = \varphi(a)$ for all $a \in G$, we have range(ψ) = range(φ).
- ψ preserves the group operation: For any $a, b \in G$, we have

$$\psi(aK \cdot bK) = \psi(abK) = \varphi(ab) = \varphi(a) \cdot \varphi(b) = \psi(aK) \cdot \psi(bK)$$

Putting it all together, the function $\psi\colon G/K \to H$ defined by $\psi(aK) = \varphi(a)$ is well-defined injective homomorphism, and we □

For example, consider the homomorphism $det\colon GL_n(\mathbb{R}) \to \mathbb{R}\backslash\{0\}$ where we view $\mathbb{R}\backslash\{0\}$ as a group under multiplication. As discussed in the previous section, we have $\ker(det) = SL_n(\mathbb{R})$. Notice that *det* is a surjective function because every nonzero real number arises as the determinant of some invertible matrix (for example, the identity matrix with the $(1, 1)$ entry replaced by r has determinant r). The First Isomorphism Theorem tells us that

$$GL_n(\mathbb{R})/SL_n(\mathbb{R}) \cong \mathbb{R}\backslash\{0\}$$

Here is what is happening intuitively. The subgroup $SL_n(\mathbb{R})$, which is the kernel of the determinant homomorphism, is the set of $n \times n$ matrices with determinant 1. This break up the $n \times n$ matrices into cosets which correspond exactly to the nonzero real numbers in the sense that all matrices of a given determinant form a coset. The First Isomorphism Theorem says that multiplication in the quotient (where you take representatives matrices from the corresponding cosets, multiply the matrices, and then expand to the resulting coset) corresponds exactly to just multiplying the real numbers which "label" the cosets.

For another example, consider the sign homomorphism $\varepsilon\colon S_n \to \{\pm 1\}$ where $n \geq 2$. As discussed, ε is a homomorphism with $\ker(\varepsilon) = A_n$. Notice that ε is surjective because $n \geq 2$ (so $\varepsilon(id) = 1$ and $\varepsilon((1\ 2)) = -1$). By the First Isomorphism Theorem we have

$$S_n/A_n \cong \{\pm 1\}$$

Now the quotient group S_n/A_n consists of two cosets: the even permutations form one coset (namely A_n) and the odd permutations form the other coset. The isomorphism above is simply saying that we can label all of the even permutations with 1 and all of the odd permutations with -1, and in this way multiplication in the quotient corresponds exactly to multiplication of the labels.

Theorem 8.3.3 (Second Isomorphism Theorem)**.** *Let G be a group, let H be a subgroup of G, and let N be a normal subgroup of G. We then have that HN is a subgroup of G, that N is a normal subgroup of HN, that $H \cap N$ is a normal subgroup of N, and that*

$$\frac{HN}{N} \cong \frac{H}{H \cap N}$$

Proof. First notice that since N is a normal subgroup of G, we know from the homework that HN is a subgroup of G. Now since $e \in H$, we have $N \subseteq HN$, and since N is a normal subgroup of G it follows that N is a normal subgroup of HN (conjugation by elements of HN is just a special case of conjugation by elements of G). Therefore, we may form the quotient HN/N. Notice that we also have $H \subseteq HN$ (because $e \in N$), so we may define $\varphi\colon H \to HN/N$ by letting $\varphi(h) = hN$. We check the following:

- φ is a homomorphism: For any $h_1, h_2 \in H$, we have

$$\varphi(h_1h_2) = h_1h_2N = h_1N \cdot h_2N = \varphi(h_1) \cdot \varphi(h_2)$$

- φ is surjective: Let $g \in HN$. We need to show that $gN \in \text{range}(\varphi)$. Fix $h \in H$ and $n \in N$ with $g = hn$. Notice that we then have $h \sim_N g$, so $hN = gN$. It follows that $\varphi(h) = hN = gN$, so $gN \in \text{range}(\varphi)$.

- $\ker(\varphi) = H \cap N$: Suppose first that $a \in H \cap N$. We then have that $a \in N$, so $aN \cap eN \neq \emptyset$ and hence $aN = eN$. It follows that $\varphi(a) = aN = eN$, so $a \in \ker(\varphi)$. It follows that $H \cap N \subseteq \ker(\varphi)$.

  Suppose conversely that $a \in \ker(\varphi)$. Notice that the domain of φ is H, so $a \in H$. Since $a \in \ker(\varphi)$, we have $\varphi(a) = eN$, so $aN = eN$. It follows that $a = e^{-1}a \in N$. Therefore, $a \in H$ and $a \in N$, so $a \in H \cap N$. It follows that $\ker(\varphi) \subseteq H \cap N$.

Therefore, $\varphi\colon H \to HN/N$ is a surjective homomorphism with kernel equal to $H \cap N$. Since $H \cap N$ is the kernel of a homomorphism with domain H, we know that $H \cap N$ is a normal subgroup of H. By the First Isomorphism Theorem, we conclude that

$$\frac{HN}{N} \cong \frac{H}{H \cap N}$$

□

Suppose that you are given a normal subgroup N of a group G. We can then form the quotient group G/N and consider the projection homomorphism $\pi\colon G \to G/N$ given by $\pi(a) = aN$. Now if we have a subgroup of the quotient G/N, then Proposition 8.2.8 tells us that we can pull this subgroup back via π to obtain a subgroup of G. Call this resulting subgroup H. Now the elements of G/N are cosets, so when we pull back we see that H will be a union of cosets of N. In particular, since eN will be an element of G/N, we will have that $N \subseteq H$.

Conversely, suppose that H is a subgroup of G with the property that $N \subseteq H$. Suppose that $a \in H$ and $b \in G$ with $a \sim_N b$. We may then fix $n \in N$ with $an = b$. Since $N \subseteq H$, we have that $n \in H$, so $b = an \in H$.

Thus, if H contains an element of a given coset of N, then H must contain every element of that coset. In other words, H is a union of cosets of N. Using Proposition 8.2.8 again, we can go across π and notice that the cosets which are subsets of H form a subgroup of the quotient.

Taken together, this is summarized in the following important result. The key ideas are discussed above, but I will omit a careful proof.

Theorem 8.3.4 (Correspondence Theorem)**.** *Let G be a group and let N be a normal subgroup of G. For every subgroup H of G with $N \subseteq H$, we have that H/N is a subgroup of G/N and the function*

$$H \mapsto H/N$$

is a bijection from subgroups of G containing N to subgroups of G/N. Furthermore, we have the following properties for any subgroups A and B of G which both contain N:

1. *A is a subgroup of B if and only if A/N is a subgroup of B/N.*
2. *A is a normal subgroup of B if and only if A/N is a normal subgroup of B/N.*
3. *If A is a subgroup of B, then $[B : A] = [B/N : A/N]$.*

Theorem 8.3.5 (Third Isomorphism Theorem)**.** *Let G be a group. Let N and K be normal subgroups of G with $N \subseteq K$. We then have that K/N is a normal subgroup of G/N and that*

$$\frac{G/N}{K/N} \cong \frac{G}{K}$$

Proof. Define $\varphi\colon G/N \to G/K$ by letting $\varphi(aN) = aK$.

- φ is well-defined: If $a, b \in G$ with $aN = bN$, then $a \sim_N b$, so $a \sim_K b$ because $N \subseteq K$, and hence $aK = bK$.
- φ is a homomorphism: For any $a, b \in G$, we have

$$\varphi(aN \cdot bN) = \varphi(abN) = abK = aK \cdot bK = \varphi(aN) \cdot \varphi(bN)$$

- φ is surjective: For any $a \in G$, we have $\varphi(aN) = aK$, so $aK \in \text{range}(\varphi)$.
- $\ker(\varphi) = N/K$: For any $a \in G$, we have

$$\begin{aligned} aN \in \ker(\varphi) &\iff \varphi(aN) = eK \\ &\iff aK = eK \\ &\iff e^{-1}a \in K \\ &\iff a \in K \end{aligned}$$

  Therefore, $\ker(\varphi) = K/N$.

By the First Isomorphism Theorem, we conclude that

$$\frac{G/N}{K/N} \cong \frac{G}{K}$$

□

Chapter 9

Group Actions

Our definition of a group involved axioms for an abstract algebraic object. However, many groups arise in a setting where the elements of the group naturally "move around" the elements of some set. For example, the elements of $GL_n(\mathbb{R})$ represent linear transformations on $\mathbb{R}^n$ and so "move around" points in n-dimensional space. When we discussed D_n, we thought of the elements of D_n as moving the vertices of a regular n-gon. The general notion of a group working on a set in this way is called a group action. Typically these sets are very geometric or combinatorial in nature, and we can understand the sets themselves by understanding the groups. Perhaps more surprising, we can turn this idea around to obtain a great deal of information about groups by understanding the sets on which they act.

The concept of understanding an algebraic object by seeing it "act" on other objects, often from different areas of mathematics (geometry, topology, analysis, combinatorics, etc.), is tremendously important part of modern mathematics. In practice, groups typically arise as symmetries of some object just like our introduction of D_n as the "symmetries" of the regular n-gon. Instead of n-gons, the objects can be graphs, manifolds, topological spaces, vector spaces, etc.

9.1 Actions, Orbits, and Stabilizers

Definition 9.1.1. *Let G be a group and let X be a (nonempty) set. A* group action *of G on X is a function $f\colon G \times X \to X$, where we write $f(g,x)$ as $g * x$, such that*

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

We describe this situation by saying that "G acts on X".

Here are several examples of group actions.

- $GL_n(\mathbb{R})$ acts on $\mathbb{R}^n$ where $A * \mathbf{x} = A\mathbf{x}$. Notice that $I_n\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$ and $A(B\mathbf{x}) = (AB)\mathbf{x})$ for all $A, B \in GL_n(\mathbb{R})$ and all $\mathbf{x} \in \mathbb{R}^n$ from linear algebra.
- S_n on $\{1, 2, \ldots, n\}$ where $\sigma * i = \sigma(i)$. Notice that $id * i = i$ for all $i \in \{1, 2, \ldots, n\}$ and for any $\sigma, \tau \in S_n$ and $i \in \{1, 2, \ldots, n\}$, we have

$$\sigma * (\tau * i) = \sigma * \tau(i) = \sigma(\tau(i)) = (\sigma \circ \tau)(i) = (\sigma \circ \tau) * i$$

- Let G be a group. G acts on G by left multiplication, i.e. $g * a = g \cdot a$. Notice that $e * a = e \cdot a = a$ for all $a \in G$. The second axiom of a group action is just associativity of the group operation because for all $g, h, a \in G$ we have

$$g * (h * a) = g * (h \cdot a) = g \cdot (h \cdot a) = (g \cdot h) \cdot a = (g \cdot h) * a$$

- Let G be a group. G acts on G by conjugation, i.e. $g * a = gag^{-1}$. Notice that $e * a = eae^{-1} = a$ for all $a \in G$. Also, for all $g, h, a \in G$ we have

$$g * (h * a) = g * (hah^{-1}) = ghah^{-1}g^{-1} = gha(gh)^{-1} = (g \cdot h) * a$$

- If G acts on X and H is a subgroup of G, then H acts on X by simply restricting the function. For example, since D_n is a subgroup of S_n, we see that D_n acts on $\{1, 2, \ldots, n\}$ via $\sigma * i = \sigma(i)$. Also, since $SL_n(\mathbb{R})$ and $O(n, \mathbb{R})$ are subgroups of $GL_n(\mathbb{R})$, they both act on $\mathbb{R}^n$ via matrix multiplication as well.

With several examples in hand, we start with the following proposition which says that given a group action, a given element of G actually permutes the elements of X.

Proposition 9.1.2. *Suppose that G acts on X. For each fixed $a \in G$, the function $\pi_a \colon X \to X$ defined by $\pi_a(x) = a * x$ is a permutation of X.*

Proof. Suppose that $\pi_a(x) = \pi_a(y)$ so that $a * x = a * y$. We then have

$$\begin{aligned} x &= e * x \\ &= (a^{-1} \cdot a) * x \\ &= a^{-1} * (a * x) \\ &= a^{-1} * (a * y) \\ &= (a^{-1} \cdot a) * y \\ &= e * y \\ &= y \end{aligned}$$

Thus, the function π_a is injective. Now fix $x \in X$. Notice that

$$\pi_a(a^{-1} * x) = a * (a^{-1} * x) = (a \cdot a^{-1}) * x = e * x = x$$

hence $x \in \text{range}(\pi_a)$. It follows that π_a is surjective. Putting this together with the above, we see that $\pi_a \colon X \to X$ is a bijection, so π_a is a permutation of X. □

Proposition 9.1.3. *Suppose that G acts on X. Define a relation $\sim$ on X by letting $x \sim y$ if there exists $a \in G$ with $a * x = y$. The relation $\sim$ is an equivalence relation on X.*

Proof. We check the properties

- Reflexive: For any $x \in X$, we have $e * x = x$, so $x \sim x$.
- Symmetric: Suppose that $x, y \in X$ with $x \sim y$. Fix $a \in G$ with $a * x = y$. We then have

$$\begin{aligned} a^{-1} * y &= a^{-1} * (a * x) \\ &= (a^{-1} \cdot a) * x \\ &= e * x \\ &= x \end{aligned}$$

so $y \sim x$.

- Transitive: Suppose that $x, y, z \in X$ with $x \sim y$ and $y \sim z$. Fix $a, b \in G$ with $a * x = y$ and $b * y = z$. We then have
$$(b \cdot a) * x = b * (a * x) = b * y = z$$
so $x \sim z$.

□

Definition 9.1.4. *Suppose that G acts on X. The equivalence class of x under the above relation $\sim$ is called the* orbit *of x. We denote this equivalence class by $\mathcal{O}_x$. Notice that $\mathcal{O}_x = \{a * x : a \in G\}$.*

If G acts on X, we know from our general theory of equivalence relations that the orbits partition X. For example, consider the case where $G = GL_2(\mathbb{R})$ and $X = \mathbb{R}^2$ with action $A * \mathbf{x} = A\mathbf{x}$. Notice that $\mathcal{O}_{(0,0)} = \{(0,0)\}$. Let's consider $\mathcal{O}_{(1,0)}$. We claim that $\mathcal{O}_{(1,0)} = \mathbb{R}^2 \backslash \{(0,0)\}$. Since the orbits partition $\mathbb{R}^2$, we know that $(0,0) \notin \mathcal{O}_{(1,0)}$. Now if $a, b \in \mathbb{R}$ with $a \neq 0$, then

$$\begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix}$$

is in $GL_2(\mathbb{R})$ since it has determinant $a \neq 0$, and

$$\begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

so $(a, b) \in \mathcal{O}_{(1,0)}$. If $b \in \mathbb{R}$ with $b \neq 0$, then

$$\begin{pmatrix} 0 & 1 \\ b & 0 \end{pmatrix}$$

is in $GL_2(\mathbb{R})$ since it has determinant $-b \neq 0$, and

$$\begin{pmatrix} 0 & 1 \\ b & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}$$

so $(0, b) \in \mathcal{O}_{(1,0)}$. Putting everything together, it follows that $\mathcal{O}_{(1,0)} = \mathbb{R}^2 \backslash \{(0,0)\}$. Since the orbits are equivalence classes, we conclude that $\mathcal{O}_{(a,b)} = \mathbb{R}^2 \backslash \{(0,0\}$ whenever $(a, b) \neq (0, 0)$. Thus, orbits partition $\mathbb{R}^2$ into the two pieces: the origin and the rest.

In contrast, consider what happens if we let the subgroup $O(2, \mathbb{R})$ acts on $\mathbb{R}^2$. Recall that the elements of $O(2, \mathbb{R})$ are the following matrices:

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

Elements of $O(2, \mathbb{R})$ preserve distance. That is, if $A \in O(2, \mathbb{R})$ and $\mathbf{x} \in \mathbb{R}^2$, then $||A\mathbf{x}|| = ||\mathbf{x}||$ (you should have seen this in linear algebra, but it can be checked directly using the above the matrices: Suppose that $(x, y) \in \mathbb{R}^2$ with $x^2 + y^2 = r^2$ and show that the same is true after you hit (x, y) with any of the above matrices). It follows that every element of $\mathcal{O}_{(1,0)}$ is on the circle of radius 1 centered at the origin. Furthermore, $\mathcal{O}_{(1,0)}$ contains all of these points because

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

which gives all points on the unit circle as you vary θ. In general, if you work through the details, you see that the orbits of this action are the circles centered at the origin.

Definition 9.1.5. *Suppose that G acts on X. For each $x \in X$, define $G_x = \{a \in G : a * x = x\}$. The set G_x is called the* stabilizer *of x.*

Proposition 9.1.6. *Suppose that G acts on X. For each $x \in X$, the set G_x is a subgroup of G.*

Proof. Let $x \in X$. Since $e * x = x$, we have $e \in G_x$. Suppose that $a, b \in G_x$. We then have that $a * x = x$ and $b * x = x$, so

$$(a \cdot b) * x = a * (b * x) = a * x = x$$

and hence $a \cdot b \in G_x$. Suppose that $a \in G_x$ so that $a * x = x$. We then have that $a^{-1} * (a * x) = a^{-1} * x$. Now

$$a^{-1} * (a * x) = (a^{-1} * a) * x = e * x = x$$

so $a^{-1} * x = x$ and hence $a^{-1} \in G_x$. Therefore, G_x is a subgroup of G. □

Since D_4 is a subgroup of S_4, and S_4 acts on $\{1, 2, 3, 4\}$ via $\sigma * i = \sigma(i)$. It follows that D_4 acts on $\{1, 2, 3, 4\}$ as well. To see how this works, we should remember our formal definitions of r and s as elements of S_n. In D_4, we have $r = (1\ 2\ 3\ 4)$ and $s = (2\ 4)$. Working out all of the elements as permutations, we see that

$$\begin{array}{cccc} id & r = (1\ 2\ 3\ 4) & r^2 = (1\ 3)(2\ 4) & r^3 = (1\ 4\ 3\ 2) \\ s = (2\ 4) & rs = (1\ 2)(3\ 4) & r^2 s = (1\ 3) & r^3 s = (1\ 4)(2\ 3) \end{array}$$

Notice that

$$G_1 = G_3 = \{id, s\} \qquad G_2 = G_4 = \{id, r^2 s\}$$

and that

$$\mathcal{O}_1 = \mathcal{O}_2 = \mathcal{O}_3 = \mathcal{O}_4 = \{1, 2, 3, 4\}$$

Theorem 9.1.7 (Orbit-Stabilizer Theorem)**.** *Suppose that G acts on X. Let $x \in X$. There is a bijection between $\mathcal{O}_x$ and the set of (left) cosets of G_x in G, so $|\mathcal{O}_x| = [G : G_x]$. In particular, if G is finite, then*

$$|G| = |\mathcal{O}_x| \cdot |G_x|$$

so $|\mathcal{O}_x|$ divides $|G|$.

Proof. We first prove the following fundamental fact. For any $a, b \in G$, we have

$$a * x = b * x \iff a \sim_{G_x} b$$

Suppose first that $a, b \in G$ with $a * x = b * x$. We then have that

$$\begin{aligned} (a^{-1} \cdot b) * x &= a^{-1} * (b * x) \\ &= a^{-1} * (a * x) \\ &= (a^{-1} \cdot a) * x \\ &= e * x \\ &= x \end{aligned}$$

Therefore, $a^{-1}b \in G_x$ and hence $a \sim_{G_x} b$.

Suppose conversely that $a, b \in G$ with $a \sim_{G_x} b$. Fix $h \in G_x$ with $ah = b$. We then have that

$$\begin{aligned} b * x &= (a \cdot h) * x \\ &= a * (h * x) \\ &= a * x \end{aligned}$$

so $a * x = b * x$.

Now for any $a, b \in G$, we have $a \sim_{G_x} b$ if and only if $aG_x = bG_x$, so it follows that

$$a * x = b * x \iff aG_x = bG_x$$

Let $\mathcal{L}_{G_x}$ be the set of left cosets of G_x in G. Define $f\colon \mathcal{L}_{G_x} \to \mathcal{O}_x$ by letting $f(aG_x) = a * x$. Now we are defining a function on cosets, but the right-to-left direction above tells us that f is well-defined. Also, the left-to-right direction above tells us that f is injective. Now if $y \in \mathbb{O}_x$, then we may fix $a \in G$ with $y = a * x$, and notice that $f(aG_x) = a * x = y$, so f is surjective. Therefore, $f\colon \mathcal{L}_{G_x} \to \mathcal{O}_x$ is a bijection and hence $|\mathcal{O}_x| = [G : G_x]$.

Suppose now that G is finite. By Lagrange's Theorem, we know that

$$[G : G_x] = \frac{|G|}{|G_x|}$$

so

$$|\mathcal{O}_x| = \frac{|G|}{|G_x|}$$

and hence

$$|G| = |\mathcal{O}_x| \cdot |G_x|$$

□

For example, consider the standard action of S_n on $\{1, 2, \ldots, n\}$. For every $i \in \{1, 2, \ldots, n\}$, we have $\mathcal{O}_i = \{1, 2, \ldots, n\}$ (because for each j, there is a permutation sending i to j), so as $|S_n| = n!$, it follows that $|G_i| = \frac{n!}{n} = (n-1)!$. In other words, there are $(n-1)!$ permutations of $\{1, 2, \ldots, n\}$ which fix a given element of $\{1, 2, \ldots, n\}$. Of course, this could have been proven directly, but it is an immediate consequence of the Orbit-Stabilizer Theorem.

In the case of D_4 acting on $\{1, 2, 3, 4\}$ discussed above, we saw that $\mathcal{O}_i = \{1, 2, 3, 4\}$ for all i. Thus, each stabilizer G_i satisfies $|G_i| = \frac{|D_4|}{4} = \frac{8}{4} = 2$, as was verified above.

9.2 The Conjugation Action and the Class Equation

Conjugacy Classes and Centralizers

Let G be a group. Throughout this section, we will consider the special case of the action of G on G given by conjugation. That is, we are considering the action $g * a = gag^{-1}$. In this case, the orbits and the stabilizers of the action are given special names.

Definition 9.2.1. *Let G be a group and consider the action of G on G given by the conjugation.*

- *The orbits of this action are called* conjugacy classes. *The conjugacy class of a is the set*

$$\{gag^{-1} : g \in G\}$$

- *For $a \in G$, the stabilizer G_a is called the* centralizer *of a in G and is denoted $C_G(a)$. Notice that*

$$g \in C_G(a) \iff g * a = a \iff gag^{-1} = a \iff ga = ag$$

Thus,

$$C_G(a) = \{g \in G : ga = ag\}$$

is the set of elements of G which commute with a.

By our general theory of group actions, we know that $C_G(a)$ is a subgroup of G for every $a \in G$. Now the conjugacy classes are orbits, so they are subsets of G which partition G, but in general they are certainly *not* subgroups of G. However, we do know that if G is finite, then the size of every conjugacy class divides $|G|$ by the Orbit-Stabilizer Theorem because the size of a conjugacy class is the index of the corresponding centralizer subgroup. In fact, in this case, the Orbit-Stabilizer Theorem says that if G is finite, then

$$|\mathcal{O}_a| \cdot |C_G(a)| = |G|$$

Notice that if $a \in G$, then we have $a \in C_G(a)$ (because a trivially commutes with a), so since $C_G(a)$ is a subgroup of G containing a, it follows that $\langle a \rangle \subseteq C_G(a)$. It is often possible to use this simple fact together with the above equality to help calculate conjugacy classes

As an example, consider the group $G = S_3$. We work out the conjugacy class and centralizer of the various elements. Notice first that $C_G(id) = G$ because every elements commutes with the identity, and the conjugacy class of id is $\{id\}$ because $\sigma \circ id \circ \sigma^{-1} = id$ for all $\sigma \in G$. Now consider the element (1 2). On the one hand, we know that $\langle (1\ 2) \rangle = \{id, (1\ 2)\}$ is a subset of $C_G((1\ 2))$, so $|C_G((1\ 2))| \geq 2$. Since $|G| = 6$, we conclude that $|\mathcal{O}_{(1\ 2)}| \leq 3$. Now we know that $(1\ 2) \in \mathcal{O}_{(1\ 2)}$ because $\mathcal{O}_{(1\ 2)}$ is the equivalence class of (1 2) and since

- $(2\ 3)(1\ 2)(2\ 3)^{-1} = (2\ 3)(1\ 2)(2\ 3) = (1\ 3)$
- $(1\ 3)(1\ 2)(1\ 3)^{-1} = (1\ 3)(1\ 2)(1\ 3) = (2\ 3)$

it follows that (1 3) and (2 3) are also in $\mathcal{O}_{(1\ 2)}$. We now have three elements of $\mathcal{O}_{(1\ 2)}$, and since $|\mathcal{O}_{(1\ 2)}| \leq 3$, we conclude that

$$\mathcal{O}_{(1\ 2)} = \{(1\ 2), (1\ 3), (2\ 3)\}$$

Notice that we can now conclude that $|C_G((1\ 2))| = 2$, so in fact we must have $C_G((1\ 2)) = \{id, (1\ 2)\}$ without doing any other calculations.

We have now found two conjugacy classes which take up 4 of the elements of $G = S_3$. Let's look at the conjugacy class of (1 2 3). We know it contains (1 2 3), and since

$$(2\ 3)(1\ 2\ 3)(2\ 3)^{-1} = (2\ 3)(1\ 2\ 3)(2\ 3) = (1\ 3\ 2)$$

it follows that (1 3 2) is there as well. Since the conjugacy classes partition G, these are the only possible elements so we conclude that

$$\mathcal{O}_{(1\ 2\ 3)} = \{(1\ 2\ 3), (1\ 3\ 2)\}$$

Using the Orbit-Stabilizer Theorem it follows that $|C_G((1\ 2\ 3))| = 3$, so since $\langle (1\ 2\ 3) \rangle \subseteq C_G((1\ 2\ 3))$ and $|\langle (1\ 2\ 3) \rangle| = 3$, we conclude that $C_G((1\ 2\ 3)) = \langle (1\ 2\ 3) \rangle$. Putting it all together, we see that S_3 breaks up into three conjugacy classes:

$$\{id\} \qquad \{(1\ 2), (1\ 3), (2\ 3)\} \qquad \{(1\ 2\ 3), (1\ 3\ 2)\}$$

The fact that the 2-cycles form one conjugacy class and the 3-cycles form another is a specific case of a general fact which we now prove.

Lemma 9.2.2. *Let $\sigma \in S_n$ be a k-cycle, say $\sigma = (a_1\ a_2\ \ldots\ a_k)$. For any $\tau \in S_n$, the permutation $\tau\sigma\tau^{-1}$ is a k-cycle and in fact*

$$\tau\sigma\tau^{-1} = (\tau(a_1)\ \tau(a_2)\ \ldots\ \tau(a_k))$$

(Note: this k-cycle may not have the smallest element first, so you may have to "rotate" it to have it in standard cycle notation).

Proof. For any i with $1 \leq i \leq k-1$, we have $\sigma(a_i) = a_{i+1}$, hence

$$(\tau\sigma\tau^{-1})(\tau(a_i)) = \tau(\sigma(\tau^{-1}(\tau(a_i)))) = \tau(\sigma(a_i)) = \tau(a_{i+1})$$

Furthermore, since $\sigma(a_k) = a_1$, we have

$$(\tau\sigma\tau^{-1})(\tau(a_k)) = \tau(\sigma(\tau^{-1}(\tau(a_k)))) = \tau(\sigma(a_k)) = \tau(a_1)$$

To finish the proof, we need to show that $\tau\sigma\tau^{-1}$ fixes all elements distinct from the $\tau(a_i)$. Suppose then that $b \neq \tau(a_i)$ for each i. We then have that $\tau^{-1}(b) \neq a_i$ for all i. Since σ fixes all elements other the a_i, it follows that σ fixes $\tau^{-1}(b)$. Therefore

$$(\tau\sigma\tau^{-1})(b) = \tau(\sigma(\tau^{-1}(b))) = \tau(\tau^{-1}(b)) = b$$

Putting it all together, we conclude that $\tau\sigma\tau^{-1} = (\tau(a_1)\ \tau(a_2)\ \ldots\ \tau(a_k))$. □

For example, suppose that $\sigma = (1\ 6\ 3\ 4)$ and $\tau = (1\ 7)(2\ 4\ 9\ 6)(5\ 8)$. To determine $\tau\sigma\tau^{-1}$, we need only apply τ to each of the elements in the cycle σ. Thus,

$$\tau\sigma\tau^{-1} = (7\ 2\ 3\ 9) = (2\ 3\ 9\ 7)$$

This result extends beyond k-cycles, and in fact you get the reverse direction as well.

Theorem 9.2.3. *Two elements of S_n are conjugate in S_n if and only if they have the same cycle structure, i.e. if and only if their cycle notations have the same number of k-cycles for each $k \in \mathbb{N}^+$.*

Proof. Let $\sigma \in S_n$ and suppose that we write σ in cycle notation as $\sigma = \pi_1\pi_2\cdots\pi_n$ with the π_i disjoint cycles. For any $\tau \in S_n$, we have

$$\tau\sigma\tau^{-1} = \tau\pi_1\pi_2\cdots\pi_n\tau^{-1} = (\tau\pi_1\tau^{-1})(\tau\pi_2\tau^{-1})\cdots(\tau\pi_n\tau^{-1})$$

By the lemma, each $\tau\pi_i\tau^{-1}$ is a cycle of the same length as π_i. Furthermore, the various cycles $\tau\pi_i\tau^{-1}$ are disjoint from each other because they are obtained by applying τ to the elements of the cycles and τ is a bijection. Therefore, $\tau\sigma\tau^{-1}$ has the same cycle structure as σ.

Suppose conversely that σ and ρ have the same cycle structure. Match up the cycles of σ with the cycles of ρ in a manner which preserves cycle length (including 1-cycles). Define $\tau\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ as follows. Given $i \in \{1, 2, \ldots, n\}$, let $\tau(i)$ be the element of the corresponding cycle in the corresponding position. Notice that τ is a bijection because the cycles in a given cycle notation are disjoing. By inserting $\tau\tau^{-1}$ between the various cycles of σ, we can use the lemma to conclude that $\tau\sigma\tau^{-1} = \rho$. □

As an illustration of the latter part of the theorem, suppose that we are working in S_8 and we have

$$\sigma = (1\ 8)(2\ 5)(4\ 7\ 6) \qquad \rho = (2\ 7\ 3)(4\ 5)(6\ 8)$$

Notice that σ and τ have the same cycle structure, so they are conjugate in S_8. We can then write

$$\sigma = (3)(1\ 8)(2\ 5)(4\ 7\ 6)$$
$$\rho = (1)(4\ 5)(6\ 8)(2\ 7\ 3)$$

The proof then defins τ by matching up the corresponding numbers, i.e. $\tau(3) = 1$, $\tau(1) = 4$, $\tau(8) = 5$, etc. Working it out, we see that we can take

$$\tau = (1\ 4\ 2\ 6\ 3)(5\ 8)(7)$$

and that it will satisfy $\tau\sigma\tau^{-1} = \rho$.

The Class Equation

Given an action of G on X, we know that the orbits partition X. In particular, the conjugacy classes of a group G partition G. In particular, if you pick a unique element b_i from each of the m conjugacy classes of G, then we have

$$|G| = |\mathcal{O}_{b_1}| + |\mathcal{O}_{b_2}| + \cdots + |\mathcal{O}_{b_m}|$$

Furthermore, since we know by the Orbit-Stabilizer Theorem the the size of an orbit divides the order of G, it follows that every summand on the right is a divisor of G.

We can get a more useful version of the above equation if we bring together the orbits of size 1. Let's begin by examining which elements of G form a conjugacy class by themselves. We have

$$\begin{aligned}\mathcal{O}_a = \{a\} &\iff gag^{-1} = a \text{ for all } g \in G \\ &\iff ga = ag \text{ for all } g \in G \\ &\iff a \in Z(G)\end{aligned}$$

Thus, the elements which form a conjugacy class by themselves are exactly the elements of $Z(G)$. Now if we bring together these elements, and pick a unique element a_i from each of the k conjugacy classes of size at least 2, we get the equation:

$$|G| = |Z(G)| + |\mathcal{O}_{a_1}| + |\mathcal{O}_{a_2}| + \cdots + |\mathcal{O}_{a_k}|$$

Since $Z(G)$ is a subgroup of G, we may use Lagrange's Theorem to conclude that each of the summands on the right is a divisor of G. Furthermore, we have $|\mathcal{O}_{a_i}| \geq 2$ for all i, although it might happen that $|Z(G)| = 1$. Finally, using the Orbit-Stabilizer Theorem, we can rewrite this latter equation as

$$|G| = |Z(G)| + [G : C_G(a_1)] + [G : C_G(a_2)] + \cdots + [G : C_G(a_k)]$$

where again each of the summands on the right is a divisor of G and $[G : C_G(a_i)] \geq 2$ for all i. These latter two equations (either variation) is known as the *Class Equation* for G.

We saw above that the class equation for S_3 reads as

$$6 = 1 + 2 + 3$$

In Problem 2c on Homework 10, you computed the number of elements of S_5 of each given cycle structure, so the class equation for S_5 reads as

$$120 = 1 + 10 + 15 + 20 + 20 + 24 + 30$$

Theorem 9.2.4. *Let $p \in \mathbb{N}^+$ be prime and let G be a group with $|G| = p^n$ for some $n \in \mathbb{N}^+$. We then have that $|Z(G)| \neq \{e\}$.*

Proof. Let $a_1, a_2, \ldots, a_k$ be representatives from the conjugacy classes of size at least 2. By the Class Equation, we know that

$$|G| = |Z(G)| + |\mathcal{O}_{a_1}| + |\mathcal{O}_{a_2}| + \cdots + |\mathcal{O}_{a_k}|$$

By the Orbit-Stabilizer Theorem, we know that $|\mathcal{O}_{a_i}|$ divides p^n for each i. By the Fundamental Theorem of Arithmetic, it follows that each $|\mathcal{O}_{a_i}|$ is one of $1, p, p^2, \ldots, p^n$. Now we know that $|\mathcal{O}_{a_i}| > 1$ for each i, so p divides each $|\mathcal{O}_{a_i}|$. Since

$$|Z(G)| = |G| - |\mathcal{O}_{a_1}| - |\mathcal{O}_{a_2}| - \cdots - |\mathcal{O}_{a_k}|$$

and p divides every term on the right-hand side, we conclude that $p \mid |Z(G)|$. In particular, $Z(G) \neq \{e\}$. □

Corollary 9.2.5. *Let $p \in \mathbb{N}^+$ be prime. Every group of order p^2 is abelian.*

Proof. Let G be a group of order p^2. By Problem 4b on Homework 16, we know that either $Z(G) = \{e\}$ or $Z(G) = G$. The former is impossible by the previous theorem, so $Z(G) = G$ and hence G is abelian. □

Proposition 9.2.6. *Let $p \in \mathbb{N}^+$ be prime. If G is a group of order p^2, then either*

$$G \cong \mathbb{Z}/p^2\mathbb{Z} \qquad \text{or} \qquad G \cong \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$$

Therefore, up to isomorphism, there are exactly two groups of order p^2.

Proof. Suppose that G is a group of order p^2. By the previous corollary, G is abelian. If there exists an element of G with order p^2, then G is cyclic and $G \cong \mathbb{Z}/p^2\mathbb{Z}$. Suppose then that G has no element of order p^2. By Lagrange's Theorem, the order of every element divides p^2, so the order of every nonidentity element of G must be p. Fix $a \in G$ with $a \neq e$, and let $H = \langle a \rangle$. Since $|H| = p < |G|$, we may fix $b \in G$ with $b \notin H$ and let $K = \langle b \rangle$. Notice that H and K are both normal subgroups of G because G is abelian. Now $H \cap K$ is a subgroup of K, so $|H \cap K|$ divides $|K| = p$. We can't have $|H \cap K| = p$, because this would imply that $H \cap K = K$, which would contradict the fact that $b \notin H$. Therefore, we must have $|H \cap K| = 1$ and hence $H \cap K = \{e\}$. Since $|H| \cdot |K| = p^2 = |G|$, we may use Corollary 8.1.17 to conclude that G is the internal direct product of H and K. Since H and K are both cyclic of order p, they are both isomorphic to $\mathbb{Z}/p\mathbb{Z}$, so

$$G \cong H \times K \cong \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$$

Finally notice that $\mathbb{Z}/p^2\mathbb{Z} \ncong \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ because the first group is cyclic while the second is not (by Problem 5 on Homework 14, for example). □

We are now in a position to extend Theorem 7.4.4 to arbitrary groups.

Theorem 9.2.7 (Cauchy's Theorem)**.** *Let $p \in \mathbb{N}^+$ be prime. If G is a group with $p \mid |G|$, then G has an element of order p.*

Proof. The proof is by induction on $|G|$. If $|G| = 1$, then the result is trivial because $p \nmid 1$ (again if you don't like this vacuous base case, simply note the if $|G| = p$, then every nonidentity element of G has order p by Lagrange's Theorem). Suppose then that G is a finite group with $p \mid |G|$, and suppose that the result is true for all groups K satisfying $p \mid |K|$ and $|K| < |G|$. Let $a_1, a_2, \ldots, a_k$ be representatives from the conjugacy classes of size at least 2. By the Class Equation, we know that

$$|G| = |Z(G)| + [G : C_G(a_1)] + [G : C_G(a_2)] + \cdots + [G : C_G(a_k)]$$

We have two cases.

- Suppose that $p \mid [G : C_G(a_i)]$ for all i. Since

  $$|Z(G)| = |G| - [G : C_G(a_1)] + [G : C_G(a_2)] + \cdots + [G : C_G(a_k)]$$

  and p divides every term on the right, it follows that $p \mid |Z(G)|$. Now $Z(G)$ is a abelian group, so Theorem 7.4.4 tells us that $Z(G)$ has an element of order p. Thus, G has an element of order p.

- Suppose that $p \nmid [G : C_G(a_i]$ for some i. Fix such an i. By Lagrange's Theorem we ahve

  $$|G| = [G : C_G(a_i)] \cdot |C_G(a_i)|$$

  Since p is a prime number with both $p \mid |G|$ and $p \nmid [G : C_G(a_i)]$, we must have that $p \mid |C_G(a_i)|$. Now $|C_G(a_i)| < |G|$ because $a_i \notin Z(G)$, so by induction, $C_G(a_i)$ has an element of order p. Therefore, G has an element of order p.

The result follows by induction. □

9.3 Simplicity of A_5

Suppose that G is a group and that H is a normal subgroup of G. We then have that $ghg^{-1} \in H$ for all $g \in G$ and all $h \in H$. Thus, if $h \in H$, then every conjugate of h in G must also be in H. In other words, if H contains an element of some conjugacy class of G, then H must contain the entire conjugacy class. All of this works in the other direction as well and we get the following fact.

Proposition 9.3.1. *Let H be a subgroup of G. We then have that H is a normal subgroup of G if and only if H is a union of conjugacy classes of G.*

If we understand the conjugacy classes of a group G, we can use this proposition to help us understand the normal subgroups of G. Let's begin such an analysis by looking at S_4. Recall in Homework 10 that we counted the number of elements of S_n of various cycle types. In particular, we showed that if $k \leq n$, then the number of k-cycles in S_n equals:

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k}$$

Also, if $n \geq 4$, we showed that the number of permutations in S_n which are the product of two disjoint 2-cycles equals

$$\frac{n(n-1)(n-2)(n-3)}{8}$$

Using these results, we see that S_4 consists of the following numbers of elements of each cycle type.

- Identity: 1
- 2-cycles: $\frac{4\cdot 3}{2} = 6$
- 3-cycles: $\frac{4\cdot 3\cdot 2}{3} = 8$
- 4-cycles: $\frac{4\cdot 3\cdot 2\cdot 1}{4} = 6$
- Product of two disjoint 2-cycles: $\frac{4\cdot 3\cdot 2\cdot 1}{8} = 3$.

Since two elements of S_4 are conjugates in S_4 exactly when they have the same cycle type, this breakdown gives the conjugacy classes of S_4. In particular, the class equation of S_4 is:

$$24 = 1 + 3 + 6 + 6 + 8$$

Using this class equation, let's examine the possible normal subgroups of S_4. We already know that A_4 is a normal subgroup of S_4 since it has index 2 in S_4. However another way to see this is that A_4 contains the identity, the 3-cycles, and the products of two disjoint 2-cycles, so it is a union of conjugacy classes of S_4. In particular, it arises from taking the 1, the 3, and the 8 in the above class equation and putting them together to form a subgroup of size $1 + 3 + 8 = 12$.

Aside from the trivial examples of $\{id\}$ and S_4 itself, are there any other normal subgroups of S_4? Suppose that H is a normal subgroup of S_4 with $\{id\} \subsetneq H \subsetneq S_4$ and $H \neq A_4$. We certainly know that $id \in H$. By Lagrange's Theorem, we know we must have that $|H| \mid 24$. We also know that H must be a union of conjugacy classes of S_4. Thus, we would need to find a way to add some collection of the various numbers in the above class equation, necessarily including the number 1, such that their sum is a divisor of 24. One way is $1 + 3 + 8 = 12$ which gave A_4. Working through the various possibilities, we see that they only other nontrivial way to make it work is $1 + 3 = 4$. This corresponds to the subset

$$\{id, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$$

Now this subset is certainly closed under conjugation, but it is not immediately obvious that it is a subgroup. Each element here has order 2, so it is closed under inverses. Performing the simple check, it turns out that it is also closed under composition, so indeed this is another normal subgroup of S_4. Thus, the normal subgroups of S_4 are $\{id\}$, S_4, A_4, and this subgroup of order 4.

Now let's examine the possible normal subgroups of A_4. We already know three examples: $\{id\}$, A_4, and the just discovered subgroup of S_4 of size 4 (it is contained in A_4, and it must be normal in A_4 because it is normal in S_4). Now the elements of A_4 are the identity, the 3-cycles, and the products of two disjoint 2-cycles. Although the set of 3-cycles forms one conjugacy class in S_4, the set of eight 3-cycles does not form one conjugacy class in A_4. We can see this immediately because $|A_4| = 12$ and $8 \nmid 12$. The problem is that the elements of S_4 which happen to conjugate one 3-cycle to another might all be odd permutations and so do not exist in A_4. How can we determine the conjugacy classes in A_4 without simply plowing through all of the calculations from scratch?

Let's try to work out the conjugacy class of (1 2 3) in A_4. First notice that we know that the conjugacy class of (1 2 3) in S_4 has size 8, so by the Orbit-Stabilizer Theorem we conclude that $|C_{S_4}((1\ 2\ 3))| = \frac{24}{8} = 3$. Since $\langle(1\ 2\ 3)\rangle \subseteq C_{S_4}((1\ 2\ 3))$ and $|\langle(1\ 2\ 3)\rangle| = 3$, it follows that $C_{S_4}((1\ 2\ 3)) = \langle(1\ 2\ 3)\rangle$. Now $\langle(1\ 2\ 3)\rangle \subseteq A_4$, so we conclude that $C_{A_4}((1\ 2\ 3)) = \langle(1\ 2\ 3)\rangle$ as well. Therefore, by the Orbit-Stabilizer Theorem, the conjugacy class of (1 2 3) in A_4 has size $\frac{12}{3} = 4$. If you want to work out what exactly it is, it suffices to find 4 conjugates of (1 2 3) in A_4. Fortunately, we know how to compute conjugates quickly in S_n using Lemma 9.2.2 and the discussing afterwards:

- $id(1\ 2\ 3)id^{-1} = (1\ 2\ 3)$
- $(1\ 2\ 4)(1\ 2\ 3)(1\ 2\ 4)^{-1} = (2\ 4\ 3)$
- $(2\ 3\ 4)(1\ 2\ 3)(2\ 3\ 4)^{-1} = (1\ 3\ 4)$
- $(1\ 2)(3\ 4)(1\ 2\ 3)[(1\ 2)(3\ 4)]^{-1} = (2\ 1\ 4) = (1\ 4\ 2)$

Thus, the conjugacy class of (1 2 3) in A_4 is:

$$\{(1\ 2\ 3), (1\ 3\ 4), (1\ 4\ 2), (2\ 4\ 3)\}$$

If you work with a 3-cycle not in this set (for example, (1 2 4)), the above argument works through to show that its conjugacy class also has size 4, so its conjugacy class must be the other four 3-cycles in A_4. Thus, we get the conjugacy class

$$\{(1\ 2\ 4), (1\ 3\ 2), (1\ 4\ 3), (2\ 3\ 4)\}$$

Finally, let's look at the conjugacy class of (1 2)(3 4) in A_4. We have

- $id(1\ 2)(3\ 4)id^{-1} = (1\ 2)(3\ 4)$
- $(2\ 3\ 4)(1\ 2)(3\ 4)(2\ 3\ 4)^{-1} = (1\ 3)(4\ 2) = (1\ 3)(2\ 4)$
- $(2\ 4\ 3)(1\ 2)(3\ 4)(2\ 4\ 3)^{-1} = (1\ 4)(2\ 3)$

so the products of two disjoint 2-cycles still form one conjugacy class in A_4. Why did this conjugacy class not break up? By the Orbit-Stabilizer Theorem, we know that $|C_{S_4}((1\ 2)(3\ 4))| = \frac{24}{3} = 8$. If you actually compute this centralizer, you will see that 4 of its elements are even permutations and four of its elements are odd permutations. Therefore, $|C_{A_4}((1\ 2)(3\ 4))| = 4$ (the four even permutations in $C_{S_4}((1\ 2)(3\ 4))$), hence using the Orbit-Stabilizter Theorem again we conclude that the conjugacy class of (1 2)(3 4) in A_4 has size $\frac{12}{4} = 3$.

Putting everything together, we see that the class equation of A_4 is:

$$12 = 1 + 3 + 4 + 4$$

Working through the possibilities as in S_4, we conclude that the three normal subgroups of A_4 we found above are indeed all of the normal subgroups of A_4 (there is no other way to add some subcollection of these numbers which includes 1 to obtain a divisor of 12). Notice that we can also conclude that following.

Proposition 9.3.2. *A_4 has no subgroup of order 6, so the converse of Lagrange's Theorem is false.*

Proof. If H was a subgroup of A_4 with $|H| = 6$, then H would have index 2 in A_4, so would be normal in A_4. However, we just saw that A_4 has no normal subgroup of order 6. □

In the case of S_4 that we just worked through, we saw that when we restrict down to A_4, some conjugacy classes split into two and others stay intact. This is a general phenomenon in S_n, as we now show. We first need the following fact.

Lemma 9.3.3. *Let H be a subgroup of S_n. If H contains an odd permutation, then $|H \cap A_n| = \frac{|H|}{2}$.*

Proof. Suppose that H contains an odd permutation, and fix such an element $\tau \in H$. Let $X = H \cap A_n$ be the set of even permutation in H and let $Y = H \backslash A_n$ be the set of odd permutations in H. Define $f\colon X \to S_n$ by letting $f(\sigma) = \sigma\tau$. We claim that f maps X bijectively onto Y. We check the following:

- f is injective: Suppose that $\sigma_1, \sigma_2 \in X$ with $f(\sigma_1) = f(\sigma_2)$. We then have $\sigma_1\tau = \sigma_2\tau$, so $\sigma_1 = \sigma_2$ by cancellation.
- $\text{range}(f) \subseteq Y$. Let $\sigma \in X$. We have $\sigma \in H$, so $\sigma\tau \in H$ because H is a subgroup of S_n. Also, σ is a even permutation and τ is an odd permutation, so $\sigma\tau$ is an odd permutation. It follows that $f(\sigma) = \sigma\tau \in Y$.
- $Y \subseteq \text{range}(f)$: Let $\rho \in Y$. We then have $\rho \in H$ and $\tau \in H$, so $\rho\tau^{-1} \in H$ because H is a subgroup of S_n. Also, we have that both ρ and τ^{-1} are odd permutations, so $\rho\tau^{-1}$ is an even permutation. It follows that $\rho\tau^{-1} \in X$ and since $f(\rho\tau^{-1}) = \rho\tau^{-1}\tau = \rho$, we conclude that $\rho \in \text{range}(f)$.

Therefore, f maps X bijectively onto Y, and hence $|X| = |Y|$. Since $H = X \cup Y$ and $X \cap Y = \emptyset$, we conclude that $|H| = |X| + |Y|$. It follows that $|H| = 2 \cdot |X|$, so $|X| = \frac{|H|}{2}$. □

Proposition 9.3.4. *Let $\sigma \in A_n$. Let X be the conjugacy class of σ in S_n, and let Y be the conjugacy class of σ in A_n.*

- *If σ commutes with some odd permutation in S_n, then $Y = X$.*
- *If σ does not commute with any odd permutation in S_n, then $Y \subseteq X$ with $|Y| = \frac{|X|}{2}$.*

Proof. First notice that $X \subseteq A_n$ because $\sigma \in A_n$ and all elements of X have the same cycle type as σ. Let $H = C_{S_n}(\sigma)$ and let $K = C_{A_n}(\sigma)$. Notice that $Y \subseteq X$ and $K = H \cap A_n$. By the Orbit-Stabilizer Theorem applied in each of S_n and A_n, we know that

$$|H| \cdot |X| = n! \qquad \text{and } |K| \cdot |Y| = \frac{n!}{2}$$

and therefore

$$2 \cdot |K| \cdot |Y| = |H| \cdot |X|$$

Suppose first that σ commutes with some odd permutation in S_n. We then have H contains an odd permutation, so by the lemma we know that $|K| = |H \cap A_n| = \frac{|H|}{2}$. Plugging this into the above equation, we see that $|H| \cdot |Y| = |H| \cdot |X|$, so $|Y| = |X|$. Since $Y \subseteq X$, it follows that $Y = X$.

Suppose now that σ does not commute with any odd permutation in S_n. We then have that $H \subseteq A_n$, so $K = H \cap A_n = H$ and hence $|K| = |H|$. Plugging this into the above equation, we see that $2 \cdot |H| \cdot |Y| = |H| \cdot |X|$, so $|Y| = \frac{|X|}{2}$. □

Let's put all of the knowledge to work in to study A_5. We begin by recalling Problem 2c on Homework 10, where you computed the number of elements of S_5 of each given cycle type:

- Identity: 1
- 2-cycles: $\frac{5\cdot 4}{2} = 10$
- 3-cycles: $\frac{5\cdot 4\cdot 3}{3} = 20$
- 4-cycles: $\frac{5\cdot 4\cdot 3\cdot 2}{4} = 30$
- 5-cycles: $\frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{5} = 24$
- Product of two disjoint 2-cycles: $\frac{5\cdot 4\cdot 3\cdot 2}{8} = 15$
- Product of a 3-cycle and a 2-cycle which are disjoint: This equals the number of 3-cycles as discussed above, which is 20 from above.

This gives the following class equation for S_5:

$$120 = 1 + 10 + 15 + 20 + 20 + 24 + 30$$

Now in A_5 we only have the identity, the 3-cycles, the 5-cycles, and the product of two disjoint 2-cycles. Let's examine what happens to these conjugacy classes in A_5. We know that each of these conjugacy classes either stays intact or breaks in half by the above proposition.

- The set of 3-cycles has size 20, and since (1 2 3) commutes with the odd permutation (4 5), this conjugacy class stays intact in A_5.
- The set of 5-cycles has size 24, and since $24 \nmid 60$, it is not possible that this conjugacy class stays intact in A_5. Therefore, the set of 5-cycles breaks up into two conjugacy classes of size 12.
- The set of products of two disjoint 2-cycles has size 15. Since this is an odd number, it is not possible that it breaks into two pieces of size $\frac{15}{2}$, so it must remain a conjugacy class in A_5.

Therefore, the class equation for A_5 is:

$$60 = 1 + 12 + 12 + 15 + 20$$

With this in hand, we can examine the normal subgroups of A_5. Of course we know that $\{id\}$ and A_5 are normal subgroups of S_5. Suppose that H is a normal subgroup of A_5 with $\{id\} \subsetneq H \subsetneq A_5$. We then have that $id \in H$, that $|H| \mid 60$, and that H is a union of conjugacy classes of A_5. Thus, we would need to find a way to add some collection of the various numbers in the above class equation, necessarily including the number 1, such that their sum is a divisor of 60. Working through the possibilities, we see that this is not possible except in the cases when we take all the numbers (corresponding to A_5) and we only take 1 (corresponding to $\{id\}$). We have proved the following important theorem.

Theorem 9.3.5. *A_5 is a simple group.*

In fact, A_n is a simple group for all $n \geq 5$, but the above method of proof falls apart for $n > 5$ (the class equations get too long and there occasionally are ways to add certain numbers to get divisors of $|A_n|$). Thus, you need some new techniques. One approach is to use induction on $n \geq 5$ starting with the base case that we just proved, but we will not go through all the details here.

9.4 Counting Orbits

Suppose that you color each vertex of a square with one of n colors. In total, there are n^4 many ways to do this because you have n choices for each of the 4 vertices. However, some of these colorings are the "same" up to symmetry. For example, if you color the upper right vertex red and the other 3 green, then you can get this by starting with the coloring the upper left vertex red, the others green, and rotating. How many possible colorings are there up to symmetry?

We can turn this question into a question about counting orbits of a given group action. We begin by letting

$$X = \{(a_1, a_2, a_3, a_4) : 1 \leq a_i \leq n \text{ for all } i\}$$

Intuitively, we are labeling the four vertices of the square with the numbers $1, 2, 3, 4$, and we are letting a_i denote the color of vertex i. If you use the labeling where upper left corner 1, the upper right 2, the lower right 3, and the lower left 4, then the above above colorings are

$$(R, G, G, G) \qquad \text{and} \qquad (G, R, G, G)$$

where R is number you associated to color red and G the number for color green. Of course, these are distinct elements of X, but we want to make them the "same". Notice that S_4 acts on X via the action:

$$\sigma * (a_1, a_2, a_3, a_4) = (a_{\sigma(1)}, a_{\sigma(2)}, a_{\sigma(3)}, a_{\sigma(4)})$$

For example, we would have

$$(1\ 2\ 3) * (a_1, a_2, a_3, a_4) = (a_2, a_3, a_1, a_4)$$

so

$$(1\ 2\ 3) * (R, B, Y, G) = (B, Y, R, G)$$

Now we really don't want to consider these two coloring the same because although you can get from one to the other via an element of S_4, you can't do it from a rigid motion of the square. However, since D_4 is a subgroup of S_4, we get restrict the above to an action of D_4 on X. Recall that when viewed as a subgroup of S_4 we can list the elements of D_4 as:

$$\begin{array}{cccc} id & r = (1\ 2\ 3\ 4) & r^2 = (1\ 3)(2\ 4) & r^3 = (1\ 4\ 3\ 2) \\ s = (2\ 4) & rs = (1\ 2)(3\ 4) & r^2 s = (1\ 3) & r^3 s = (1\ 4)(2\ 3) \end{array}$$

The key insight is that two colorings are the same exactly when they are in the same orbit of this action by D_4. For example, we have the following orbits:

- $\mathcal{O}_{(R,R,R,R)} = \{(R, R, R, R)\}$
- $\mathcal{O}_{(R,G,G,G)} = \{(R, G, G, G), (G, R, G, G), (G, G, R, G), (G, G, G, R)\}$
- $\mathcal{O}_{(R,G,R,G)} = \{(R, G, R, G), (G, R, G, R)\}$

Thus, to count the number of colorings up to symmetry, we want to count the number of orbits of this action. The problem in attacking this problem directly is that the orbits have different sizes, so you can not simply divide n^4 by the common size of the orbits. We need a better way to count the number of orbits of an action.

Definition 9.4.1. *Suppose that G acts on X. For each $g \in G$, let $X_g = \{x \in X : g * x = x\}$. The set X_g is called the* fixed-point set *of g.*

Theorem 9.4.2 (Burnside's Lemma - due originally to Cauchy - sometimes also attributed to Frobenius)**.** *Suppose that G acts on X and that both G and X are finite. If k is the number of orbits of the action, then*

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

Thus, the number of orbits is the average number of elements fixed by each $g \in G$.

Proof. We the count the set $A = \{(g, x) \in G \times X : g * x = x\}$ in two different ways. On the one hand, for each $g \in G$, there are $|X_g|$ many elements of A in the "row" corresponding to g, so $|A| = \sum_{g\in G}|X_g|$. On the other hand, for each $x \in X$, there are $|G_x|$ many elements of A in the "column" corresponding to x, so $|A| = \sum_{x\in X}|G_x|$. Using the Orbit-Stabilizer Theorem, we know that

$$\sum_{g\in G}|X_g| = |A| = \sum_{x\in X}|G_x| = \sum_{x\in X}\frac{|G|}{|\mathcal{O}_x|} = |G|\cdot\sum_{x\in X}\frac{1}{|\mathcal{O}_x|}$$

and therefore

$$\frac{1}{|G|}\sum_{g\in G}|X_g| = \sum_{x\in X}\frac{1}{|\mathcal{O}_x|}$$

Let's examine this latter sum. Let $P_1, P_2, \ldots, P_k$ be the distinct orbits of X. We then have that

$$\sum_{x\in X}\frac{1}{|\mathcal{O}_x|} = \sum_{i=1}^{k}\sum_{x\in P_i}\frac{1}{|P_i|} = \sum_{i=1}^{k}|P_i|\cdot\frac{1}{|P_i|} = \sum_{i=1}^{k}1 = k$$

Therefore,

$$\frac{1}{|G|}\sum_{g\in G}|X_g| = \sum_{x\in X}\frac{1}{|\mathcal{O}_x|} = k$$

□

Of course, Burnside's Lemma will only be useful if it is not hard to compute the various values $|X_g|$. Let's return to our example of D_4 acting on the set X above. First notice that $X_{id} = X$, so $|X_{id}| = n^4$. Let move on to $|X_r|$ where $r = (1\ 2\ 3\ 4)$. Which elements (a_1, a_2, a_3, a_4) are fixed by r? We have $r * (a_1, a_2, a_3, a_4) = (a_2, a_3, a_4, a_1)$, so we need $a_1 = a_2$, $a_2 = a_3$, $a_3 = a_4$, and $a_4 = a_1$. Thus, an element (a_1, a_2, a_3, a_4) is fixed by r exactly when all the a_i are equal. There are n such choices (because we can pick a_1 arbitrarily, and then all the others are determined), so $|X_r| = n$. In general, given any $\sigma \in D_4$, an element of X is in X_σ exactly when all of the entries in each cycle of σ get the same color. Therefore, we have $|X_\sigma| = n^d$ where d is the number of cycles in the cycles notation of σ assuming that you include the 1-cycles. For example, we have $|X_{r^2}| = n^2$ and $|X_s| = n^3$. Working this out in the above cases and using the fact that $|D_4| = 8$, we conclude from Burnside's Lemma that the number of ways to color the vertices of the square with n colors up to symmetry is:

$$\frac{1}{8}(n^4 + n + n^2 + n + n^3 + n^2 + n^3 + n^2) = \frac{1}{8}(n^4 + 2n^3 + 3n^2 + 2n)$$

Let's examine the problem of coloring the faces of a cube. We will label the faces of a cube as a 6-sided die is labeled so that opposing faces sum to 7. For example, you could take the top face to be 1, the bottom 6, the front 2, the back 5, the right 3, and the left 4. With this labeling, the symmetries of the cube forms a subgroup of S_6. Letting G be this subgroup, notice that $|G| = 24$ because we can put any of the 6 faces on

top, and then rotate around the top in 4 distinct ways. If you work through the actually elements, you see that G equals the following subset of S_6:

id	(2 3 5 4)	(2 5)(3 4)	(2 4 5 3)
(1 2)(3 4)(5 6)	(1 3 2)(4 5 6)	(1 5 6 2)	(1 4 2)(3 5 6)
(1 2 3)(4 6 5)	(1 3)(2 5)(4 6)	(1 5 3)(2 4 6)	(1 4 6 3)
(1 2 4)(3 6 5)	(1 3 6 4)	(1 5 4)(2 3 6)	(1 4)(2 5)(3 6)
(1 2 6 5)	(1 3 5)(2 6 4)	(1 5)(2 6)(3 4)	(1 4 5)(2 6 3)
(1 6)(3 4)	(1 6)(2 3)(4 5)	(1 6)(2 5)	(1 6)(2 4)(3 5)

Let's examine how these elements arise.

- Identity: 1 of these, and it fixes all n^6 many colorings.
- 4-cycles: These are 90^o rotations around the line through the center of two opposing faces. There are 6 of these (3 such lines, and can rotate in either direction), and each fixes n^3 many colorings.
- Product of two 2-cycles: These are 180^o rotations around the line through the center of two opposing faces. There are 3 of these, and each fixes n^4 many colorings.
- Product of two 3-cycles: These are rotations around the line through opposite corners of the cube. There are 8 of these (there are four pairs of opposing corners, and then you can rotate either 120^o or 240^o for each), and each fixes n^2 many colorings .
- Product of three 2-cycles: These are 180^o rotations around a line through the middle of opposing edges of the cube. There are 6 of these (there are 6 such pairs of opposing edges), and each fixes n^3 many colorings.

Using Burnside's Lemma, we conclude that the total number of colorings of the faces of a cube using n colors up to symmetry is:

$$\frac{1}{24}(n^6 + 3n^4 + 12n^3 + 8n^2)$$

Chapter 10

Ring Theory

10.1 Definitions and Examples

Definition 10.1.1. *A* ring *is a set R equipped with two binary operations $+$ and $\cdot$ and two elements $0, 1 \in R$ such that*

1. *$a + (b + c) = (a + b) + c$ for all $a, b, c \in R$.*
2. *$a + b = b + a$ for all $a, b \in R$.*
3. *$a + 0 = a = 0 + a$ for all $a \in R$.*
4. *For all $a \in R$, there exists $b \in R$ with $a + b = 0 = b + a$.*
5. *$a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in R$.*
6. *$a \cdot 1 = a$ and $1 \cdot a = a$ for all $a \in R$.*
7. *$a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in R$.*
8. *$(a + b) \cdot c = a \cdot c + b \cdot c$ for all $a, b, c \in R$.*

Notice that the first four axioms simply say that $(R, +, 0)$ is an abelian group.

Some people omit property 6 and so do not require that their rings have a multiplicative identity. They call our rings either "rings with identity" or "rings with 1". We will not discuss such objects here, and for us "ring" implies that there is a multiplicative identity. Notice that we are requiring that $+$ is commutative but we do *not* require that $\cdot$ is commutative. Also, we are *not* requiring that elements have multiplicative inverses.

For example, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all rings with the standard notions of addition and multiplication along with the usual 0 and 1. For each $n \in \mathbb{N}$, the set $M_n(\mathbb{R})$ of *all* $n \times n$ matrices with entries from $\mathbb{R}$ is a ring where $+$ and $\cdot$ are the usual matrix addition and multiplication, 0 is the zero matrix, and $1 = I_n$ is the $n \times n$ identity matrix. Certainly some matrices in $M_n(\mathbb{R})$ fail to be invertible, but that is not a problem because the ring axioms say nothing about the existence of multiplicative inverses. Notice that the multiplicative *group* $GL_n(\mathbb{R})$ is not a ring because it is not closed under addition. For example

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

but this latter matrix is not an element of $GL_2(\mathbb{R})$. The next proposition gives some examples of finite rings.

Proposition 10.1.2. *Let $n \in \mathbb{N}^+$. The set $\mathbb{Z}/n\mathbb{Z}$ of equivalence classes of the relation $\equiv_n$ with operations*

$$\overline{a} + \overline{b} = \overline{a+b} \qquad \text{and} \qquad \overline{a} \cdot \overline{b} = \overline{ab}$$

is a ring with additive identity $\overline{0}$ and multiplicative identity $\overline{1}$.

Proof. We already know that $(\mathbb{Z}/n\mathbb{Z}, +, \overline{0})$ is an abelian group. We proved that multiplication of equivalence classes given by

$$\overline{a} \cdot \overline{b} = \overline{ab}$$

is well-defined in Proposition 3.3.4, so it gives a binary operation on $\mathbb{Z}/n\mathbb{Z}$. It is now straightforward to check that $\overline{1}$ is a multiplicative identity, that $\cdot$ is associative, and that $\cdot$ distributes over addition by appealing to these facts in $\mathbb{Z}$. □

You may find our requirement that $+$ is commutative in rings a bit surprising. In fact, this axiom follows from the others (so logically we could omit it). You can see this as follows. Suppose that we satisfy all of the above except possible for 2. Let $a, b \in R$. We then have

$$(1+1)\cdot(a+b) = (1+1)\cdot a + (1+1)\cdot b = 1\cdot a + 1\cdot a + 1\cdot b + 1\cdot b = a + a + b + b$$

and also

$$(1+1)\cdot(a+b) = 1\cdot(a+b) + 1\cdot(a+b) = a + b + a + b$$

hence

$$a + a + b + b = a + b + a + b$$

Now we are assuming that $(R, +, 0)$ is a group because we have axioms 1, 3, and 4, so by left and right cancellation we conclude that $a + b = b + a$.

If R is a ring, then we know that $(R, +, 0)$ is a group. In particular, every $a \in R$ has a unique additive inverse. Since we have another binary operation in multiplication, we choose different notation new notation for the additive inverse of a (rather than use the old a^{-1} which looks multiplicative).

Definition 10.1.3. *Let R be a ring. Given $a \in R$, we let $-a$ be the unique additive inverse of a, so $a + (-a) = 0$ and $(-a) + a = 0$.*

Proposition 10.1.4. *Let R be a ring.*

1. *For all $a \in R$, we have $a \cdot 0 = 0 = 0 \cdot a$.*
2. *For all $a, b \in R$, we have $a \cdot (-b) = -(a \cdot b) = (-a) \cdot b$.*
3. *For all $a, b \in R$, we have $(-a) \cdot (-b) = a \cdot b$.*
4. *For all $a \in R$, we have $-a = (-1) \cdot a$.*

Proof.

1. Let $a \in R$. We have

   $$a \cdot 0 = a \cdot (0 + 0) = (a \cdot 0) + (a \cdot 0)$$

   Therefore

   $$0 + (a \cdot 0) = (a \cdot 0) + (a \cdot 0)$$

   By cancellation in the group $(R, +, 0)$, it follows that $a \cdot 0 = 0$. Similarly we have

   $$0 \cdot a = (0 + 0) \cdot a = (0 \cdot a) + (0 \cdot a)$$

   Therefore

   $$0 + (0 \cdot a) = (0 \cdot a) + (0 \cdot a)$$

   By cancellation in the group $(R, +, 0)$, it follows that $0 \cdot a = 0$.

2. Let $a, b \in R$. We have
$$0 = a \cdot 0 = a \cdot (b + (-b)) = (a \cdot b) + (a \cdot (-b))$$
Therefore, $a \cdot (-b)$ is the additive inverse of $a \cdot b$, which is to say that $a \cdot (-b) = -(a \cdot b)$. Similarly
$$0 = 0 \cdot b = (a + (-a)) \cdot b = (a \cdot b) + ((-a) \cdot b)$$
so $(-a) \cdot b$ is the additive inverse of $a \cdot b$, which is to say that $(-a) \cdot b = -(a \cdot b)$.

3. Let $a, b \in R$. Using 2, we have
$$(-a) \cdot (-b) = -(a \cdot (-b)) = -(-(a \cdot b)) = a \cdot b$$
where we have used the group theoretic fact that the inverse of the inverse is the original element.

4. Let $a \in R$. We have
$$(-1) \cdot a = -(1 \cdot a) = -a$$

□

Definition 10.1.5. *Let R be a ring. Given $a, b \in R$, we define $a - b = a + (-b)$.*

Proposition 10.1.6. *If R is a ring with $1 = 0$, then $R = \{0\}$.*

Proof. Suppose that $1 = 0$. For every $a \in R$, we have
$$a = a \cdot 1 = a \cdot 0 = 0$$
Thus, $a = 0$ for all $a \in R$. It follows that $R = \{0\}$. □

Definition 10.1.7. *A* commutative ring *is a ring R such that $\cdot$ is commutative, i.e. such that $a \cdot b = b \cdot a$ for all $a, b \in R$.*

Definition 10.1.8. *Let R be a ring. A* subring *of R is a subset $S \subseteq R$ such that*

- *S is an additive subgroup of R*
- *$1 \in S$*
- *$ab \in S$ whenever $a \in S$ and $b \in S$.*

Example 10.1.9. *Here are two important examples of subrings. The reason why the notation of one ring involves parentheses while notation for the other involves square brackets will be explained later.*

- *$\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ is a subring of $\mathbb{R}$. To see this, first notice that $0 = 0 + 0\sqrt{2} \in \mathbb{Q}(\sqrt{2})$ and $1 = 1 + 0\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. Suppose that $x, y \in \mathbb{Q}(\sqrt{2})$ and fix $a, b, c, d \in \mathbb{Q}$ with $x = a + b\sqrt{2}$ and $y = c + d\sqrt{2}$. We then have that*
$$x + y = (a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$$
and
$$-x = -(a + b\sqrt{2}) = (-a) + (-b)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$$
and
$$xy = (a + b\sqrt{2})(c + d\sqrt{2}) = ac + ad\sqrt{2} + bc\sqrt{2} + 2bd = (ac + 2bd) + (ad + bc)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$$
Therefore, $\mathbb{Q}(\sqrt{2})$ is a subring of $\mathbb{R}$.

- $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ *is a subring of* $\mathbb{C}$. *To see this, first notice that* $0 = 0 + 0i \in \mathbb{Z}[i]$ *and* $1 = 1 + 0i \in \mathbb{Z}[i]$. *Suppose that* $x, y \in \mathbb{Z}[i]$ *and fix* $a, b, c, d \in \mathbb{Z}$ *with* $x = a + bi$ *and* $y = c + di$. *We then have that*

$$x + y = (a + bi) + (c + di) = (a + c) + (b + d)i \in \mathbb{Z}[i]$$

and

$$-x = -(a + bi) = (-a) + (-b)i \in \mathbb{Z}[i]$$

and

$$xy = (a + bi)(c + di) = ac + adi + bci + bdi^2 = (ac - bd) + (ad + bc)i \in \mathbb{Z}[i]$$

Therefore, $\mathbb{Z}[i]$ *is a subring of* $\mathbb{C}$. *The ring* $\mathbb{Z}[i]$ *is called the ring of* Gaussian Integers

In our work on group theory, we made extensive use of subgroups of a given group G to understand G. In our discussion of rings, the concept of subrings will play a much smaller role. Some rings of independent interest can be seen as subrings of larger rings, such as $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Z}[i]$ above. However, we will not typically try to understand R by looking its subrings. Partly, this is due to the fact that we will spend a large amount of time working with infinite rings (as opposed to the significant amount of time we spent on finite groups, where Lagrange's Theorem played a key role). As we will see, our attention will turn toward certain subsets of a ring called *ideals* which play a role similar to normal subgroups of a group.

Finally, one small note. Some authors use a slightly different definition of a subring in that they do not require that $1 \in S$. They use the idea that a subring of R should be a subset $S \subseteq R$ which forms a ring with the inherited operations, but the multiplicative identity of S could be different from the multiplicative identity of R. Again, since subrings will not play a particularly important role for us, we will not dwell on this distinction.

10.2 Units and Zero Divisors

As we mentioned, our ring axioms do not require that elements have multiplicative inverses. Those elements which do have multiplicative inverses are given special names.

Definition 10.2.1. *Let* R *be a ring. An element* $u \in R$ *is a* unit *if it has a multiplicative inverse, i.e. there exists* $v \in R$ *with* $uv = 1 = vu$.

For example, the units in $\mathbb{Z}$ are $\{\pm 1\}$ and the units in $\mathbb{Q}$ are $\mathbb{Q}\backslash\{0\}$ (and similarly for $\mathbb{R}$ and $\mathbb{C}$). The units in $M_n(\mathbb{R})$ are the invertible $n \times n$ matrices. As in group theory, when inverses exist, they are unique.

Proposition 10.2.2. *Let* R *be a ring and let* $u \in R$ *be a unit. There is a unique* $v \in R$ *with* $uv = 1$ *and* $vu = 1$. *We denote unique element* u^{-1}.

Proof. Existence follows from the assumption that u is a unit. Suppose that v and w both work. We then have

$$\begin{aligned} v &= v \cdot 1 \\ &= v \cdot (u \cdot w) \\ &= (v \cdot u) \cdot w \\ &= 1 \cdot w \\ &= w \end{aligned}$$

hence $v = w$. □

Proposition 10.2.3. *Let* R *be a ring. The set of units of* R, *denoted* $U(R)$, *forms an group under multiplication with identity* 1.

Proof. Clearly $1 \in U(R)$ because $1 \cdot 1 = 1$. Suppose that $u \in U(R)$. We then have that $uu^{-1} = 1 = u^{-1}u$, so u^{-1} is a unit (since u works as a multiplicative inverse). Therefore, $U(R)$ is closed under inverses. Suppose that $u, v \in U(R)$. We then have

$$(uv)(v^{-1}u^{-1}) = u(vv^{-1})u^{-1} = u1u^{-1} = uu^{-1} = 1$$

and also

$$(v^{-1}u^{-1})(uv) = v^{-1}(u^{-1}u)v = v^{-1}1v = v^{-1}v = 1$$

so $uv \in U(R)$. Thus, $U(R)$ is closed under multiplication.

Since $U(R)$ is closed under multiplication, we see that $\cdot$ is a binary operation on $U(R)$. Now $\cdot$ is associative on $U(R)$ because it is associative on all of R, 1 is clearly an identity for $U(R)$, and every element of $U(R)$ has an inverse because $U(R)$ is closed under inverses. Therefore, $U(R)$ is a group under multiplication. □

This finally explains our notation for the group $U(\mathbb{Z}/n\mathbb{Z})$; namely, we are considering $\mathbb{Z}/n\mathbb{Z}$ as a ring and forming the corresponding unit group of that ring. Notice that for any $n \in \mathbb{N}^+$, we have $GL_n(\mathbb{R}) = U(M_n(\mathbb{R}))$.

Let's recall the multiplication table of $\mathbb{Z}/6\mathbb{Z}$:

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{0}$	$\overline{2}$	$\overline{4}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{2}$	$\overline{0}$	$\overline{4}$	$\overline{2}$
$\overline{5}$	$\overline{0}$	$\overline{5}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

Looking at the table, we see that $U(\mathbb{Z}/6\mathbb{Z}) = \{\overline{1}, \overline{5}\}$ as we already know. As remarked when we first saw this table, there are some interesting things. For example, we have $\overline{2} \cdot \overline{3} = \overline{0}$, so it is possible to have the product of two nonzero elements result in 0. Elements which are part of such a pair are given a name.

Definition 10.2.4. *Let R be ring. A nonzero element $a \in R$ is a* zero divisor *if there exists a nonzero $b \in R$ such that* either *$ab = 0$ or $ba = 0$.*

In the above case of $\mathbb{Z}/6\mathbb{Z}$, we see that the zero divisors are $\{\overline{2}, \overline{3}, \overline{4}\}$. The concept of unit and zero divisor are antithetical as we now see.

Proposition 10.2.5. *Let R be a ring. No element is both a unit and zero divisor.*

Proof. Suppose that R is a ring and that $a \in R$ is a unit. If $b \in R$ satisfies $ab = 0$, then

$$\begin{aligned} b &= 1 \cdot b \\ &= (a^{-1}a) \cdot b \\ &= a^{-1} \cdot (ab) \\ &= a^{-1} \cdot 0 \\ &= 0 \end{aligned}$$

Similarly, if $b \in R$ satisfies $ba = 0$, then

$$\begin{aligned} b &= b \cdot 1 \\ &= b \cdot (aa^{-1}) \\ &= (ba) \cdot a^{-1} \\ &= 0 \cdot a^{-1} \\ &= 0 \end{aligned}$$

Therefore, there is no nonzero $b \in R$ which satisfies either $ab = 0$ or $ba = 0$. It follows that a is not a zero divisor. □

Proposition 10.2.6. *Let $n \in \mathbb{N}^+$ and let $a \in \mathbb{Z}$.*

- *If $\gcd(a, n) = 1$, then $\overline{a}$ is a unit in $\mathbb{Z}/n\mathbb{Z}$.*
- *If $\gcd(a, n) > 1$ and $\overline{a} \neq \overline{0}$, then $\overline{a}$ is a zero divisor in $\mathbb{Z}/n\mathbb{Z}$.*

Proof. The first part is given by Proposition 3.4.3. Suppose that $\gcd(a, n) > 1$ and $\overline{a} \neq \overline{0}$. Let $d = \gcd(a, n)$ and notice that $0 < d < n$ because $n \nmid a$ (since we are assuming $\overline{a} \neq \overline{0}$). We have $d \mid n$ and $d \mid a$, so we may fix $b, c \in \mathbb{Z}$ with $db = n$ and $dc = a$. Notice that $0 < b < n$ because $d, n \in \mathbb{N}$ and $d > 1$, hence $\overline{b} \neq 0$. Now

$$ab = (dc)b = (db)c = nc$$

so $n \mid ab$. It follows that $\overline{a} \cdot \overline{b} = \overline{ab} = \overline{0}$. Since $\overline{b} \neq 0$, we conclude that $\overline{a}$ is a zero divisor. □

Therefore, every nonzero element of $\mathbb{Z}/n\mathbb{Z}$ is either a unit or a zero divisor. This is not true in every ring however. The units in $\mathbb{Z}$ are $\{\pm 1\}$ but there are no zero divisors in $\mathbb{Z}$. We now define three important classes of rings.

Definition 10.2.7. *Let R be a ring.*

- *R is a* division ring *if $1 \neq 0$ and every nonzero element of R is a unit.*
- *R is a* field *if R is a commutative division ring. Thus, a field is a commutative ring with $1 \neq 0$ for which every nonzero element is a unit.*
- *R is an* integral domain *if R is a commutative ring with $1 \neq 0$ which has no zero divisors. Equivalently, an integral domain is a commutative ring R with $1 \neq 0$ such that whenever $ab = 0$, either $a = 0$ or $b = 0$.*

Clearly every field is a division ring. We also have the following.

Proposition 10.2.8. *Every field is an integral domain.*

Proof. Suppose that R is a field. If $a \in R$ is nonzero, then it is a unit, so it is not a zero divisor. □

For example, each of $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields. The ring $\mathbb{Z}$ is an example of an integral domain which is not a field, so the concept of integral domain is strictly weaker. There also exist division rings which are not fields (one example is the *Hamiltonion Quaternions*, which are an extension of the complex numbers), but we will not digress to construct such objects now.

If $n \in \mathbb{N}^+$ is composite, then the ring $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring with zero divisors, so it is not an integral domain. However, for each prime p, the ring $\mathbb{Z}/p\mathbb{Z}$ is field (as see above, every nonzero element has a multiplicative inverse). This gives us an infinite supply of finite fields. Here are the addition and multiplication tables of $\mathbb{Z}/5\mathbb{Z}$ to get a picture of one of one of these objects.

$+$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{0}$	$\overline{1}$
$\overline{3}$	$\overline{3}$	$\overline{4}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{4}$	$\overline{4}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{1}$	$\overline{3}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{1}$	$\overline{4}$	$\overline{2}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

Another example of a field is the ring $\mathbb{Q}(\sqrt{2})$ discussed in the last section. To see that it is a field, we need only check that every nonzero element has an inverse. Suppose that $a, b \in \mathbb{Q}$ and that $a + b\sqrt{2} \neq 0$. Notice that we must have $a - b\sqrt{2} \neq 0$ because otherwise would $\sqrt{2}$ would be rational. Now

$$\begin{aligned}\frac{1}{a+b\sqrt{2}} &= \frac{1}{a+b\sqrt{2}} \cdot \frac{a-b\sqrt{2}}{a-b\sqrt{2}} \\ &= \frac{a-b\sqrt{2}}{a^2-2b^2} \\ &= \frac{1}{a^2-2b^2} + \frac{-b}{a^2-2b^2}\sqrt{2}\end{aligned}$$

Since $a, b \in \mathbb{Q}$, we have both $\frac{1}{a^2-2b^2} \in \mathbb{Q}$ and $\frac{-b}{a^2-2b^2} \in \mathbb{Q}$, so $\frac{1}{a+b\sqrt{2}} \in \mathbb{Q}(\sqrt{2})$.

Although we will spend a bit of time discussing noncommutative rings, the focus of our study will be commutative rings and often we will be working with integral domains (and sometimes more specifically with fields). The next proposition is a fundamental tool when working in integral domains. Notice that it can fail in arbitrary commutative rings. For example, in $\mathbb{Z}/6\mathbb{Z}$ we have $\overline{3} \cdot \overline{2} = \overline{3} \cdot \overline{4}$, but $\overline{2} \neq \overline{4}$

Proposition 10.2.9. *Suppose that R is an integral domain and that $ab = ac$ with $a \neq 0$. We then have that $b = c$.*

Proof. Since $ab = ac$, we have $ab - ac = 0$. Using the distributive law, we see that $a(b - c) = 0$ (more formally, we have $ab + (-ac) = 0$, so $ab + a(-c) = 0$, hence $a(b + (-c)) = 0$, and thus $a(b - c) = 0$. Since R is an integral domain, either $a = 0$ or $b - c = 0$. Now the former is impossible by assumption, so we conclude that $b - c = 0$. Adding c to both sides, we conclude that $b = c$. □

Although $\mathbb{Z}$ is an example of integral domain which is not a field, it turns out that all such examples are infinite.

Proposition 10.2.10. *Every finite integral domain is a field.*

Proof. Suppose that R is a finite integral domain. Let $a \in R$ with $a \neq 0$. Define $\lambda_a \colon R \to R$ by letting $\lambda_a(b) = ab$. Now if $b, c \in R$ with $\lambda_a(b) = \lambda_a(c)$, then $ab = ac$, so $b = c$ by the previous proposition. Therefore, $\lambda_a \colon R \to R$ is injective. Since R is finite, it must be the case that λ_a is surjective. Thus, there exists $b \in R$ with $\lambda_a(b) = 1$, which is to say that $ab = 1$. Therefore, a is a unit in R. Since $a \in R$ with $a \neq 0$ was arbitrary, it follows that R is a field. □

10.3 Polynomial Rings, Power Series Rings, and Matrix Rings

Polynomial Rings

Given a ring R, we show in this section how to build new rings from R. Our first example of a new ring intuitively equals the set of all "polynomials with coefficient in R". For example, you are used to working with polynomials over the real number $\mathbb{R}$ which look like

$$3x^7 + \sqrt{5}x^4 - 2x^2 + 142x - \pi$$

In this light, you add polynomials in the obvious way:

$$(4x^2 + 3x - 2) + (8x^3 - 2x^2 - 81x + 14) = 8x^3 + 2x^2 - 78x + 12$$

and you multiply polynomials in the "obvious" way:

$$(2x^2 + 5x - 3)(-7x^2 + 2x + 1) = -14x^4 - 31x^3 + 33x^2 - x - 3$$

In more detail, we multiplied by collecting like powers of x as follows:

$$(2\cdot(-7))x^4+(2\cdot 2+5\cdot(-7))x^3+(2\cdot 1+5\cdot 2+(-3)\cdot(-7))x^2+(5\cdot 1+(-3)\cdot 2)+(-3)\cdot 1$$

In general, we aim to carry over the above idea except we will allow our coefficients to come from a general ring R. Thus, a typical element should look like:

$$a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$$

where each $a_i\in R$. Before we dive into giving a formal definition, we first discuss polynomials a little more deeply.

In your previous work, you typically thought of a polynomial as providing describing a certain function from $\mathbb{R}$ to $\mathbb{R}$ resulting from "plugging in for x". Thus, when you wrote the polynomial $4x^2+3x-2$, you were probably thinking of it as the function which sends 1 to $4+3-2=5$, sends 2 to $16+6-2=20$, etc. In contrast, we will consciously avoid defining a polynomial as the resulting function obtained by "plugging in for x". To see why this distinction matters, consider the case where your ring $R=\mathbb{Z}/2\mathbb{Z}$ and you have two polynomials $\overline{1}x^2+\overline{1}x+\overline{1}$ and $\overline{1}$. These look like different polynomials on the face of it. However, notice that

$$\overline{1}\cdot\overline{0}^2+\overline{1}\cdot\overline{0}+\overline{1}=\overline{1}$$

and also

$$\overline{1}\cdot\overline{1}^2+\overline{1}\cdot\overline{1}+\overline{1}=\overline{1}$$

so these two distinct polynomials "evaluate" to same thing whenever you plug in elements of R (namely they both produce the function which outputs $\overline{1}$ for every input). Therefore, the resulting functions are indeed equal as functions despite the fact the polynomials have distinct forms.

We will enforce this distinction between polynomials and the resulting functions, so we will simply define our ring in a manner that two different looking polynomials are really distinct elements of our ring. In order, to do this carefully, let's go back and look at our polynomials. For example, consider the polynomial with real coefficients given by $5x^3+9x-2$. Since we are not "plugging in" for x, this polynomial is determined by its sequence of coefficients. In other words, if we order the sequence from coefficients of smaller powers to coefficients of larger powers, we can represent this polynomial as the sequence $(-2,9,0,5)$. Performing this step gets rid of the superfluous x which was really just serving as a placeholder and honestly did not have any real meaning. We will adopt this perspective and simply *define* a polynomial to be such a sequence. However, since polynomials can have arbitrarily large degree, these finite sequences would all of different lengths. We get around this problem by defining polynomials as infinite sequences of elements of R in which only finitely many of the terms are nonzero. Here's the formal definition.

Definition 10.3.1. *Let R be a ring. We define a new ring denoted $R[x]$ whose elements are the set of all infinite sequences $\{a_n\}$ where each $a_n\in R$ and $\{n\in\mathbb{N}:a_n\neq 0\}$ is finite. We define two binary operations on $R[x]$ as follows.*

$$\{a_n\}+\{b_n\}=\{a_n+b_n\}$$

and

$$\begin{aligned}\{a_n\}\cdot\{b_n\}&=\{a_0b_n+a_1b_{n-1}+\cdots+a_{n-1}b_1+a_nb_0\}\\&=\{\sum_{k=0}^{n}a_kb_{n-k}\}\\&=\{\sum_{i+j=n}a_ib_j\}\end{aligned}$$

We will see below that this makes $R[x]$ into a ring called the polynomial ring *over R.*

Let's pause for a moment to ensure that the above definition makes sense. Suppose that $\{a_n\}$ and $\{b_n\}$ are both infinite sequences of elements of R for which only finitely many nonzero terms. Fix $M, N \in \mathbb{N}$ such that $a_n = 0$ for all $n > M$ and $b_n = 0$ for all $n > N$. We then have that $a_n + b_n = 0$ for all $n > \max\{M, N\}$, so the infinite sequence $\{a_n + b_n\}$ only has finitely many nonzero terms. Also, we have

$$\sum_{i+j=n} a_i b_j = 0$$

for all $n > M + N$ (because if $n > M + N$ and $i + j = n$, then either $i > M$ or $j > N$, so either $a_i = 0$ or $b_j = 0$), hence the infinite sequence $\{\sum_{i+j=n} a_i b_j\}$ only has finitely many nonzero terms. We now check that these operations turn $R[x]$ into a ring.

Theorem 10.3.2. *Let R be a ring. The set $R[x]$ with the above operations is a ring with additive identity the infinite sequence $0, 0, 0, 0, \ldots$ and multiplicative identity the infinite sequence $1, 0, 0, 0, \ldots$. Furthermore, if R is commutative, then $R[x]$ is commutative.*

Proof. Many of these checks are routine. For example, $+$ is associative on $R[x]$ because we have

$$\begin{aligned}
\{a_n\} + (\{b_n\} + \{c_n\}) &= \{a_n\} + \{b_n + c_n\} \\
&= \{a_n + (b_n + c_n)\} \\
&= \{(a_n + b_n) + c_n\} && \text{(since } + \text{ is associative on } R\text{)} \\
&= \{a_n + b_n\} + \{c_n\} \\
&= (\{a_n\} + \{b_n\}) + \{c_n\}
\end{aligned}$$

The other ring axioms involving addition are completely analogous. However, when we get to multiplication life gets a bit more interesting because our multiplication operation is much more complicated than componentwise multiplication. For example, let $\{c_n\}$ be the infinite sequence $1, 0, 0, 0, \ldots$, i.e.

$$e_n = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n > 0 \end{cases}$$

Now for any elements $\{a_n\}$ of our ring, we have

$$\{e_n\} \cdot \{a_n\} = \{\sum_{k=0}^{n} e_k a_{n-k}\} = \{e_0 a_{n-0}\} = \{1 \cdot a_{n-0}\} = \{a_n\}$$

where we have used the fact that $c_k = 0$ whenever $k > 0$ We also have

$$\{a_n\} \cdot \{e_n\} = \{\sum_{k=0}^{n} a_k e_{n-k}\} = \{a_n e_0\} = \{a_n \cdot 1\} = \{a_n\}$$

where we have used the fact that if $0 \leq k < n$, then $n - k > 1$, so $e_{n-k} = 0$. Therefore, the infinite sequence $1, 0, 0, 0, \ldots$ is indeed a multiplicative identity of $R[x]$.

The most interesting (i.e. difficult) check is that $\cdot$ is associative on $R[x]$. We have

$$\begin{aligned}
\{a_n\} \cdot (\{b_n\} \cdot \{c_n\}) &= \{a_n\} \cdot \{ \sum_{j+k=n} b_j c_k \} \\
&= \{ \sum_{i+\ell=n} a_i \cdot (\sum_{j+k=\ell} b_j c_k) \} \\
&= \{ \sum_{i+\ell=n} \sum_{j+k=\ell} a_i (b_j c_k)) \} \\
&= \{ \sum_{i+(j+k)=n} a_i (b_j c_k)) \} \\
&= \{ \sum_{i+(j+k)=n} (a_i b_j) c_k \} \\
&= \{ \sum_{(i+j)+k=n} (a_i b_j) c_k \} \\
&= \{ \sum_{\ell+k=n} \sum_{i+j=\ell} (a_i b_j) c_k \} \\
&= \{ \sum_{\ell+k=n} (\sum_{i+j=\ell} a_i b_j) \cdot c_k \} \\
&= \{ \sum_{i+j=n} a_i b_j \} \cdot \{c_n\} \\
&= (\{a_n\} \cdot \{b_n\}) \cdot \{c_n\}
\end{aligned}$$

One also needs to check the two distributive laws and that $R[x]$ is commutative if R is commutative, but these are notably easier than associative of multiplication and are left to an exercise. □

Now the above formal definition of $R[x]$ is precise and useful when trying to prove theorems about polynomials, but it is terrible to work with intuitively. Thus, when discussing elements of $R[x]$, we will typically use standard polynomial notation. For example, if we are dealing with $\mathbb{Q}[x]$, we will simply write the formal element

$$(5, 0, 0, -\frac{1}{3}, 7, 0, \frac{22}{7}, 0, 0, 0, \dots)$$

as

$$\frac{22}{7}x^6 + 7x^4 - \frac{1}{3}x^3 + 5$$

and we will treat the x as a meaningless placeholder symbol called an *indeterminate*. This is where the x comes from in the notation $R[x]$ (if for some reason you want to use a different indeterminate, say t, you can instead use the notation $R[t]$). Formally, $R[x]$ will be the set of infinite sequences, but we use this more straightforward notation in the future when working with polynomials. However, as emphasized above, you should keep a clear distinction in your mind between an element of $R[x]$ and the function it represents via "evaluation" that we discuss below.

If you want to be very pedantic, you can connect up the formal definition of $R[x]$ (as infinite sequences of elements of R with finitely many nonzero terms) and the more gentle standard notation of polynomials as follows. Every element $a \in R$ naturally appears in $R[x]$ as the sequence $(a, 0, 0, 0, \dots)$. If you simply *define* x to be the sequence $(0, 1, 0, 0, 0, \dots)$, then working in the ring R it is not difficult to check that $x^2 = (0, 0, 1, 0, 0, \dots)$, that $x^3 = (0, 0, 0, 1, 0, \dots)$, etc. With these identifications, if you interpret the additions and multiplications implicit in the polynomial

$$\frac{22}{7}x^6 + 7x^4 - \frac{1}{3}x^3 + 5$$

as their formal counterparts defined above, then everything matches up as you would expect. Now when working in $R[x]$ with this notation, we will typically call elements of $R[x]$ by names like $p(x)$ and $q(x)$ and write something like "Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be an element of $R[x]$". Since I can't say this enough, do *not* simply view this as saying that $p(x)$ is the resulting function.

Definition 10.3.3. *Let R be a ring. Given a* nonzero *element $\{a_n\} \in R[x]$, we define the* degree *of $\{a_n\}$ to be* $\max\{n \in \mathbb{N} : a_n \neq 0\}$. *In the more relaxed notation, given a* nonzero *polynomial $p(x) \in R[x]$, say*

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

with $a_n \neq 0$, we define the degree *of $p(x)$ to be n. We write* $\deg(p(x))$ *for the degree of $p(x)$. Notice that we do not define the degree of the zero polynomial.*

The next proposition gives some relationship between the degrees of polynomials and the degrees of their sum/product.

Proposition 10.3.4. *Let R be a ring and let $p(x), q(x) \in R[x]$ be nonzero.*

- *Either $p(x) + q(x) = 0$ or $\deg(p(x) + q(x)) \leq \max\{\deg(p(x)), \deg(q(x))\}$*
- *Either $p(x) \cdot q(x) = 0$ or $\deg(p(x)q(x)) \leq \deg(p(x)) + \deg(q(x))$*

Proof. We give an argument using our formal notions. Let $p(x)$ be the sequence $\{a_n\}$ and let $q(x)$ be the sequence $\{b_n\}$. Let $M = \deg(p(x))$ so that $a_M \neq 0$ and $a_n = 0$ for all $n > M$. Let $N = \deg(q(x))$ so that $a_N \geq 0$ and $a_N = 0$ for all $n > N$.

Suppose that $p(x)+q(x) \neq 0$. For any $n > \max\{M, N\}$, we have $a_n + b_n = 0+0 = 0$, so $\deg(p(x)+q(x)) \leq \max\{M, N\}$.

Suppose that $p(x)q(x) \neq 0$. Let $n > M + N$ and consider the sum

$$\sum_{k=0}^{n} a_k b_{n-k} = 0$$

Notice that if $k > M$, then $a_k = 0$ so $a_k b_{n-k} = 0$. Also, if $k < M$, then $n-k > M+N-k = N+(M-k) > N$, hence $b_{n-k} = 0$ and so $a_k b_{n-k} = 0$. Thus, $a_k b_{n-k} = 0$ for all k with $0 \leq k \leq n$, so it follows that

$$\sum_{k=0}^{n} a_k b_{n-k} = 0$$

Therefore, $\deg(p(x)q(x)) \leq M + N$. □

Notice that in the ring $\mathbb{Z}/6\mathbb{Z}[x]$, we have

$$(\overline{2}x^2 + \overline{5}x + \overline{4})(\overline{3}x + \overline{1}) = \overline{1}x^2 + \overline{5}x + \overline{4}$$

so the product of a degree 2 polynomial and a degree 1 polynomial resulting in a degree 2 polynomials. It follows that we can indeed have a strict inequality in the latter case. In fact, we can have the product of two nonzero polynomials result in the zero polynomial, i.e. there may exist zero divisors in $R[x]$. For example, working in $\mathbb{Z}/6\mathbb{Z}[x]$ again we have

$$(\overline{4}x + \overline{2}) \cdot \overline{3}x^2 = \overline{0}$$

Fortunately, for well-behaved rings, the degree of the product always equals the sum of the degrees.

Proposition 10.3.5. *Let R be an integral domain. We then have that $R[x]$ is an integral domain. Furthermore, if $p(x), q(x) \in R[x]$ are both nonzero, then $p(x)q(x) \neq 0$ and $\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x))$.*

Proof. We again give an argument using our formal notions. Let $p(x)$ be the sequence $\{a_n\}$ and let $q(x)$ be the sequence $\{b_n\}$. Let $M = \deg(p(x))$ so that $a_M \neq 0$ and $a_n = 0$ for all $n > M$. Let $N = \deg(q(x))$ so that $a_N \geq 0$ and $a_N = 0$ for all $n > N$. Now consider

$$\sum_{k=0}^{M+N} a_k b_{M+N-k}$$

Notice that if $k > M$, then $a_k = 0$ so $a_k b_{M+N-k} = 0$. Also, if $k < M$, then $M + N - k = N + (M - k) > N$, hence $b_{M+N-k} = 0$ and so $a_k b_{M+N-k} = 0$. Therefore, we have

$$\sum_{k=0}^{M+N} a_k b_{M+N-k} = a_M b_{M+N-M} = a_M b_N$$

Now we have $a_M \neq 0$ and $b_N \neq 0$, so since R is an integral domain we know that $a_M b_N \neq 0$. Therefore,

$$\sum_{k=0}^{M+N} a_k b_{M+N-k} = a_M b_N \neq 0$$

It follows that $p(x)q(x) \neq 0$. Furthermore, we have $\deg(p(x)q(x)) \geq M + N$. Since we already know that $\deg(p(x)q(x)) \leq M + N$, we conclude that $\deg(p(x)q(x)) = M + N$. □

Power Series Rings

When we defined $R[x]$, the ring of polynomials with coefficients in R, we restricted our infinite sequences to have only a finite number of nonzero terms so that they "correspond" to polynomials. However, the entire apparatus we constructed goes through without a hitch if we take the set of all infinite sequences with no restriction. Intuitively, an infinite sequence $\{a_n\}$ corresponds to the "power series"

$$a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots$$

The nice thing about defining our objects as infinite sequences is that there is no confusion at all about "plugging in for x" because there is no x. Thus, there are no issues about convergence or whether this is a well-defined function at all. We define our $+$ and $\cdot$ on these infinite sequences exactly as in the polynomial ring case, and the proofs of the ring axioms follows word for word (actually, the proof is a bit easier because you don't have to worry about the resulting sequences having only a finite number of nonzero terms).

Definition 10.3.6. *Let R be a ring. We define a new ring denoted $R[[x]]$ whose elements are the set of all infinite sequences $\{a_n\}$ where each $a_n \in R$. We define two binary operations on $R[x]$ as follows.*

$$\{a_n\} + \{b_n\} = \{a_n + b_n\}$$

and

$$\begin{aligned}\{a_n\} \cdot \{b_n\} &= \{a_0 b_n + a_1 b_{n-1} + \cdots + a_{n-1} b_1 + a_n b_0\} \\ &= \{\sum_{k=0}^{n} a_k b_{n-k}\} \\ &= \{\sum_{i+j=n} a_i b_j\}\end{aligned}$$

These operations make $R[[x]]$ into a ring called the ring of formal power series *over R, or simply the* power series ring *over R.*

If you have worked with generating series in combinatorics as strictly combinatorial objects (i.e. you did not worry about values of x where convergence made sense, or work with resulting function on the restricted domain), then you were in fact working in the ring $\mathbb{R}[[x]]$. If you have worked with infinite series, then you know that

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$$

whenever $|x| < 1$. You have probably used this fact when you worked with generating functions as well, even if you weren't thinking about these as functions. If you are not thinking about the above equality in terms of functions (with worries about convergence on the right), then what you are doing is that you are working in the ring $R[[x]]$ and saying that $1 - x$ is a unit with inverse $1 + x + x^2 + x^3 + \dots$. We now formally verify this fact.

Proposition 10.3.7. *Let R be a ring. Define two elements $\{a_n\}$ and $\{b_n\}$ of $R[[x]]$ by*

$$a_n = \begin{cases} 1 & \text{if } n = 0 \\ -1 & \text{if } n = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad b_n = 1 \text{ for all } n$$

Letting e_n be the multiplicative identity of the ring $R[[x]]$, i.e.

$$e_n = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases}$$

we then have that

$$\{a_n\} \cdot \{b_n\} = \{e_n\} \qquad \text{and} \qquad \{b_n\} \cdot \{a_n\} = \{e_n\}$$

Thus, $\{a_n\}$ and $\{b_n\}$ are units in $R[[x]]$ and are inverses of each other. Less formally, working in the ring $R[[x]]$, we have

$$(1-x)(1 + x + x^2 + x^3 + \dots) = 1 \qquad \text{and} \qquad (1 + x + x^2 + x^3 + \dots)(1-x) = 1$$

Proof. Notice that $a_0 \cdot b_0 = 1 \cdot 1 = 1 = e_0$. Also, for any $n \in \mathbb{N}^+$, we have

$$\begin{aligned} \sum_{k=0}^{n} a_k b_{n-k} &= a_0 b_n + a_1 b_{n-1} && \text{(since } a_k = 0 \text{ if } k \geq 2\text{)} \\ &= 1 \cdot 1 + (-1) \cdot 1 \\ &= 1 - 1 \\ &= 0 \\ &= e_n \end{aligned}$$

Therefore $\{a_n\} \cdot \{b_n\} = \{e_n\}$. The proof that $\{b_n\} \cdot \{a_n\} = \{e_n\}$ is similar. □

Matrix Rings

Definition 10.3.8. *Let R be a ring and let $n \in \mathbb{N}^+$. We let $M_n(R)$ be the ring of all $n \times n$ matrices with entries in R with the following operations. Writing an element of R as $[a_{ij}]$, we define*

$$[a_{ij}] + [b_{ij}] = [a_{ij} + b_{ij}] \qquad \text{and} \qquad [a_{ij}] \cdot [b_{ij}] = [\sum_{k=1}^{n} a_{ik} b_{kj}]$$

With these operations, $M_n(R)$ is a ring with additive identity the matrix of all zeros and multiplicative identity the matrix with all zeros except for ones on the diagonal, i.e.

$$e_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The verification of the ring axioms on $M_n(R)$ are mostly straightforward (you probably did them in the case where $R = \mathbb{R}$ in linear algebra). The hardest check once again is that $\cdot$ is associative. We formally carry that out now. Given matrices $[a_{ij}]$, $[b_{ij}]$, and $[c_{ij}]$ in $M_n(R)$, we have

$$\begin{aligned}
[a_{ij}] \cdot ([b_{ij}] \cdot [c_{ij}]) &= [a_{ij}] \cdot [\sum_{\ell=1}^{n} b_{i\ell} c_{\ell j}] \\
&= [\sum_{k=1}^{n} a_{ik} \cdot (\sum_{\ell=1}^{n} b_{k\ell} c_{\ell j})] \\
&= [\sum_{k=1}^{n} \sum_{\ell=1}^{n} a_{ik} (b_{k\ell} c_{\ell j})] \\
&= [\sum_{k=1}^{n} \sum_{\ell=1}^{n} (a_{ik} b_{k\ell}) c_{\ell j}] \\
&= [\sum_{\ell=1}^{n} \sum_{k=1}^{n} (a_{ik} b_{k\ell}) c_{\ell j}] \\
&= [\sum_{\ell=1}^{n} (\sum_{k=1}^{n} a_{ik} b_{k\ell}) \cdot c_{\ell j}] \\
&= [\sum_{k=1}^{n} a_{ik} b_{kj}] \cdot [c_{ij}] \\
&= ([a_{ij}] \cdot [b_{ij}]) \cdot [c_{ij}]
\end{aligned}$$

One nice thing about this more general construction of matrix rings is that it supplies us with a decent supply of noncommutative rings.

Proposition 10.3.9. *Let R be a ring with $1 \neq 0$. For each $n \geq 2$, the ring $M_n(R)$ is noncommutative.*

Proof. Check that the matrix of all zeros except for a 1 in the $(1, 2)$ position does not commute with the matrix of all zeroes except for a 1 in the $(2, 1)$ position. □

In particular, this construction gives us examples of finite noncommutative rings. For example, the ring $M_2(\mathbb{Z}/2\mathbb{Z})$ is a noncommutative ring with $2^4 = 16$ elements.

10.4 Ideals, Quotients, and Homomorphisms

Ideals

Suppose that R is a ring and that I is a subset of R which is an additive subgroup. We know that when we look only at addition, we have that I breaks up R into (additive) cosets of the form

$$r + I = \{r + a : a \in I\}$$

These cosets are the equivalence classes of the equivalence relation $\sim_I$ on R defined by $r \sim_I s$ if there exists $a \in I$ with $r + a = s$. Since $(R, +, 0)$ is abelian, we know that the additive subgroup I of R is normal in R, so we can take the quotient R/I as additive groups. In this quotient, we know from our general theory of quotient groups that addition is well-defined by:

$$(r + I) + (s + I) = (r + s) + I$$

Now if we want to turn the resulting quotient into a ring (rather than just an abelian group), we would certainly require that multiplication of cosets is well-defined as well. In other words, we would need to know that if $r, s, t, u \in R$ with $r \sim_I t$ and $s \sim_I u$, then $rs \sim_I tu$. A first guess might be that we will should require that I is closed under multiplication as well. Before jumping to conclusions, let's work out whether this happens for free, or if it looks grim, then what additional conditions on I might we want to require.

Suppose then that $r, s, t, u \in R$ with $r \sim_I t$ and $s \sim_I u$. Fix $a, b \in I$ with $r + a = t$ and $s + b = u$. We then have

$$tu = (r + a)(s + b) = r(s + b) + a(s + b) = rs + rb + as + ab$$

Now in order for $rs \sim_I tu$, we would want $rb + as + ab \in I$. In order to ensure this, it would suffice to know that $rb \in I$, $as \in I$, and $ab \in I$ because we are assuming that I is an additive subgroup. The last of these, that $ab \in I$, would follow if add the additional assumption that I is closed under multiplication as we guessed above. However, a glance at the other two suggests that we might need to require more. These other summands suggest that we want I to be closed under "super multiplication", i.e. that if we take an element of I and multiply it by any element of R on either side, then we stay in I. If we have these conditions on I (which at this point looks like an awful lot to ask), then everything should work out fine. We give a special name to subsets of R which have this property.

Definition 10.4.1. *Let R be a ring. An ideal of R is a subset $I \subseteq R$ such that*

- *I is an additive subgroup of R.*
- *$ra \in I$ whenever $r \in R$ and $a \in I$.*
- *$ar \in I$ whenever $r \in R$ and $a \in I$.*

For example, consider the ring $R = \mathbb{Z}$. Suppose $n \in \mathbb{N}$ and let $I = n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$. We then have that I is an ideal of R. To see this, first notice that we already know that $n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Now for any $m \in \mathbb{Z}$ and $k \in \mathbb{Z}$, we have

$$m \cdot (nk) = n \cdot (mk) \in n\mathbb{Z} \qquad \text{and} \qquad (nk) \cdot m = n \cdot (km) \in n\mathbb{Z}$$

Therefore, $I = n\mathbb{Z}$ does satisfy the additional conditions necessary, so $I = n\mathbb{Z}$ is an ideal of R. Before going further, let's note one small simplification in the definition of an ideal.

Proposition 10.4.2. *Let R be a ring. Suppose that $I \subseteq R$ satisfies*

- *$0 \in I$.*
- *$a + b \in I$ whenever $a \in I$ and $b \in I$.*
- *$ra \in I$ whenever $r \in R$ and $a \in I$.*
- *$ar \in I$ whenever $r \in R$ and $a \in I$.*

We then have that I is an ideal of R.

Proof. The only condition that is missing is that I is closed under additive inverses. For any $a \in R$, we have $-a = (-1) \cdot a$, so $-a \in R$ by the third condition (notice that we are using the fact that all our rings have a multiplicative identity). □

For another example of an ideal, consider the ring $R = \mathbb{Z}[x]$. Let I be set of all polynomials with 0 constant term. Formally, we are letting I be the set of infinite sequences $\{a_n\}$ with $a_0 = 0$. Let's prove that I is an ideal using a more informal approach to the polynomial ring (make sure you know how to translate everything we are saying into formal terms). Notice that the zero polynomial is trivially in I and that I is closed under addition because the constant term of the sum of two polynomials is the sum of their constant terms. Finally, if $f(x) \in R$ and $p(x) \in I$, then we can write

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

and

$$p(x) = b_m x^n + b_{n-1}x^{n-1} + \cdots + b_1 x$$

Multiplying out the polynomials, we see that the constant term of $f(x)p(x)$ is $a_0 \cdot 0 = 0$ and similarly the constant term of $p(x)f(x)$ is $0 \cdot a_0 = 0$. Therefore, we have both $f(x)p(x) \in I$ and $p(x)f(x) \in I$. It follows that I is an ideal of R.

From the above discussion, it appears that if I is an ideal of R, then it makes sense to take the quotient of R by I and have both addition and multiplication of cosets be well-defined. However, our definition was motivated by what conditions would make checking that multiplication is well-defined easy. As in our definition of a normal subgroup, it is a pleasant surprise that the implication can be reversed so our conditions are precisely what is needed for the quotient to make sense.

Proposition 10.4.3. *Let R be a ring and let $I \subseteq R$ be an additive subgroup of $(R, +, 0)$. The following are equivalent.*

- *I is an ideal of R.*
- *Whenever $r, s, t, u \in R$ with $r \sim_I t$ and $s \sim_I u$, we have $rs \sim_I tu$.*

Proof. We first prove that 1 implies 2. $r, s, t, u \in R$ with $r \sim_I t$ and $s \sim_I u$. Fix $a, b \in I$ with $r + a = t$ and $s + b = u$. We then have

$$tu = (r + a)(s + b) = r(s + b) + a(s + b) = rs + rb + as + ab$$

Since I is an ideal of R and $b \in I$, we know that both $rb \in I$ and $ab \in I$. Similarly, since I is an ideal of R and $a \in I$, we know that $as \in I$. Now I is an ideal of R, so it is an additive subgroup of R, and hence $rb + as + ab \in I$. Since

$$tu = rs + (rb + as + ab)$$

we conclude that $rs \sim_I tu$.

We now prove that 2 implies 1. We are assuming that I is an additive subgroup of R and condition 2. Let $r \in R$ and let $a \in I$. Now $r \sim_I r$ because $r + 0 = r$ and $0 \in I$. Also, we have $a \sim_I 0$ because $a + (-a) = 0$ and $-a \in I$ (as $a \in I$ and I is an additive subgroup).

- Since $r \sim_I r$ and $a \sim_I 0$, we may use condition 2 to conclude that $ra \sim_I r0$, which is to say that $ra \sim_I 0$. Thus, we may fix $b \in I$ with $ra + b = 0$. Since $b \in I$ and I is an additive subgroup, it follows that $ra = -b \in I$.
- Since $a \sim_I 0$ and $r \sim_I r$, we may use condition 2 to conclude that $ar \sim_I 0r$, which is to say that $ar \sim_I 0$. Thus, we may fix $b \in I$ with $ar + b = 0$. Since $b \in I$ and I is an additive subgroup, it follows that $ar = -b \in I$.

Therefore, we have both $ra \in I$ and $ar \in I$. Since $r \in R$ and $a \in I$ were arbitrary, we conclude that I is an ideal of R. □

We are now ready to formally define quotient rings.

Definition 10.4.4. *Let R be a ring and let I be an ideal of R. Let R/I be the set of additive cosets of I in R, i.e. the set of equivalence classes of R under $\sim_I$. Define operation on R/I by letting*

$$(a+I)+(b+I)=(a+b)+I \qquad (a+I)\cdot(b+I)=ab+I$$

With these operations, the set R/I becomes a ring with additive identity $0+I$ and multiplicative identity $1+I$ (we did the hard part of checking that the operations are well-defined, and from here the ring axioms follow in a completely straightforward manner by pushing them all up to R). Furthermore, if R is commutative, then R/I is also commutative.

Let $n \in \mathbb{N}^+$. As discussed above, the set $n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$ is an ideal of $\mathbb{Z}$. Our familiar ring $\mathbb{Z}/n\mathbb{Z}$ defined in the first section is precisely the quotient of $\mathbb{Z}$ by this ideal $n\mathbb{Z}$, hence the notation again.

As we have seen, ideals in rings correspond exactly to normal subgroups of a group in that they are the "special" subsets for which it makes sense to take a quotient. There is one small but important note to be made here. Of course, every normal subgroup of a group G is a subgroup of G. However, it is *not* true that every ideal of a ring R is a subring of R. The reason is that for I to be an ideal of R, it is *not* required that $1 \in I$. In fact, if I is an ideal of R and $1 \in R$, then $r = r \cdot 1 \in I$ for all $r \in R$, so $I = R$. Since every subring S of R must satisfy $1 \in S$, it follows that the only ideal of R which is a subring of R is the whole ring itself. This might seem to be quite a nuisance, but as mentioned above, we will pay very little attention to subrings of a given ring, and the vast majority of our focus will be on ideals.

We end with our discussion of ideals in general rings with a simple characterization for when two elements of a ring R represent the same coset.

Proposition 10.4.5. *Let R be a ring and let I be an ideal of R. Let $r, s \in R$. The following are equivalent.*

1. $r + I = s + I$
2. $r \sim_I s$
3. $r - s \in I$
4. $s - r \in I$

Proof. Notice that 1 and 2 are equivalent from our general theory of equivalence relations because $r + I$ is simply the equivalence class of r under the relation $\sim_I$.

2 implies 3: Suppose that $r \sim_I s$, and fix $a \in I$ with $r + a = s$. Subtracting s and a from both sides, it follows that $r - s = -a$ (you should work through the details of this if you are nervous). Now $a \in I$ and I is an additive subgroup of R, so $-a \in I$. It follows that $r - s \in I$.

3 implies 4: Suppose that $r - s \in I$. Since I is an additive subgroup of R, we know that $-(r - s) \in I$. Since $-(r - s) = s - r$, it follows that $s - r \in I$.

4 implies 2: Suppose that $s - r \in I$. Since $r + (s - r) = s$, it follows that $r \sim_I s$. □

Ring Homomorphisms

Definition 10.4.6. *Let R and S be rings. A (ring) homomorphism from R to S is a function $\varphi\colon R \to S$ such that*

- *$\varphi(r + s) = \varphi(r) + \varphi(s)$ for all $r, s \in R$.*
- *$\varphi(r \cdot s) = \varphi(r) \cdot \varphi(s)$ for all $r, s \in R$.*

- $\varphi(1_R) = 1_S$.

A (ring) isomorphism from R to S is a homomorphism $\varphi\colon R \to S$ which is a bijection.

Definition 10.4.7. *Given two rings R and S, we say that R and* isomorphic*, and write $R \cong S$, if there exists an isomorphism $\varphi\colon R \to S$.*

Notice that we have the additional requirement that $\varphi(1_R) = 1_S$. When we discussed group homomorphisms, we derived $\varphi(e_G) = e_H$ rather than explicitly require it. Unfortunately, it does not follow for free from the other two conditions in the ring case. If you go back and look at the proof that $\varphi(e_G) = e_H$ in the group case, you will see that we used the fact that $\varphi(e_G)$ has an inverse but it is not true in rings that every element must have a multiplicative inverse. To see an example where the condition can fail consider the ring $\mathbb{Z} \times \mathbb{Z}$ (we have not formally defined the direct product of rings, but it works in the same way). Define $\varphi\colon \mathbb{Z} \to \mathbb{Z} \times \mathbb{Z}$ by $\varphi(n) = (n, 0)$. It is not hard to check that φ satisfies the first two conditions for a ring homomorphism, but $\varphi(1) = (1, 0)$ and the identity of $\mathbb{Z} \times \mathbb{Z}$ is $(1, 1)$.

Definition 10.4.8. *Let R be a ring and let $c \in R$. Define $Ev_c\colon R[x] \to R$ by letting*

$$Ev_c(\{a_n\}) = \sum_n a_n c^n$$

Notice that the above sum makes because elements $\{a_n\} \in R[x]$ have only finitely many nonzero terms (if $\{a_n\}$ is nonzero, we can stop the sum at $M = \deg(\{a_n\})$). Intuitively, we are defining

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

Thus, Ev_c is the function which says "evaluate the polynomial at c".

The next proposition is really fundamental. Intuitively, it says the following. Suppose that you are given a *commutative* ring R, and you take two polynomials in R. Given any $c \in R$, you get the same thing if add (multiply) the polynomials first and plug c into the result as you would if you plug c into each polynomial and add (multiply) the results.

Proposition 10.4.9. *Let R be a commutative ring and let $c \in R$. The function Ev_c is a ring homomorphism.*

Proof. We clearly have $Ev_c(1) = 1$ (where the 1 in parenthesis is the constant polynomial 1). For any $\{a_n\}, \{b_n\} \in R[x]$, we have

$$\begin{aligned} Ev_c(\{a_n\} + \{b_n\}) &= Ev_c(\{a_n + b_n\}) \\ &= \sum_n (a_n + b_n)c^n \\ &= \sum_n (a_n c^n + b_n c^n) \\ &= \sum_n a_n c^n + \sum_n b_n c^n \\ &= Ev_c(\{a_n\}) + Ev_c(\{b_n\}) \end{aligned}$$

Thus, Ev_c preserves addition. We now check that Ev_c preserves multiplication. Suppose that $\{a_n\}, \{b_n\} \in R[x]$. Let $M_1 = \deg(\{a_n\})$ and $M_2 = \deg(\{b_n\})$ (for this argument, let $M_i = 0$ if the corresponding

polynomial is the zero polynomial). We know that $M_1 + M_2 \geq \deg(\{a_n\} \cdot \{b_n\})$, so

$$\begin{aligned}
Ev_c(\{a_n\} \cdot \{b_n\}) &= Ev_c(\{ \sum_{i+j=n} a_i b_j\}) \\
&= \sum_{n=0}^{M_1+M_2} (\sum_{i+j=n} a_i b_j) c^n \\
&= \sum_{n=0}^{M_1+M_2} \sum_{i+j=n} a_i b_j c^n \\
&= \sum_{n=0}^{M_1+M_2} \sum_{i+j=n} a_i b_j c^{i+j} \\
&= \sum_{n=0}^{M_1+M_2} \sum_{i+j=n} a_i b_j c^i c^j \\
&= \sum_{n=0}^{M_1+M_2} \sum_{i+j=n} a_i c^i b_j c^j && \text{(since } R \text{ is commutative)} \\
&= \sum_{i=0}^{M_1} \sum_{j=0}^{M_2} a_i c^i b_j c^j && \text{(since } a_i = 0 \text{ if } i > M_1 \text{ and } b_j = 0 \text{ if } j > M_2) \\
&= \sum_{i=0}^{M_1} (a_i c^i \cdot (\sum_{j=0}^{M_2} b_j c^j)) \\
&= (\sum_{i=0}^{M_1} a_i c^i) \cdot (\sum_{j=0}^{M_2} b_j c^j) \\
&= Ev_c(\{a_n\}) \cdot Ev_c(\{b_n\})
\end{aligned}$$

Thus, Ev_c preserves multiplication. It follows that Ev_c is a ring homomorphism. □

Let's take a look at what can fail when R is noncommutative. Consider the polynomials $p(x) = ax$ and $q(x) = bx$. We then have that $p(x)q(x) = abx^2$. Now if $c \in R$, then

$$Ev_c(p(x)q(x)) = Ev_c(abx^2) = abc^2$$

and

$$Ev_c(p(x)) \cdot Ev_c(q(x)) = acbc$$

It seems impossible to argue that these are equal in general if you can not commute b with c. To find a specific counterexample, if suffices to find a noncommutative ring with two elements b and c such that $bc^2 \neq cbc$ (because then you can take $a = 1$). It's not hard to find two matrices which satisfy this, so the corresponding Ev_c will not be a ring homomorphism.

In the future, given $p(x) \in R[x]$ and $c \in R$, we will tend to write the informal $p(c)$ for the formal notion $Ev_c(p(x))$. In this informal notation, the proposition says that if R is commutative, $c \in R$, and $p(x), q(x) \in R[x]$, then

$$(p + q)(c) = p(c) + q(c) \qquad \text{and} \qquad (p \cdot q)(c) = p(c) \cdot q(c)$$

Definition 10.4.10. *Let $\varphi\colon R \to S$ be a ring homomorphism. We define $ker(\varphi)$ to be the kernel of φ when viewed as a homomorphism of the additive groups, i.e. $ker(\varphi) = \{a \in R : \varphi(R) = 0_S\}$.*

Recall that the normal subgroups of a group G are precisely the kernels of group homomorphisms with domain G. Continuing on the analogy between normal subgroups of a group and ideal of a ring, we might hope that the ideals of a ring R are precisely the kernels of ring homomorphisms with domain R. The next propositions confirm this.

Proposition 10.4.11. *If $\varphi\colon R \to S$ is a ring homomorphism, then ker(φ) is an ideal of R*

Proof. Let $K = \ker(\varphi)$. Since we know in particular in that φ is a additive group homomorphism, we know from our work on groups that K is an additive subgroup of R. Suppose that $a \in K$ and $r \in R$. Since $a \in K$, we have $\varphi(a) = 0$. Therefore

$$\varphi(ra) = \varphi(r) \cdot \varphi(a) = \varphi(r) \cdot 0_S = 0_S$$

so $ra \in K$ and

$$\varphi(ar) = \varphi(a) \cdot \varphi(r) = 0_S \cdot \varphi(r) = 0_S$$

so $ar \in K$. Therefore, K is an ideal of R. □

Proposition 10.4.12. *Let R be a ring and let I be an ideal of R. There exists a ring S and a ring homomorphism $\varphi\colon R \to S$ such that $I = ker(\varphi)$.*

Proof. Consider the ring $S = R/I$ and the projection $\pi\colon R \to R/I$ defined by $\pi(a) = a + I$. As in the group case, it follows that π is a ring homomorphism and that $\ker(\pi) = I$. □

Now many of the results about group homomorphism carry over to ring homomorphism. For example, we have the following.

Proposition 10.4.13. *Let $\varphi\colon R \to S$ be a ring homomorphism. φ is injective if and only if* $\ker(\varphi) = \{0_R\}$.

Proof. Notice that φ is in particular a homomorphism of additive groups. Thus, the result follows from the corresponding result about groups. However, let's prove it again because it is an important result.

Suppose first that φ is injective. We know that $\varphi(0_R) = 0_S$, so $0_R \in \ker(\varphi)$. If $a \in \ker(\varphi)$, then $\varphi(a) = 0_S = \varphi(0_R)$, so $a = 0_R$ because φ is injective. Therefore, $\ker(\varphi) = \{0_R\}$.

Suppose conversely that $\ker(\varphi) = \{0_R\}$. Let $r, s \in R$ with $\varphi(r) = \varphi(s)$. We then have

$$\begin{aligned}\varphi(r - s) &= \varphi(r + (-s)) \\ &= \varphi(r) + \varphi(-s) \\ &= \varphi(r) - \varphi(s) \\ &= 0_S\end{aligned}$$

so $r - s \in \ker(\varphi)$. We are assuming that $\ker(\varphi) = \{0_R\}$, so $r - s = 0_R$ and hence $r = s$. □

Next we discuss the ring theoretic analogues of Proposition 8.2.8. If you replace "subgroup" by "subring", then everything works out as you would expect. There is a little more work here than just directly appealing to the group-theoretic fact because you need to check that the corresponding object is closed under multiplication. However, we will omit the details.

Proposition 10.4.14. *Let $\varphi\colon R_1 \to R_2$ be a homomorphism. We have the following*

1. *For all subrings S_1 of R_1, we have $\varphi(S_1) = \{\varphi(a) : a \in S_1\}$ is a subring of R_2. In particular, range(φ) $= \varphi(R_1)$ is a subring of R_2.*
2. *For all subgroups S_2 of R_2, we have $\varphi^{-1}(S_2) = \{a \in R_1 : \varphi(a) \in S_2\}$ is a subring of R_1.*

Since ideals are *not* subrings in general, you might wonder if the above result works if you replace that S_i's by ideals. The second statement is true for ideals, but the first may fail. Intuitively, if you push an ideal across a homomorphism and there are are elements of R_2 which are not hit by φ, then there is no reason to believe that the resulting subset of R_2 will be closed under multiplication from that element. However, we have the following.

Proposition 10.4.15. *Let $\varphi\colon R_1 \to R_2$ be a homomorphism. We have the following*

1. *Suppose that φ is surjective. For all ideals I_1 of R_1, we have $\varphi(I_1) = \{\varphi(a) : a \in I_1\}$ is an ideal of R_2*
2. *For all ideals I_2 of R_2, we have $\varphi^{-1}(I_2) = \{a \in R_1 : \varphi(a) \in I_2\}$ is an ideal of R_1.*

We end with brief statements of the Isomorphism and Correspondence Theorems. As you would expect, the corresponding results from group theory do a lot of heavy lifting in the following proofs, but you still need to check a few extra things in each case (like the corresponding functions preserve multiplication).

Theorem 10.4.16 (First Isomorphism Theorem)**.** *Let $\varphi\colon R \to S$ be a ring homomorphism and let $K = ker(\varphi)$. Define a function $\psi\colon R/K \to S$ by letting $\psi(a + K) = \varphi(a)$. We then have that ψ is a well-defined function which is a ring isomorphism onto the subring range(φ) of S. Therefore*

$$R/ker(\varphi) \cong range(\varphi)$$

Let's see the First Isomorphism Theorem in action. Let $R = \mathbb{Z}[x]$ and let I be the ideal of all polynomials with 0 constant term. Consider the ring R/I. Intuitively, taking the quotient by I you "kill off" all polynomials without a constant term. Thus, intuitively, two polynomials will be in the same coset in the quotient when they have the same constant term (because then the difference of the two polynomials will be in I). Thus, you might expect that $R/I \cong \mathbb{Z}$. We can see this formally by appealing to the first Isomorphism Theorem. Recall that $Ev_0\colon \mathbb{Z}[x] \to \mathbb{Z}$ is a ring homomorphism from our previous work (because $\mathbb{Z}$ is commutative). Notice that Ev_0 is surjective because $Ev_0(a) = a$ (where you interpret the a in parentheses as the constant polynomial a) and that $\ker(Ev_0) = I$. Therefore, we conclude that $R/I \cong \mathbb{Z}$ using the First Isomorphism Theorem.

We state the remaining theorems without comment.

Theorem 10.4.17 (Second Isomorphism Theorem)**.** *Let R be a ring, let S be a subring of R, and let I be an ideal of R. We then have that $S + I = \{r + a : r \in S, a \in I\}$ is a subring of R, that I is an ideal of $S + I$, that $S \cap I$ is an ideal of S, and that*

$$\frac{S+I}{I} \cong \frac{S}{S \cap I}$$

Theorem 10.4.18 (Correspondence Theorem)**.** *Let R be a ring and let I be an ideal of R. For every subring S of R with $I \subseteq S$, we have that S/I is a subring of R/I and the function*

$$S \mapsto S/I$$

is a bijection from subrings of R containing I to subrings of R/I. Also, for every ideal J of R with $I \subseteq J$, we have that J/I is an ideal of R/I and the function

$$J \mapsto J/I$$

is a bijection from ideals of R containing J to ideals of R/J.

Theorem 10.4.19 (Third Isomorphism Theorem)**.** *Let R be a ring. Let I and J be ideals of R with $I \subseteq J$. We then have that J/I is an ideal of R/I and that*

$$\frac{R/I}{J/I} \cong \frac{R}{J}$$

10.5 Ideals in Commutative Rings

Throughout this section, we work in the special case when R is a commutative ring. We begin by noting that we can make one other tiny simplification in the definition of an ideal.

Proposition 10.5.1. *Let R be a commutative ring. Suppose that $I \subseteq R$ satisfies*

- $0 \in I$.
- $a + b \in I$ *whenever* $a \in I$ *and* $b \in I$.
- $ra \in I$ *whenever* $r \in R$ *and* $a \in I$.

We then have that I is an ideal of R.

Proof. Given $r \in R$ and $a \in I$, we have $ar = ra \in I$ because R is commutative. The result follows from Proposition 10.4.2. □

We next give the ideal-theoretic analogue of cyclic subgroups of a given group.

Definition 10.5.2. *Let R be a commutative ring and let $a \in R$. We define $\langle a \rangle = \{ra : r \in R\}$. The ideal $\langle a \rangle$ is called the* ideal generated by *a.*

Proposition 10.5.3. *Let R be a ring and let $a \in R$. Let $I = \langle a \rangle$.*

1. *I is an ideal of R with $a \in I$.*
2. *If J is an ideal of R with $a \in J$, then $I \subseteq J$.*

Proof. We first prove 1. We begin by noting that $a = 1 \cdot a \in I$ and that $0 = 0 \cdot a \in I$. For any $r, s \in R$, we have

$$ra + sa = (r + s)a \in I$$

so I is closed under addition. For any $r, s \in R$, we ahve

$$r \cdot (sa) = (rs) \cdot a \in I$$

Therefore I is an ideal of R with $a \in I$ by Proposition 10.5.1

We now prove 2. Let J be an ideal of R with $a \in J$. For any $r \in R$, we have $ra \in J$ because J is a ideal of R and $a \in J$. It follows that $I \subseteq J$. □

Definition 10.5.4. *Let R be a ring and let I be an ideal of R. We say that I is a* principal ideal *if there exists $a \in R$ with $I = \langle a \rangle$.*

Proposition 10.5.5. *The ideals of $\mathbb{Z}$ are precisely the sets $n\mathbb{Z} = \langle n \rangle$ for every $n \in \mathbb{N}$. Thus, every ideal of $\mathbb{Z}$ is principal.*

Proof. Let I be an ideal of R. We know that $0 \in I$. If $I = \{0\}$, then $I = \langle 0 \rangle$, and we are done. Suppose then that $I \neq \{0\}$. Notice that if $k \in I$, then $-k \in I$ as well, hence $I \cap \mathbb{N}^+ \neq \emptyset$. By well-ordering, we may let $n = \min(I \cap \mathbb{N}^+)$. We claim that $I = \langle n \rangle$.

First notice that since $n \in I$ and I is an ideal of R, it follows that $\langle n \rangle \subseteq I$. Suppose now that $k \in I$. Fix $q, r \in \mathbb{Z}$ with $k = qn + r$ and $0 \leq r < n$. We then have that $r = k - qn = k + (-q)n$. Since $k, n \in I$ and I is an ideal, we know that $(-q)n \in I$ and so $r = k + (-q)n \in I$. Now $0 \leq r < n$ and $n = \min(I \cap \mathbb{N}^+)$, so we must have $r = 0$. It follows that $k = qn$, so $k \in \langle n \rangle$. Now $k \in I$ was arbitrary, so $I \subseteq \langle n \rangle$. Putting this together with the above, we conclude that $I = \langle n \rangle$. □

Lemma 10.5.6. *Let R be a commutative ring and let I be an ideal of R. We have that $I = R$ if and only if I contains a unit of R.*

Proof. If $I = R$, then $1 \in I$, so I contains a unit. Suppose conversely that I contains a unit, and fix such a unit $u \in I$. Since u is a unit, we may fix $v \in R$ with $vu = 1$. Since $u \in I$ and I is an ideal, we conclude that $1 \in I$. Now for any $r \in R$, we have $r = r \cdot 1 \in I$ again because I is an ideal of R. Thus, $R \subseteq I$, and since $I \subseteq R$ trivially, it follows that $I = R$. □

Proposition 10.5.7. *Let R be a commutative ring. The following are equivalent.*

1. *R is a field.*
2. *The only ideals of R are $\{0\}$ and R.*

Proof. We first prove that 1 implies 2. Suppose that R is field. Let I be an ideal of R with $I \neq \{0\}$. Fix $a \in I$ with $a \neq 0$. Since R is a field, every nonzero element of R is a unit, so a is a unit. Since $a \in I$, we may use the lemma to conclude that $I = R$.

We now prove that 2 implies 1 by proving the contrapositive. Suppose that R is not a field. Fix a nonzero element $a \in R$ such that a is not a unit. Let $I = \langle a \rangle = \{ra : r \in R\}$. We know from above that I is an ideal of R. If $1 \in I$, then we may fix $r \in R$ with $ra = 1$, which implies that a is a unit (remember that we are assuming that R is commutative). Therefore, $1 \notin I$, and hence $I \neq R$. Since $a \neq 0$ and $a \in I$, we have $I = \{0\}$. Therefore, I is an ideal of R distinct from $\{0\}$ and R. □

We now define two very special kinds of ideals in commutative rings.

Definition 10.5.8. *Let R be a commutative ring. A* prime ideal *of R is an ideal $P \subseteq R$ such that $P \neq R$ and whenever $ab \in P$, then either $a \in P$ or $b \in P$.*

For example, $\{0\}$ is a prime ideal of $\mathbb{Z}$ (because $\mathbb{Z}$ is an integral domain), and $p\mathbb{Z}$ is a prime ideal of $\mathbb{Z}$ for every prime number $p \in \mathbb{N}^+$. To see this latter fact, fix a prime number $p \in \mathbb{N}^+$. Notice that

$$a \in p\mathbb{Z} \iff \text{There exists k} \in \mathbb{Z} \text{ with } a = pk \iff p \mid a$$

In particular, we have $1 \notin p\mathbb{Z}$, so $p\mathbb{Z} \neq \mathbb{Z}$. Suppose now that $a, b \in \mathbb{Z}$ with $ab \in p\mathbb{Z}$. From above, we have $p \mid ab$, so as p prime, either $p \mid a$ or $p \mid b$. In the former case, we have $a \in p\mathbb{Z}$ while in the latter we have $b \in p\mathbb{Z}$. Therefore, $p\mathbb{Z}$ is a prime ideal of $\mathbb{Z}$ for every prime $p \in \mathbb{N}^+$.

In contrast, $6\mathbb{Z}$ is not a prime ideal of $\mathbb{Z}$ because $2 \cdot 3 \in 6\mathbb{Z}$ but $3 \notin 6\mathbb{Z}$ and $2 \notin 3\mathbb{Z}$. A similar argument shows that $n\mathbb{Z}$ is not a prime ideal of $\mathbb{Z}$ whenever $n \in \mathbb{N}^+$ is composite.

Definition 10.5.9. *Let R be a commutative ring. A* maximal ideal *of R is an ideal $M \subseteq R$ such that $M \neq R$ and there exists no ideal I of R with $M \subsetneq I \subsetneq R$.*

Theorem 10.5.10. *Let R be a commutative ring and let P be an ideal of R. P is a prime ideal of R if and only if R/P is an integral domain.*

Proof. Suppose first that P is a prime ideal of R. Since R is commutative, we know that R/P is commutative. By definition, we have $P \neq R$, so $1 \notin P$, and hence $1 + P \neq 0 + P$. Finally, suppose that $a, b \in R$ with $(a + P) \cdot (b + P) = 0 + P$. We then have that $ab + P = 0 + P$, so $ab \in P$. Since P is a prime ideal, either $a \in P$ or $b \in P$. Therefore, either $a + P = 0 + P$ or $b + P = 0 + P$. It follows that R/P is an integral domain.

Suppose conversely that R/P is an integral domain. We then have that $1 + P \neq 0 + P$ by definition of an integral domain, hence $1 \notin P$ and so $P \neq R$. Suppose that $a, b \in R$ with $ab \in P$. We then have

$$(a + P) \cdot (b + P) = ab + P = 0 + P$$

Since R/P is an integral domain, we conclude that either $a + P = 0 + P$ or $b + P = 0 + P$. Therefore, either $a \in P$ or $b \in P$. It follows that P is a prime ideal of R. □

Theorem 10.5.11. *Let R be a commutative ring and let M be an ideal of R. M is a maximal ideal of R if and only if R/M is a field.*

Proof. We give two proofs. The first is the slick "highbrow" proof. Using the Correspondence Theorem and Proposition 10.5.7, we have

$$\begin{aligned} M \text{ is a maximal ideal of } R &\iff \text{There are no ideals } I \text{ of } R \text{ with } M \subsetneq I \subseteq R \\ &\iff \text{There are no ideals of } R/M \text{ other than } \{0+I\} \text{ and } R/M \\ &\iff R/M \text{ is a field} \end{aligned}$$

If you don't like appealing to the Correspondence Theorem (which is a shame, because it's awesome), we can prove it directly via a "lowbrow" proof.

Suppose first that M is a maximal ideal of R. Fix a nonzero element $a+M \in R/M$. Since $a+M \neq 0+M$, we have that $a \notin M$. Let $I = \{sa+m : s \in R, m \in M\}$. We then have that I is an ideal of M (check it) with $M \subseteq I$ and $a \in I$. Since $a \notin M$, we have $M \subsetneq I$, so as M is maximal it follows that $I = R$. Thus, we may fix $S \in R$ and $m \in M$ with $sa+m = 1$. We then have $sa - 1 = -m \in M$, so $sa + M = 1 + M$. It follows that $(s+M)(a+M) = 1+M$, so $a+M$ has an inverse in R/M (recall that R and hence R/M is commutative, so we only need an inverse on one side).

Suppose conversely that R/M is a field. Since R/M is a field, we have $1+M \neq 0+M$, so $1 \notin M$ and hence $M \subsetneq R$. Fix an ideal I of R with $M \subsetneq I$. Since $M \subsetneq I$, we may fix $a \in I \backslash M$. Since $a \notin M$, we have $a+M \neq 0+M$, and using the fact that R/M is a field we may fix $b+M \in R/M$ with $(a+M)(b+M) = 1+M$. We then have $ab+M = 1+M$, so $ab-1 \in M$. Fixing $m \in M$ with $ab-1 = m$, we then have $ab - m = 1$. Now $a \in I$, so $ab \in I$ as I is an ideal. Also, we have $m \in M \subseteq I$. It follows that $1 = ab - m \in I$, and thus $I = R$. Therefore, M is a maximal ideal of R. □

From the theorem, we see that $p\mathbb{Z}$ is a maximal ideal of $\mathbb{Z}$ for every prime $p \in \mathbb{N}^+$ because we know that $\mathbb{Z}/p\mathbb{Z}$ is a field (you could have proven this directly as well). We also get the following nice corollary.

Corollary 10.5.12. *Let R be a commutative ring. Every maximal ideal of R is a prime ideal of R.*

Proof. Suppose that M is a maximal ideal of R. We then have that R/M is a field by Theorem 10.5.11, so R/M is an integral domain by Proposition 10.2.8. Therefore, M is a prime ideal of R by Theorem 10.5.10. □

The converse is not true. As we've seen, the ideal $\{0\}$ is a prime ideal of $\mathbb{Z}$, but it is certainly not a maximal ideal of $\mathbb{Z}$. For later use, we now generalize the idea of a principal ideal.

Definition 10.5.13. *Let R be a commutative ring and let $a_1, a_2, \ldots, a_n \in R$. We define*

$$\langle a_1, a_2, \ldots, a_n \rangle = \{r_1a_1 + r_2a_2 + \cdots + r_na_n : r_i \in R \text{ for each } i\}$$

This set is an ideal called the ideal generated by $a_1, a_2, \ldots, a_n$.

It is not hard to check that $\langle a_1, a_2, \ldots, a_n \rangle$ is the smallest ideal containing each of the a_i.

10.6 The Characteristic of a Ring

Let R be a ring (not necessarily commutative). We know that R has an element 1, so it makes sense to consider $1+1$, $1+1+1$, etc. It is natural to call these resulting ring elements 2, 3, etc., but you need to be careful in interpreting these. For example, in the ring $\mathbb{Z}/2\mathbb{Z}$, the multiplicative identity is $\overline{1}$, and we have $\overline{1}+\overline{1} = \overline{0}$. Thus, in the above notation, we would have $2 = 0$ in the ring $\mathbb{Z}/2\mathbb{Z}$. To avoid such confusion, we introduce new notation by putting an underline beneath a number n to denote the corresponding result of adding $1 \in R$ to itself n times. Here is the formal definition.

Definition 10.6.1. *Let R be a ring. For each $n \in \mathbb{Z}$, we define an element $\underline{n}$ recursively as follows. Let*

- $\underline{0} = 0$
- $\underline{n+1} = \underline{n} + 1$ *for all* $n \in \mathbb{N}$
- $\underline{n} = -(\underline{-n})$ *if* $n \in \mathbb{Z}$ *with* $n < 0$

For example, if we unravel the above definition, we have

$$\underline{4} = 1 + 1 + 1 + 1 \qquad \underline{-3} = -\underline{3} = -(1 + 1 + 1) = (-1) + (-1) + (-1)$$

As the above example illustrate, if n is negative, the element $\underline{n}$ is just the result of adding $-1 \in R$ to itself $|n|$ many times. Let's analyze how to add and multiply elements of the form $\underline{m}$ and $\underline{n}$. Notice that we have

$$\underline{3} + \underline{4} = (1 + 1 + 1) + (1 + 1 + 1 + 1) = 1 + 1 + 1 + 1 + 1 + 1 + 1 = \underline{7}$$

and using the distributive law we have

$$\begin{aligned}
\underline{3} \cdot \underline{4} &= (1 + 1 + 1)(1 + 1 + 1 + 1) \\
&= (1 + 1 + 1) \cdot 1 + (1 + 1 + 1) \cdot 1 + (1 + 1 + 1) \cdot 1 + (1 + 1 + 1) \cdot 1 \\
&= (1 + 1 + 1) + (1 + 1 + 1) + (1 + 1 + 1) + (1 + 1 + 1) \\
&= 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 \\
&= \underline{12}
\end{aligned}$$

At least for positive m and n, these computations illustrate the following result.

Proposition 10.6.2. *Let R be a ring. For any $m, n \in \mathbb{Z}$, we have*

- $\underline{m} + \underline{n} = \underline{m + n}$
- $\underline{m \cdot n} = \underline{m} \cdot \underline{n}$

Therefore the function $\varphi \colon \mathbb{Z} \to R$ defined by $\varphi(n) = \underline{n}$ is a ring homomorphism.

Proof. As mentioned, the above examples illustrate why this should be true for positive m, n, and a little playing around with negative values will suggest it works there. A formal proof fixes $m \in \mathbb{Z}$ and begins by proving the result for all $n \in \mathbb{N}$ by induction, and then handles negative values of n from there. However, I will leave the details for you. □

Proposition 10.6.3. *Let R be a ring. For all $a \in R$ and $n \in \mathbb{Z}$, we have $\underline{n} \cdot a = a \cdot \underline{n}$.*

Proof. Let $a \in R$. Notice that

$$\underline{0} \cdot a = 0 \cdot a = 0 = a \cdot 0 = a \cdot \underline{0}$$

If $n \in \mathbb{N}^+$, we have

$$\begin{aligned}
\underline{n} \cdot a &= (1 + 1 + \cdots + 1) \cdot a \\
&= 1 \cdot a + 1 \cdot a + \cdots + 1 \cdot a \\
&= a + a + \cdots + a \\
&= a \cdot 1 + a \cdot 1 + \cdots + a \cdot 1 \\
&= a \cdot (1 + 1 + \cdots + 1) \\
&= a \cdot \underline{n}
\end{aligned}$$

where each of the above sums have n terms (if you find the ... not sufficiently formal, you can again give a formal inductive argument). Suppose now that $n \in \mathbb{Z}$ with $n < 0$. We then have $-n > 0$, hence

$$\begin{aligned} \underline{n} \cdot a &= (-(\underline{-n})) \cdot a \\ &= -((\underline{-n}) \cdot a) \\ &= -(a \cdot (\underline{-n})) && \text{(from above)} \\ &= a \cdot (-(\underline{-n})) \\ &= a \cdot \underline{n} \end{aligned}$$

□

Definition 10.6.4. *Let R be a ring. We define the* characteristic *of R, denoted $char(R)$, as follows. If there exists $n \in \mathbb{N}^+$ with $\underline{n} = 0$, we define $char(R)$ to be the least such n. If no such n exists, we define $char(R) = 0$.*

Notice that $char(R)$ is the order of 1 when viewed as an element of the abelian group $(R, +, 0)$ unless that order is infinite (in which case we defined $char(R) = 0$). For example, $char(\mathbb{Z}/n\mathbb{Z}) = n$ for all $n \in \mathbb{N}^+$ and $char(\mathbb{Z}) = 0$. Also, we have $char(\mathbb{Q}) = 0$ and $char(\mathbb{R}) = 0$. For an example of an infinite ring with nonzero characteristic, notice that $char(\mathbb{Z}/n\mathbb{Z}[x]) = n$ for all $n \in \mathbb{N}^+$.

Proposition 10.6.5. *Let R be a ring with $char(R) = n$ and let $\varphi\colon \mathbb{Z} \to R$ be the ring homomorphism $\varphi(m) = \underline{m}$. Let S be the subgroup of $(R, +, 0)$ generated by 1. We then have that $S = range(\varphi)$ is a subring of R. Furthermore:*

- *If $n \neq 0$, then $\ker(\varphi) = n\mathbb{Z}$ and so by the First Isomorphism Theorem it follows that $\mathbb{Z}/n\mathbb{Z} \cong S$.*
- *If $n = 0$, then $\ker(\varphi) = \{0\}$, so φ is injective and hence $\mathbb{Z} \cong S$*

In particular, if $char(R) = n \neq 0$, then R has a subring isomorphic to $\mathbb{Z}/n\mathbb{Z}$, while if $char(R) = 0$, then R has a subring isomorphic to $\mathbb{Z}$.

Proof. The subgroup of $(R, +, 0)$ generated by 1 is

$$S = \{\underline{m} : m \in \mathbb{Z}\} = \{\varphi(m) : m \in \mathbb{Z}\} = \text{range}(\varphi)$$

Since S is the range of ring homomorphism, we have that S is a subring of R. Now if $n = 0$, then $\underline{m} \neq 0$ for all nonzero $m \in \mathbb{Z}$, so $\ker(\varphi) = \{0\}$. Suppose that $n \neq 0$. To see that $\ker(\varphi) = n\mathbb{Z}$, note that the order of 1 when viewed as an element of the abelian group $(R, +, 0)$ equals n, so $\underline{m} = 0$ if and only if $n \mid m$ by Proposition 4.2.5. □

Proposition 10.6.6. *If R is an integral domain, then either $char(R) = 0$ or $char(R)$ is prime.*

Proof. Let R be an integral domain and let $n = char(R)$. Notice that $n \neq 1$ because $1 \neq 0$ as R is an integral domain. Suppose that $n \geq 2$ and that n is composite. Fix $k, \ell \in \mathbb{N}$ with $2 \leq k, \ell < n$ and $n = k\ell$. We then have

$$0 = \underline{n} = \underline{k \cdot \ell} = \underline{k} \cdot \underline{\ell}$$

Since R is an integral domain, either $\underline{k} = 0$ or $\underline{\ell} = 0$. However, this is a contradiction because $k, \ell < n$ and $n = char(R)$ is the least positive value of m with $\underline{m} = 0$. Therefore, either $n = 0$ or n is prime. □

Chapter 11

Integral Domains

We now have several example of fields, like $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}/p\mathbb{Z}$ for primes p. By Proposition 10.2.8 each of these is also an integral domain. Our primary example of an integral domain which is not a field is $\mathbb{Z}$. We spent a lot of time developing arithmetic in $\mathbb{Z}$ early on in Section 2, and there we worked through the notions of divisibility, greatest common divisors, and prime factorizations. All of these concepts turn out to be trivial in a field (every nonzero element "divides" every other in a field because every nonzero element has an inverse), but we saw a rich and interesting theory in $\mathbb{Z}$.

We want more examples of integral domains which are not fields to see what happens in those cases. Proposition 10.3.5 lets us build new integral domain from old because it says that $R[x]$ is an integral domain whenever R is an integral domain. In particular, $F[x]$ is an integral domain whenever F is a field. As we will see, the polynomial ring $F[x]$ for a field F behaves in a great many ways like $\mathbb{Z}$. This is one of the major insights that abstract algebra provides because it lets us view two objects that look very different (like $\mathbb{R}[x]$ and $\mathbb{Z}$) as having very similar algebraic properties. Thus, intuition and results about $\mathbb{Z}$ can be translated to learn things about polynomials and vice versa.

11.1 Divisibility

We now begin the development of divisibility in commutative rings (it's possible but more involved to discuss these concepts in the noncommutative case, but you need to constantly refer to which side you are working on and this brings attention away from our primary concerns). As we will see, divisibility in general commutative rings has some undesirable properties as well, so we quickly start assuming that we are working in an integral domain. Again, the general commutative case is worthy of investigation, but the fundamental examples for us will be integral domains so we will focus our attention on those.

Definition 11.1.1. *Let R be a commutative ring and let $a, b \in R$. We say that a* divides *b, and write $a \mid b$, if there exists $d \in R$ with $b = ad$.*

Of course, this is just our definition of divisibility in $\mathbb{Z}$ generalized to an arbitrary ring R. For example, in the ring $\mathbb{Z}[x]$, we have

$$x^2 + 3x - 1 \mid x^4 - x^3 - 11x^2 + 10x - 2$$

because

$$(x^2 + 3x - 1)(x^2 - 4x + 2) = x^4 - x^3 - 11x^2 + 10x - 2$$

As noted above, if R is a field, then for any $a, b \in R$ with $a \neq 0$, we have $a \mid b$ because $b = a(a^{-1}b)$. Thus, divisibility is trivial in fields. Also notice that in a general ring R, we have that $a \in R$ is a unit if and only if $a \mid 1$.

Proposition 11.1.2. *Let R be a commutative ring and let $a, b, c \in R$.*

1. *If $a \mid b$ and $b \mid c$, then $a \mid c$.*
2. *If $a \mid b$ and $a \mid c$, then $a \mid (bx + cy)$ for all $x, y \in R$.*

Proof.

1. Fix $k, m \in R$ with $b = ak$ and $c = bm$. We then have $c = bm = (ak)m = a(km)$, so $a \mid c$.
2. Fix $k, m \in R$ with $b = ak$ and $c = am$. Let $x, y \in R$. We then have
$$bx + cy = (ak)x + (am)y = a(kx) + a(my) = a(kx + my)$$
so $a \mid (bx + cy)$.

□

For a first example of where working in an integral domain makes divisibility "easier", consider the following proposition. It is false in general commutative rings. For example, if you let $R = \mathbb{Z} \times \mathbb{Z}$, notice that $(2, 0) \mid (6, 0)$ via both $(2, 0) \cdot (3, 0) = (6, 0)$ and also $(2, 0) \cdot (3, 5) = (6, 0)$.

Proposition 11.1.3. *Let R be an integral domain. Suppose that $a, b \in R$ with $a \neq 0$ and that $a \mid b$. There exists a unique $d \in R$ such that $ad = b$.*

Proof. The existence of a d follows immediately from the definition of $a \mid b$. Suppose that $c, d \in R$ with $ac = b$ and $ad = b$. We then have that $ac = ad$. Since $a \neq 0$ and R is an integral domain, we may cancel the a's to conclude that $c = d$. □

Proposition 11.1.4. *Let R be a commutative ring and let $a, b \in R$. Define $\sim$ on R by $a \sim b$ if there exists a unit $u \in R$ such that $b = au$. The relation $\sim$ is an equivalence relation.*

Proof. We check the properties.

- Reflexive: Given $a \in R$, we have $a = a \cdot 1$, so since $1 \in R$ is a unit it follows that $a \sim a$.
- Symmetric: Suppose that $a, b \in R$ with $a \sim b$. Fix a unit $u \in R$ with $b = au$. Multiplying on the right by u^{-1}, we see that $a = bu^{-1}$. Since u^{-1} is also a unit (because $uu^{-1} = 1$), it follows that $b \sim a$.
- Transitive: Suppose that $a, b, c \in R$ with $a \sim b$ and $b \sim c$. Fix units $u, v \in R$ with $b = au$ and $c = bv$. We then have $c = bv = (au)v = a(uv)$. Since uv is a unit (its inverse is $v^{-1}u^{-1}$), it follows that $a \sim c$.

Therefore, $\sim$ is an equivalence relation. □

Definition 11.1.5. *Let R be a commutative ring. Elements of the same equivalence class are called* associates*. In other words, given $a, b \in R$, then a and b are associates if there exists a unit $u \in R$ with $b = au$.*

For example, the units of $\mathbb{Z}$ are ± 1. Thus, the associates of a given $n \in \mathbb{Z}$ are $\pm n$. These are equivalence classes and they partition $\mathbb{Z}$ into the equivalence classes $\{0\}$, $\{\pm 1\}$, $\{\pm 2\}$, ...

Proposition 11.1.6. *Let R be an integral domain and let $a, b \in R$. The following are equivalent.*

1. *a and b are associates in R.*
2. *Both $a \mid b$ and $b \mid a$.*

Proof. Suppose first that a and b are associates. Fix a unit u with $b = au$. We then clearly have $a \mid b$, and since $a = bu^{-1}$ we have $b \mid a$.

Suppose conversely that both $a \mid b$ and $b \mid a$. Fix $c, d \in R$ with $b = ac$ and $a = bd$. Notice that if $a = 0$, then $b = ac = 0c = 0$, so $a = b1$ and a and b are associates. Suppose instead that $a \neq 0$. We then have

$$a1 = a = bd = (ac)d = acd$$

Since R is an integral domain and $a \neq 0$, if follows that $cd = 1$ so both c and d are units in R. Therefore, as $b = ac$, it follows that a and b are associates. □

Our next goal is to generalize the notion of a prime number to an arbitrary ring. When we were working with $\mathbb{Z}$, we noticed the fundamental fact about prime numbers that we used again and again was that if p is a prime and $p \mid ab$, then either $p \mid a$ or $p \mid b$. It's not obvious that this is equivalent to the usual idea of being unable to factor it nontrivially, so we introduce two different notions.

Definition 11.1.7. *Let R be an integral domain and let $p \in R$ be nonzero and not a unit.*

- *We say that p is* irreducible *if whenever $p = ab$, then either a is a unit or b is a unit.*
- *We say that p is* prime *if whenever $p \mid ab$, either $p \mid a$ or $p \mid b$.*

Now in the ring $\mathbb{Z}$ we know that an element is prime exactly when it is irreducible. As we will see, one direction of this equivalence holds in every integral domain but the other does not. Before giving this direction, here's another characterization of the irreducible elements of a ring R.

Proposition 11.1.8. *Let R be an integral domain. A nonzero nonunit $p \in R$ is irreducible if and only if the only divisors of p are the units and the associates of p.*

Proof. Let $p \in R$ be nonzero and not a unit.

Suppose that p is irreducible. Let $a \in R$ with $a \mid p$. Fix $d \in R$ with $p = ad$. Since p is irreducible, we conclude that either a is a unit or d is a unit. If a is a unit, we are done. If d is a unit, then have that p and a are associates.

Suppose conversely that the only divisors of p are the units and associates of p. Suppose that $p = ab$ and that a is not a unit. Notice that $a \neq 0$ because $p \neq 0$. We have that $a \mid p$, so since a is not a unit, we must have that a is an associate of p. Fix a unit u with $p = au$. We then have $ab = p = au$, so since R is in integral domain and $a \neq 0$, it follows that $b = u$. Therefore b is a unit. □

Proposition 11.1.9. *Let R be an integral domain. If p is prime, then p is irreducible.*

Proof. Suppose that p is prime. Let $a, b \in R$ with $p = ab$. We then have $p1 = ab$, so $p \mid ab$. Since p is prime, we conclude that either $p \mid a$ or $p \mid b$. Suppose that $p \mid a$. Fix $c \in R$ with $a = pc$. We then have

$$p1 = p = ab = (pc)b = p(cb)$$

Since R is an integral domain and $p \neq 0$, it follows that $1 = cb$, so b is a unit. Suppose instead that $p \mid b$. Fix $d \in R$ with $b = pd$. We then have

$$p1 = p = ab = a(pd) = p(ad)$$

Since R is an integral domain and $p \neq 0$, it follows that $1 = ad$, so a is a unit. □

Proposition 11.1.10. *Let R be an integral domain and let $p \in R$ be nonzero. The ideal $\langle p \rangle$ is a prime ideal of R if and only if p is a prime element of R.*

Proof. Suppose first that $\langle p \rangle$ is a prime ideal of R. Notice that $p \neq 0$ by assumption and that p is not a unit because $\langle p \rangle \neq R$. Suppose that $a, b \in R$ and $p \mid ab$. We then have that $ab \in \langle p \rangle$, so as $\langle p \rangle$ is a prime ideal we know that either $a \in \langle p \rangle$ or $b \in \langle p \rangle$. In the former case, we conclude that $p \mid a$, and in the latter case we conclude that $p \mid b$. Since $a, b \in R$ were arbitrary, it follows that p is a prime element of R.

Suppose conversely that p is a prime element of R. By definition, we know that p is not a unit, so $1 \notin \langle p \rangle$ and hence $\langle p \rangle \neq R$. Suppose that $a, b \in R$ and $ab \in \langle p \rangle$. We then have that $p \mid ab$, so as p is a prime element we know that either $p \mid a$ or $p \mid b$. In the former case, we conclude that $a \in \langle p \rangle$ and in the latter case we conclude that $b \in \langle p \rangle$. Since $a, b \in R$ were arbitrary, it follows that $\langle p \rangle$ is a prime ideal of R. □

Finally, we set ourselves up for a discussion of greatest common divisors in general rings. The fundamental results about them will have to wait until we have built up some more theory. If you recall our discussion in $\mathbb{Z}$, we avoided defining it as the *largest* common divisor of a and b, and instead said that it had the property that every other common divisor of a and b was also a divisor of it. Taking this approach was useful in $\mathbb{Z}$ (after all, $\gcd(a, b)$ had this much stronger property anyway and otherwise $\gcd(0, 0)$ wouldn't make sense), but it is absolutely essential when we try to generalize it to other rings since a general ring has no notion of "order" or "largest".

Definition 11.1.11. *Let R be an integral domain and let $a, b \in R$.*

1. *A common divisor of a and b is an element $c \in R$ such that $c \mid a$ and $c \mid b$.*
2. *An element $d \in R$ is called a* greatest common divisor *of a and b if*
   - *d is a common divisor of a and b.*
   - *For every common divisor c of a and b, we have $c \mid d$.*

When we worked in $\mathbb{Z}$, we added the additional requirement that $\gcd(a, b)$ was nonnegative so that it would be unique. However, just like with "order", there is no notion of "positive" in a general ring. Thus, we will have to live with a lack of uniqueness in general. Fortunately, any two greatest common divisors are associates, as we now prove.

Proposition 11.1.12. *Suppose that R is an integral domain and $a, b \in R$.*

- *If d is a greatest common divisor of a and b, then every associate of d is also a greatest common divisor of a and b.*
- *If d and d' are both greatest common divisors of a and b, then d and d' are associates.*

Proof. Suppose that d is a greatest common divisor of a and b. Suppose that d' is an associate of d. We then have that $d' \mid d$, so since d is a common divisor of a and b and divisibility is transitive, it follows that d' is a common divisor of a and b. Let c be a common divisor of a and b. Since d is a greatest common divisor of a and b, we know that $c \mid d$. Now $d \mid d'$ because d and d' are associates, so by transitivity of divisibility, we conclude that $c \mid d'$. Therefore, every other common divisor of a and b divides d', and hence d' is a greatest common divisor of a and b.

Suppose that d and d' are both greatest common divisors of a and b. We then have that d' is a common divisor of a and b, so $d' \mid d$ because d is a greatest common divisor. Similarly, we have that d is a common divisor of a and b, so $d \mid d'$. From above, we conclude that either d and d' are associates. □

So far, we have danced around the fundamental question: Given an integral domain R and elements $a, b \in R$, must there exist a greatest common divisor of a and b? The answer is **no** in general, so proving existence in "nice" integral domains will be a top priority for us.

11.2 Polynomial Rings over Fields

As discussed in the introduction, if F is a field, then F is an integral domain, so $F[x]$ is an integral domain by Proposition 10.3.5. Thus, we can apply the concepts of the previous section to these rings. We begin by classifying the units in these polynomial rings.

Proposition 11.2.1. *Let F be a field. The units in $F[x]$ are precisely the nonzero constant polynomials. In other words, if you identify an element of F with the corresponding constant polynomial, then $U(F[x]) = F\backslash\{0\}$.*

Proof. Recall that the identity of $F[x]$ is the constant polynomial 1. If $a \in F$ is nonzero, then considering a and a^{-1} as constant polynomials in $F[x]$ we have $aa^{-1} = 1$, so $a \in U(F[x])$. Suppose conversely that $p(x) \in F[x]$ is a unit. Fix $q(x) \in F[x]$ with $p(x) \cdot q(x) = 1$. Notice that we must have both $p(x)$ and $q(x)$ be nonzero because $1 \neq 0$. Using Proposition 10.3.5, we then have

$$0 = \deg(1) = \deg(p(x) \cdot q(x)) = \deg(p(x)) + \deg(q(x))$$

Since $\deg(p(x)), \deg(q(x)) \in \mathbb{N}$, it follows that $\deg(p(x)) = 0 = \deg(q(x))$. Therefore, $p(x)$ is a nonzero constant polynomial. □

The above proposition can be false if you only assume that R is a commutative ring. It is possible to have constant polynomials not be units in $R[x]$, and it is also possible to have nonconstant polynomials be units in $R[x]$ (see the homework). However, if R is an integral domain (but possibly not a field), then the above argument does show that $U(R[x]) = U(R)$.

Now that we have classified the units in $F[x]$, we can get a characterization of the irreducible elements of $F[x]$.

Proposition 11.2.2. *Let F be a field. A polynomial $f(x) \in F[x]$ is irreducible in $F[x]$ if and only if $\deg(f(x)) \geq 1$ and it is not possible to write $f(x) = g(x) \cdot h(x)$ with $g(x), h(x) \in F[x]$ and both $\deg(g(x)) < \deg(f(x))$ and $\deg(h(x)) < \deg(f(x))$.*

Proof. Suppose first that $f(x) \in F[x]$ is a polynomial which is irreducible in $F[x]$. Since $f(x)$ is irreducible, and it nonzero and not a unit, so we must have $\deg(f(x)) \geq 1$. Suppose that $g(x), h(x) \in F[x]$ with $f(x) = g(x) \cdot h(x)$. Since $f(x)$ is irreducible, we know that either $g(x)$ is a unit or $h(x)$ is a unit, so by the previous proposition we know that either $\deg(g(x)) = 0$ or $\deg(h(x)) = 0$. Using Proposition 10.3.5, we know that

$$\deg(f(x)) = \deg(g(x) \cdot h(x)) = \deg(g(x)) + \deg(h(x))$$

It follows that either $\deg(g(x)) = \deg(f(x))$ or $\deg(h(x)) = \deg(f(x))$. Therefore, it is not possible to write $f(x) = g(x) \cdot h(x)$ with both $g(x), h(x) \in F[x]$ and both $\deg(g(x)) < \deg(f(x))$ and $\deg(h(x)) < \deg(f(x))$.

Suppose conversely that $f(x)$ is nonconstant polynomial which fails to be irreducible in $F[x]$. Fix $g(x), h(x) \in F[x]$ both nonunits with $f(x) = g(x) \cdot h(x)$. Since $f(x) \neq 0$, we know that both $g(x) \neq 0$ and $h(x) \neq 0$. Also, since both $g(x)$ and $h(x)$ are nonunits, we know from the previous proposition that they are not constant polynomials, so $\deg(g(x)) \geq 1$ and $\deg(h(x)) \geq 1$. Using Proposition 10.3.5, we know that

$$\deg(f(x)) = \deg(g(x) \cdot h(x)) = \deg(g(x)) + \deg(h(x))$$

Since $\deg(g(x)) \geq 1$, we must have $\deg(h(x)) < \deg(f(x))$ and similarly since $\deg(h(x)) \geq 1$ we must have $\deg(g(x)) < \deg(f(x))$. Thus, we have shown that it is possible to write $f(x) = g(x) \cdot h(x)$ with both $g(x), h(x) \in F[x]$ and both $\deg(g(x)) < \deg(f(x))$ and $\deg(h(x)) < \deg(f(x))$. □

As usual, life is not as pretty if you are not working over a field. For example, suppose that you are working over the nice integral domain $\mathbb{Z}$. Notice that $2x = 2 \cdot x$ and both 2 and x are not units in $\mathbb{Z}[x]$, so $2x$

fails to be irreducible in $\mathbb{Z}[x]$. However, it is not possible to write $2x = g(x) \cdot h(x)$ with both $\deg(g(x)) < 1$ and $\deg(h(x)) < 1$.

Thinking back to our work on divisibility in $\mathbb{Z}$, one of our primary tools was the ability to divide any integer a by a nonzero integer b to get a quotient q and remainder r in which the remainder was "smaller" (in that case, we had $0 \leq r < |b|$). This simple fact was the foundation for our work with greatest common divisors via the Euclidean algorithm and the proof that the greatest common divisor was the least element of the set

$$\{am + bn : m, n \in \mathbb{Z}\}$$

We would like to carry over these insights about $\mathbb{Z}$ to the ring $F[x]$. Our first step is getting an analogue of "division with remainder". If you are familiar with polynomial long division, that is exactly what the next proposition provides. If you unwrap the proof, it is precisely the algorithm of polynomial long division.

Theorem 11.2.3. *Let F be a field. Let $f(x), g(x) \in F[x]$ with $g(x) \neq 0$. There exist unique $q(x), r(x) \in F[x]$ with $f(x) = q(x) \cdot g(x) + r(x)$ and either $r(x) = 0$ or $\deg(r(x)) < \deg(g(x))$.*

Proof. Fix $g(x) \in F[x]$ with $g(x) \neq 0$. We first prove the existence of $q(x)$ and $r(x)$ for all $f(x) \in F[x]$. We begin by handling some simple cases which will serve as base cases for an induction. Notice first that if $f(x) = 0$, then we may take $q(x) = 0$ and $r(x) = 0$ because

$$f(x) = 0 \cdot g(x) + 0$$

and we are done. Also, if $f(x) \neq 0$ but $\deg(f(x)) < \deg(g(x))$, then we may take $q(x) = 0$ and $r(x) = f(x)$ because

$$f(x) = 0 \cdot g(x) + f(x)$$

We handle all other polynomials $f(x)$ using induction on $\deg(f(x))$. Suppose then that $\deg(f(x)) \geq \deg(g(x))$ and that we know the existence result is true for all $p(x) \in F[x]$ with either $p(x) = 0$ or $\deg(p(x)) < \deg(f(x))$. Suppose that

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

and

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$$

with $a_n \neq 0$ and $b_m \neq 0$. Since we are assuming that $\deg(f(x)) \geq \deg(g(x))$, we have $n \geq m$. Consider the polynomial

$$p(x) = f(x) - a_n b_m^{-1} x^{n-m} g(x)$$

We have

$$\begin{aligned}
p(x) &= f(x) - a_n b_m^{-1} x^{n-m} g(x) \\
&= (a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0) - a_n b_m^{-1} x^{n-m} \cdot (b_m x^m + b_{m-1}x^{m-1} + \cdots + m_0) \\
&= (a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0) - (a_n b_m^{-1} b_m x^n + a_n b_m^{-1} b_{m-1} x^{n-1} + \cdots + a_n b_m^{-1} b_0 x^{n-m}) \\
&= (a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0) - (a_n x^n + a_n b_m^{-1} b_{m-1} x^{n-1} + \cdots + a_n b_m^{-1} b_0 x^{n-m}) \\
&= (a_{n-1} + a_n b_m^{-1} b_{m-1}) x^{n-1} + \cdots + (a_0 + a_n b_m^{-1} b_0)
\end{aligned}$$

Therefore, either $p(x) = 0$ or $\deg(p(x)) < n = \deg(f(x))$. By induction, we may fix $q^*(x)$ and $r^*(x)$ with

$$p(x) = q^*(x) \cdot g(x) + r^*(x)$$

where either $r^*(x) = 0$ or $\deg(r^*(x)) < \deg(g(x))$. We then have

$$f(x) - a_n b_m^{-1} x^{n-m} \cdot g(x) = q^*(x) \cdot g(x) + r^*(x)$$

hence

$$\begin{aligned} f(x) &= a_n b_m^{-1} x^{n-m} \cdot g(x) + q^*(x) \cdot g(x) + r^*(x) \\ &= (a_n b_m^{-1} x^{n-m} + q^*(x)) \cdot g(x) + r^*(x) \end{aligned}$$

Thus, if we let $q(x) = (a_n b_m^{-1} x^{n-m} + q^*(x))$ and $r(x) = r^*(x)$, we have either $r(x) = 0$ or $\deg(r(x)) < \deg(g(x))$, so we have proven existence for $f(x)$. The existence result follows by induction.

We next prove uniqueness. Suppose that

$$q_1(x) \cdot g(x) + r_1(x) = f(x) = q_2(x) \cdot g(x) + r_2(x)$$

where $q_i(x), r_i(x) \in F[x]$ and either $r_i(x) = 0$ or $\deg(r_i(x)) < \deg(g(x))$ for each i. We then have

$$q_1(x) \cdot g(x) - q_2(x) \cdot g(x) = r_2(x) - r_1(x)$$

hence

$$(q_1(x) - q_2(x)) \cdot g(x) = r_2(x) - r_1(x)$$

Suppose that $r_2(x) - r_1(x) \neq 0$. Since $F[x]$ is an integral domain, we then must have $q_1(x) - q_2(x) \neq 0$ and $g(x) \neq 0$, and also

$$\deg(r_2(x) - r_1(x)) = \deg(q_1(x) - q_2(x)) + \deg(g(x)) \geq \deg(g(x))$$

However, this is a contradiction because $\deg(r_2(x) - r_1(x)) < \deg(g(x))$ since for each i we have either $r_i(x) = 0$ or $\deg(r_i(x)) < \deg(g(x))$. We conclude that we must have $r_2(x) - r_1(x) = 0$ and thus $r_1(x) = r_2(x)$. Canceling this from

$$q_1(x) \cdot g(x) - q_2(x) \cdot g(x) = r_2(x) - r_1(x)$$

we conclude that

$$q_1(x) \cdot g(x) = q_2(x) \cdot g(x)$$

Since $g(x) \neq 0$ and $F[x]$ is an integral domain, it follows that $q_1(x) = q_2(x)$ as well. This finishes the proof. □

Let's compute an example working over the field $\mathbb{Z}/7\mathbb{Z}$. Working in $\mathbb{Z}/7\mathbb{Z}$, let

$$f(x) = \overline{3}x^4 + \overline{6}x^3 + \overline{1}x^2 + \overline{2}x + \overline{2}$$

and let

$$g(x) = \overline{2}x^2 + \overline{5}x + \overline{1}$$

We perform long division, i.e. follow the proof, to find $q(x)$ and $r(x)$. Notice that the leading coefficient of $g(x)$ is $\overline{2}$ and that in $\mathbb{Z}/7\mathbb{Z}$ we have $\overline{2}^{-1} = \overline{4}$. We begin by computing

$$\overline{3} \cdot \overline{4} \cdot x^{4-2} = \overline{5}x^2$$

This will be the first term in our resulting quotient. We then multiply this by $\overline{5}x^2 \cdot g(x)$ and subtract from $f(x)$ to obtain

$$\begin{aligned} f(x) - \overline{5}x^2 \cdot g(x) &= (\overline{3}x^4 + \overline{6}x^3 + \overline{1}x^2 + \overline{2}x + \overline{2}) - \overline{5}x^2 \cdot (\overline{2}x^2 + \overline{5}x + \overline{1}) \\ &= (\overline{3}x^4 + \overline{6}x^3 + \overline{1}x^2 + \overline{2}x + \overline{2}) - (\overline{3}x^4 + \overline{4}x^3 + \overline{5}x^2) \\ &= \overline{2}x^3 + \overline{3}x^2 + \overline{2}x + \overline{2} \end{aligned}$$

We now continue on with this new polynomial (this is where we appealed to induction in the proof) as our "new" $f(x)$. We follow the proof recursively and compute

$$\overline{2} \cdot \overline{4} \cdot x = \overline{1}x$$

This will be our next term in the quotient. We now subtract $\overline{1}x \cdot g(x)$ from our current polynomial to obtain

$$\begin{aligned}(\overline{2}x^3 + \overline{3}x^2 + \overline{2}x + \overline{2}) - \overline{x} \cdot g(x) &= (\overline{2}x^3 + \overline{3}x^2 + \overline{2}x + \overline{2}) - \overline{x} \cdot (\overline{2}x^2 + \overline{5}x + \overline{1}) \\ &= (\overline{2}x^3 + \overline{3}x^2 + \overline{2}x + \overline{2}) - (\overline{2}x^3 + \overline{5}x^2 + \overline{1}x) \\ &= \overline{5}x^2 + \overline{1}x + \overline{2}\end{aligned}$$

Continuing on, we compute

$$\overline{5} \cdot \overline{4} = \overline{6}$$

and add this to our quotient. We now subtract $\overline{6} \cdot g(x)$ from our current polynomial to obtain

$$\begin{aligned}(\overline{5}x^2 + \overline{1}x + \overline{2}) - \overline{6} \cdot g(x) &= (\overline{5}x^2 + \overline{1}x + \overline{2}) - \overline{6} \cdot (\overline{2}x^2 + \overline{5}x + \overline{1}) \\ &= (\overline{5}x^2 + \overline{1}x + \overline{2}) - (\overline{5}x^2 + \overline{2}x + \overline{6}) \\ &= \overline{6}x + \overline{3}\end{aligned}$$

We have arrived at a point with our polynomial has degree less than that of $g(x)$, so we have bottomed out in the above proof at a base case. Adding up our contributions to the quotient gives.

$$q(x) = \overline{5}x^2 + \overline{1}x + \overline{2}$$

and we are left with the remainder

$$r(x) = \overline{6}x + \overline{3}$$

Therefore, we have written

$$\overline{3}x^4 + \overline{6}x^3 + \overline{1}x^2 + \overline{2}x + \overline{2} = (\overline{5}x^2 + \overline{1}x + \overline{2}) \cdot (\overline{2}x^2 + \overline{5}x + \overline{1}) + (\overline{6}x + \overline{3})$$

Our first theoretical use of the the ability to divide and get a remainder roots of polynomials.

Definition 11.2.4. *Let F be a field. A* root *of a polynomial $f(x) \in F[x]$ is an element $a \in F$ such that $f(a) = 0$ (or more formally $Ev_a(f(x)) = 0$).*

Proposition 11.2.5. *Let F be a field and let $a \in F$. For each polynomial $f(x) \in F[x]$, the following are equivalent:*

1. *a is a root of $f(x)$.*
2. *$(x - a) \mid f(x)$ in $F[x]$*

Proof. Suppose first that $(x - a) \mid f(x)$ in $F[x]$. Fix $g(x) \in F[x]$ with $f(x) = (x - a) \cdot g(x)$. We then have

$$f(a) = (a - a) \cdot g(a) = 0 \cdot g(a) = 0$$

or more formally:

$$\begin{aligned}Ev_a(f(x)) &= Ev_a((x - a) \cdot g(x)) \\ &= Ev_a(x - a) \cdot Ev_a(g(x)) && \text{(since } Ev_a \text{ is a ring homomorphism)} \\ &= 0 \cdot Ev_a(g(x)) \\ &= 0\end{aligned}$$

Therefore, a is a root of $f(x)$.

Suppose conversely that a is a root of $f(x)$. Fix $q(x), r(x) \in F[x]$ with

$$f(x) = q(x) \cdot (x - a) + r(x)$$

and either $r(x) = 0$ or $\deg(r(x)) < \deg(x - a)$. Since $\deg(x - a) = 1$, we have that $r(x)$ is a constant polynomial, so writing $r(x) = c$ for some $c \in F$ we have

$$f(x) = q(x) \cdot (x - a) + c$$

Now plugging in a (or applying Ev_a) we see that

$$f(a) = q(a) \cdot (a - a) + c = q(a) \cdot 0 + c = c$$

Since we are assuming that $f(a) = 0$, it follows that $c = 0$, and thus

$$f(x) = q(x) \cdot (x - a)$$

Therefore, $(x - a) \mid f(x)$. □

Proposition 11.2.6. *Let F be a field and let $f(x) \in F[x]$ be a nonzero polynomial. The polynomial $f(x)$ has at most $\deg(f(x))$ many roots in F. In particular, every nonzero polynomial in $F[x]$ has finitely many roots.*

Proof. We prove the result by induction on $\deg(f(x))$. If $\deg(f(x)) = 0$, then $f(x)$ is a nonzero constant polynomial, so has 0 roots in F. Suppose that the result is true for all polynomials of degree n, and let $f(x) \in F[x]$ with $\deg(f(x)) = n+1$. If $f(x)$ has no roots in F, then we are done because $0 \leq n+1$. Suppose then that $f(x)$ has at least one root in F, and fix such a root $a \in F$. From above we know that $(x-a) \mid f(x)$, so we may fix $g(x) \in F[x]$ with

$$f(x) = (x - a) \cdot g(x)$$

Notice that

$$\begin{aligned} n + 1 &= \deg(f(x)) \\ &= \deg((x - a) \cdot g(x)) \\ &= \deg(x - a) + \deg(g(x)) \\ &= 1 + \deg(g(x)) \end{aligned}$$

so $\deg(g(x)) = n$. By induction, we know that $g(x)$ has at most n roots in F. Notice that if b is a root of $f(x)$, then

$$0 = f(b) = (b - a) \cdot g(b)$$

so either $b - a = 0$ or $g(b) = 0$ (because F is an integral domain), and hence either $b = a$ or b is a root of $g(x)$. Therefore, $f(x)$ has at most $n + 1$ roots in F, namely the roots of $g(x)$ together with a. The result follows by induction. □

Back when we defined the polynomial ring $R[x]$, we were very careful to define an element as a sequence of coefficients rather than as the function defined from R to R given by evaluation. As we saw, if $F = \mathbb{Z}/2\mathbb{Z}$, then the two distinct polynomials

$$\overline{1}x^2 + \overline{1}x + \overline{1} \qquad\qquad \overline{1}$$

given the same function on $\mathbb{Z}/2\mathbb{Z}$ because they evaluate to the same value for each element of $\mathbb{Z}/2\mathbb{Z}$. We now see that such a situation is impossible for an infinite field.

Proposition 11.2.7. *Let F be an infinite field and let $f(x), g(x) \in F[x]$. If $f(a) = g(a)$ for all $a \in F$, then $f(x) = g(x)$. Thus, if F is infinite, then distinct polynomials give different functions from F to F.*

Proof. Suppose that $f(a) = g(a)$ for all $a \in F$. Consider the polynomial $h(x) = f(x) - g(x) \in F[x]$. For every $a \in F$, we have $h(a) = f(a) - g(a) = 0$. Since F is infinite, we conclude that $h(x)$ has infinitely many roots, so by the previous proposition we must have that $h(x) = 0$. Thus, $f(x) - g(x) = 0$, and adding $g(x)$ to both sides we conclude that $f(x) = g(x)$. □

11.3 Euclidean Domains

In the last section, we established an analog of division with remainder in the ring $F[x]$ of polynomials over a field F. We immediately mined some consequences of that fact, and we will continue to do so in the coming sections. However, before we jump into these, we will define a class of integral domains based on the idea of allowing "division with remainder" so that our results will be as general as possible.

Definition 11.3.1. *Let R be an integral domain. A function $N \colon R\backslash\{0\} \to \mathbb{N}$ is called a Euclidean function on R if for all $a, b \in R$ with $b \neq 0$, there exist $q, r \in R$ such that*

$$a = qb + r$$

and either $r = 0$ or $N(r) < N(b)$.

Definition 11.3.2. *An integral domain R is a* Euclidean domain *if there exists a Euclidean function on R.*

Example 11.3.3. *Our work so far has established the following.*

- *The function $N \colon \mathbb{Z}\backslash\{0\} \to \mathbb{N}$ defined by $N(a) = |a|$ is a Euclidean function on $\mathbb{Z}$, so $\mathbb{Z}$ is a Euclidean domain.*
- *Let F be a field. The function $N \colon F[x]\backslash\{0\} \to \mathbb{N}$ defined by $N(f(x)) = \deg(f(x))$ is a Euclidean function on $F[x]$, so $F[x]$ is a Euclidean domain.*

Notice that we do not require the uniqueness of q and r in our definition of a Euclidean function. Although it was certainly a nice perk to have some aspect of uniqueness in $\mathbb{Z}$ and $F[x]$, it turns out to be be unnecessary for the theoretical results of interest about Euclidean domains. Furthermore, there are some natural Euclidean functions on integral domains for which uniqueness fails, and we want to be as general as possible.

The name *Euclidean domain* comes from the fact that any such integral domain supports the ability to find greatest common divisors via the Euclidean algorithm. In particular, the notion of "size" given by a Euclidean function $N \colon R \to \mathbb{N}$ allows us to use induction to prove the existence of greatest common divisors. We begin with the following generalization of a simple result we proved about $\mathbb{Z}$ which works in any integral domain (even any commutative ring).

Proposition 11.3.4. *Let R be an integral domain. Let $a, b, q, r \in R$ with $a = qb + r$. For any $d \in R$, we have that d is a common divisor of a and b if and only if d is a common divisor of b and r, i.e.*

$$\{d \in R : d \text{ is a common divisor of } a \text{ and } b\} = \{d \in R : d \text{ is a common divisor of } b \text{ and } r\}$$

Proof. Suppose first that d is a common divisor of b and r. Since $d \mid b$, $d \mid r$, and $a = qb + r = bq + r1$, it follows that $d \mid a$.

Conversely, suppose that d is a common divisor of a and b. Since $d \mid a$, $d \mid b$, and $r = a - qb = a1 + b(-q)$, it follows that $d \mid r$. □

Theorem 11.3.5. *Let R be a Euclidean domain. Every pair of elements $a, b \in R$ has a greatest common divisor.*

Proof. Since R is a Euclidean domain, we may fix a Euclidean function $N\colon R\backslash\{0\} \to \mathbb{N}$. We use (strong) induction on $N(b) \in \mathbb{N}$ to prove the result. We begin by noting that if $b = 0$, then the set of common divisors of a and b equals the set of divisors of a (because every integer divides 0), so a satisfies the requirement of a greatest common divisor. Suppose then that $b \in R$ is nonzero and we know the result for all pairs $x, y \in R$ with either $y = 0$ or $N(y) < N(b)$. Fix $q, r \in R$ with $a = qb + r$ and either $r = 0$ or $N(r) < N(b)$. By (strong) induction, we know that b and r have a greatest common divisor d. By the Proposition 11.3.4, the set of common divisors of a and b equals the set of common divisors of b and r. It follows that d is a greatest common divisor of a and b. □

As an example, consider working in the ring $\mathbb{Q}[x]$ and trying to find a greatest common divisor of the following two polynomials:

$$f(x) = x^5 + 3x^3 + 2x^2 + 6 \qquad g(x) = x^4 - x^3 + 4x^2 - 3x + 3$$

We apply the Euclidean Algorithm as follows (we suppress the computations of the long divisions):

$$\begin{aligned} x^5 + 3x^3 + 2x^2 + 6 &= (x+1)(x^4 - x^3 + 4x^2 - 3x + 3) + (x^2 + 3) \\ x^4 - x^3 + 4x^2 - 3x + 3 &= (x^2 - x + 1)(x^2 + 3) + 0 \end{aligned}$$

Thus, the set of common of $f(x)$ and $g(x)$ equals the set of common divisors of x^2+3 and 0, which is just the set of divisors of x^2+3. Therefore, x^2+3 is a greatest common divisor of $f(x)$ and $g(x)$. Now this is not the only greatest common divisor because we know that any associate of $x^2 + 3$ will also be a greatest common divisor of $f(x)$ and $g(x)$. The units in $\mathbb{Q}[x]$ are the nonzero constants, so other greatest common divisors are $2x^2 + 6$, $\frac{5}{6}x^2 + \frac{5}{2}$, etc. We would like to have a canonical choice for which to pick, akin to choosing the nonnegative value when working in $\mathbb{Z}$.

Definition 11.3.6. *Let F be a field. A* monic *polynomial in $F[x]$ is a nonzero polynomial whose leading term is* 1*.*

Notice that every nonzero polynomial in $F[x]$ is an associate with a unique monic polynomial (if the leading term is $a \neq 0$, just multiply by a^{-1} to get a monic associate, and notice that this is the only way to multiply by a nonzero constant to make it monic). By restricting to monic polynomials, we get a canonical choice for a greatest common divisor.

Definition 11.3.7. *Let F be a field and let $f(x), g(x) \in F[x]$ be polynomials. If at least one of $f(x)$ and $g(x)$ is nonzero, we define* $\gcd(f(x), g(x))$ *to be the unique monic polynomial which is a greatest common divisor of $f(x)$ and $g(x)$. Notice that if both $f(x)$ and $g(x)$ are the zero polynomial, then* 0 *is the only greatest common divisor of $f(x)$ and $g(x)$, so we define* $\gcd(f(x), g(x)) = 0$*.*

Now $x^2 + 3$ is monic, so from the above computations, we have

$$\gcd(x^5 + 3x^3 + 2x^2 + 6, x^4 - x^3 + 4x^2 - 3x + 3) = x^2 + 3$$

We end this section by showing that the Gaussian Integers $\mathbb{Z}[i]$ are also a Euclidean domain.

Definition 11.3.8. *Working in the field $\mathbb{C}$, we define the following.*

- $\mathbb{Q}(i) = \{q + ri : q, r \in \mathbb{Q}\}$
- $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$

Notice that $\mathbb{Q}(i)$ is a subfield of $\mathbb{C}$ and $\mathbb{Z}[i]$ is a subring of $\mathbb{Q}(i)$ and thus of $\mathbb{C}$. The ring $\mathbb{Z}[i]$ is called the Gaussian Integers.

To see that $\mathbb{Q}(i)$ is a field, suppose that $\alpha \in \mathbb{Q}(i)$ is nonzero and write $\alpha = q + ri$. We then have that either $q \neq 0$ or $r \neq 0$, so

$$\begin{aligned}\frac{1}{\alpha} &= \frac{1}{q+ri} \\ &= \frac{1}{q+ri} \cdot \frac{q-ri}{q-ri} \\ &= \frac{q-ri}{q^2+r^2} \\ &= \frac{q}{q^2+r^2} + \frac{-r}{q^2+r^2} \cdot i\end{aligned}$$

Since both $\frac{q}{q^2+r^2}$ and $\frac{-r}{q^2+r^2}$ are elements of $\mathbb{Q}$, it follows that $\frac{1}{\alpha} \in \mathbb{Q}(i)$.

Definition 11.3.9. *We define a function $N\colon \mathbb{Q}(i) \to \mathbb{Q}$ by letting $N(q+ri) = q^2 + r^2$. The function N is called the* norm *on the field $\mathbb{Q}(i)$.*

Proposition 11.3.10. *For the function $N(q+ri) = q^2 + r^2$ defined on $\mathbb{Q}(i)$, we have*

1. *$N(\alpha) \geq 0$ for all $\alpha \in \mathbb{Q}(i)$.*
2. *$N(\alpha) = 0$ if and only if $\alpha = 0$.*
3. *$N(q) = q^2$ for all $q \in \mathbb{Q}$.*
4. *$N(\alpha) \in \mathbb{N}$ for all $\alpha \in \mathbb{Z}[i]$.*
5. *$N(\alpha\beta) = N(\alpha) \cdot N(\beta)$ for all $\alpha, \beta \in \mathbb{Q}(i)$.*

Proof. The first four are all immediate from the definition. Suppose that $\alpha, \beta \in \mathbb{Q}(i)$ and write $\alpha = q + ri$ and $\beta = s + ti$. We have

$$\begin{aligned}N(\alpha\beta) &= N((q+ri)(s+ti)) \\ &= N(qs + rsi + qti - rt) \\ &= N((qs-rt) + (rs+qt)i) \\ &= (qs-rt)^2 + (rs+qt)^2 \\ &= q^2s^2 - 2qsrt + r^2t^2 + r^2s^2 + 2rsqt + q^2t^2 \\ &= q^2s^2 + r^2s^2 + q^2t^2 + r^2t^2 \\ &= (q^2+r^2) \cdot (s^2+t^2) \\ &= N(q+ri) \cdot N(s+ti) \\ &= N(\alpha) \cdot N(\beta)\end{aligned}$$

□

Theorem 11.3.11. *$\mathbb{Z}[i]$ is a Euclidean domain with Euclidean function $N(a+bi) = a^2 + b^2$.*

Proof. We already know that $\mathbb{Z}[i]$ is an integral domain. Suppose that $\alpha, \beta \in \mathbb{Z}[i]$ with $\beta \neq 0$. When we divide α by β in the field $\mathbb{Q}(i)$ we get $\frac{\alpha}{\beta} = s + ti$ for some $s, t \in \mathbb{Q}$. Fix integers $m, n \in \mathbb{Z}$ closest to $s, t \in \mathbb{Q}$ respectively, i.e. fix $m, n \in \mathbb{Z}$ so that $|m - s| \leq \frac{1}{2}$ and $|n - t| \leq \frac{1}{2}$. Let $\gamma = m + ni \in \mathbb{Z}[i]$, and let $\rho = \alpha - \beta\gamma \in \mathbb{Z}[i]$. We then have that $\alpha = \beta\gamma + \rho$, so we need only show that $N(\rho) < N(\beta)$. Now

$$\begin{aligned} N(\rho) &= N(\alpha - \beta\gamma) \\ &= N(\beta \cdot (s + ti) - \beta \cdot \gamma) \\ &= N(\beta \cdot ((s + ti) - (m + ni)) \\ &= N(\beta \cdot ((s - m) + (t - n)i) \\ &= N(\beta) \cdot N((s - m) + (t - n)i) \\ &= N(\beta) \cdot ((s - m)^2 + (t - n)^2) \\ &\leq N(\beta) \cdot (\frac{1}{4} + \frac{1}{4}) \\ &= \frac{1}{2} \cdot N(\beta) \\ &< N(\beta) \end{aligned}$$

where the last line follows because $N(\beta) > 0$. □

We work out an example of finding a greatest common of $8 + 9i$ and $10 - 5i$ in $\mathbb{Z}[i]$. We follow the proof to find quotients and remainders. Notice that

$$\begin{aligned} \frac{8 + 9i}{10 - 5i} &= \frac{8 + 9i}{10 - 5i} \cdot \frac{10 + 5i}{10 + 5i} \\ &= \frac{80 + 40i + 90i - 45}{100 + 25} \\ &= \frac{35 + 130i}{125} \\ &= \frac{7}{25} + \frac{26}{25} \cdot i \end{aligned}$$

Following the proof (where we take the closest integers to $\frac{7}{25}$ and $\frac{26}{25}$), we should use the quotient i and determine the remainder from there. We thus write

$$8 + 9i = i \cdot (10 - 5i) + (3 - i)$$

Notice that $N(3 - i) = 9 + 1 = 10$ which is less than $N(10 - 5i) = 100 + 25 = 125$. Following the Euclidean algorithm, we next calculate

$$\begin{aligned} \frac{10 - 5i}{3 - i} &= \frac{10 - 5i}{3 - i} \cdot \frac{3 + i}{3 + i} \\ &= \frac{30 + 10i - 15i + 5}{9 + 1} \\ &= \frac{35 - 5i}{10} \\ &= \frac{7}{2} - \frac{1}{2} \cdot i \end{aligned}$$

Following the proof (where we now have many choices because $\frac{7}{2}$ is equally close to 3 and 4 and $-\frac{1}{2}$ is equally close to -1 and 0), we choose to take the quotient 3. We then write

$$10 - 5i = 3 \cdot (3 - i) + (1 - 2i)$$

Notice that $N(1-2i) = 1+4 = 5$ which is less than $N(3-i) = 9+1 = 10$. Going to the next step, we calculate

$$\begin{aligned}\frac{3-i}{1-2i} &= \frac{3-i}{1-2i}\cdot\frac{1+2i}{1+2i}\\ &= \frac{3+6i-i+2}{1+4}\\ &= \frac{5+5i}{5}\\ &= 1+i\end{aligned}$$

Therefore, we have

$$3-i = (1+i)\cdot(1-2i)+0$$

Putting together the various divisions, we see the Euclidean algorithm as:

$$\begin{aligned}8+9i &= i\cdot(10-5i)+(3-i)\\ 10-5i &= 3\cdot(3-i)+(1-2i)\\ 3-i &= (1+i)\cdot(1-2i)+0\end{aligned}$$

Thus, the set of common divisors of $8+9i$ and $10-5i$ equals the set of common divisors of $1-2i$ and 0, which is just the set of divisors of $1-2i$. Since a greatest common divisor is unique up to associates and the units of $\mathbb{Z}[i]$ are $1, -1, i, -i$, it follows the set of greatest common divisors of $8+9i$ and $10-5i$ is

$$\{1-2i, -1+2i, 2+i, -2-i\}$$

11.4 Principal Ideal Domains

We chose our definition of a Euclidean domain to abstract away the fundamental fact about $\mathbb{Z}$ that we can always divide in such a way to get a quotient along with a "smaller" remainder. As we have seen, this ability allows us to carry over to these more general rings the existence of greatest common divisors and the method of finding them via the Euclidean Algorithm.

Recall back when we working with $\mathbb{Z}$ that we had another characterization of (and proof of existence for) the greatest common divisor. We proved that the greatest common divisor of two nonzero integers a and b was the least positive number of the form $ma+nb$ where $m, n \in \mathbb{Z}$. Now the "least" part will have no analogue in a general integral domain, so we will have to change that. Perhaps surprisingly, it turns out that the way to generalize this construction is to work with ideals. As we will see, in hindsight, what makes this approach to greatest common divisors work in $\mathbb{Z}$ is the fact that every ideal of $\mathbb{Z}$ is principal (from Proposition 10.5.5). We give the integral domains which have this property a special name.

Definition 11.4.1. *A* principal ideal domain*, or* PID*, is an integral domain in which every ideal is principal.*

Before working with these rings on their own terms, we first prove that every Euclidean domains is a PID so that we have a decent supply of examples. Our proof generalizes the one for $\mathbb{Z}$ in the sense that instead of looking for a smallest positive element of the ideal we simply look for an element of smallest "size" according to a given Euclidean function.

Theorem 11.4.2. *Every Euclidean domain is a PID.*

Proof. Let R be a Euclidean domain, and fix a Euclidean function $N\colon R\backslash\{0\} \to \mathbb{N}$. Suppose that I is an ideal of R. If $I = \{0\}$, then $I = \langle 0\rangle$. Suppose then that $I \neq \{0\}$. The set

$$\{N(a) : a \in I\backslash\{0\}\}$$

is a nonempty subset of $\mathbb{N}$. By the well-ordering property of $\mathbb{N}$, the set has a least element m. Fix $b \in I$ with $N(b) = m$. Since $b \in I$, we clearly have $\langle b \rangle \subseteq I$. Suppose now that $a \in I$. Fix $q, r \in R$ with

$$a = qb + r$$

and either $r = 0$ or $N(r) < N(b)$. Since $r = a - qb$ and both $a, d \in I$, it follows that $r \in I$. Now if $r \neq 0$, then $N(r) < N(b) = m$ contradicting our minimality of m. Therefore, we must have $r = 0$ and so $a = qb$. It follows that $a \in \langle b \rangle$. Since $a \in I$ was arbitrary, we conclude that $I \subseteq \langle b \rangle$. Therefore, $I = \langle b \rangle$. □

Corollary 11.4.3. *$\mathbb{Z}$, $F[x]$ for F a field, and $\mathbb{Z}[i]$ are all PIDs.*

Notice also that all fields F are also PIDs for the trivial reason that the only ideals of F are $\{0\} = \langle 0 \rangle$ and $F = \langle 1 \rangle$. In fact, all fields are also trivially Euclidean domain via absolutely any function $N \colon F \backslash \{0\} \to \mathbb{N}$ because you can always divide by a nonzero element with zero as a remainder.

It turns out that there are PIDs which are not Euclidean domains, but we will not construct examples of such rings now. Returning to our other characterization of greatest common divisors in $\mathbb{Z}$, we had that if $a, b \in \mathbb{Z}$ not both nonzero, then we considered the set

$$\{ma + nb : m, n \in \mathbb{Z}\}$$

and proved that the least positive element of this set was the greatest common divisor. In our current ring-theoretic language, the above set is the ideal $\langle a, b \rangle$ of $\mathbb{Z}$, and a generator of this ideal is a greatest common divisor. With this change in perspective/language, we can carry this argument over to an arbitrary PID.

Theorem 11.4.4. *Let R be a PID and let $a, b \in R$.*

1. *There exists a greatest common divisor of a and b.*
2. *If d is a greatest common divisor of a and b, then there exists $r, s \in R$ with $d = ra + sb$.*

Proof.

1. Let $a, b \in R$. Consider the ideal

   $$I = \langle a, b \rangle = \{ra + sb : r, s \in R\}$$

   Since R is a PID, the ideal I is principal, so we may fix $d \in R$ with $I = \langle d \rangle$. Since $d \in \langle d \rangle = \langle a, b \rangle$, we may fix $r, s \in R$ with $ra + sb = d$. We claim that d is a greatest common divisor of a and b.

   First notice that $a \in I$ since $a = 1a + 0b$, so $a \in \langle d \rangle$, and hence $d \mid a$. Also, we have $b \in I$ because $b = 0a + 1b$, so $b \in \langle d \rangle$, and hence $d \mid b$. Thus, d is a common divisor of a and b.

   Suppose now that c is a common divisor of a and b. Fix $m, n \in R$ with $a = cm$ and $b = cn$. We then have

   $$\begin{aligned} d &= ra + sb \\ &= r(cm) + s(cn) \\ &= c(rm + sn) \end{aligned}$$

   Thus, $c \mid d$. Putting it all together, we conclude that d is a greatest common divisor of a and b.

2. For the d in part 1, we showed in the proof that there exist $r, s \in R$ with $d = ra + sb$. Let d' be any other greatest common divisor of a and b, and fix a unit u with $d' = du$. We then have

   $$d' = du = (ra + sb)u = a(ru) + b(su)$$

□

If you are given $a, b \in R$ and you know a greatest common divisor d of a and b, how can you explicitly calculate $r, s \in R$ with $ra + sb = d$? In a general PID, this can be very hard. However, suppose you are in the special case where R is a Euclidean domain. Assuming that we can explicitly calculate quotients and remainders for repeated division (as we could in $\mathbb{Z}$, $F[x]$, and $\mathbb{Z}[i]$), we can calculate a greatest common divisor d of a and b by "winding up" the Euclidean algorithm backwards as in $\mathbb{Z}$.

For example, working in the Euclidean domain $\mathbb{Z}[i]$, we computed in the last section that $1 - 2i$ is a greatest common divisor of $8 + 9i$ and $10 - 5i$ by applying the Euclidean algorithm to obtain:

$$\begin{aligned} 8 + 9i &= i \cdot (10 - 5i) + (3 - i) \\ 10 - 5i &= 3 \cdot (3 - i) + (1 - 2i) \\ 3 - i &= (1 + i) \cdot (1 - 2i) + 0 \end{aligned}$$

Working backwards, we see that

$$\begin{aligned} 1 - 2i &= 1 \cdot (10 - 5i) + (-3) \cdot (3 - i) \\ &= 1 \cdot (10 - 5i) + (-3) \cdot [(8 + 9i) - i \cdot (10 - 5i)] \\ &= (1 + 3i) \cdot (10 - 5i) + (-3) \cdot (8 + 9i) \end{aligned}$$

It is possible to define a greatest common divisor of elements $a_1, a_2, \ldots, a_n \in R$ completely analogously to our definition for pairs of elements. If you do so, even in the case of a nice Euclidean domain, you can't immediately generalize the idea of the Euclidean Algorithm to many elements without doing a kind of repeated nesting that gets complicated. However, notice that you can very easily generalize our PID arguments to prove that greatest common divisors exist and are unique up to associates by following the above proofs and simply replacing the ideal $\langle a, b \rangle$ with the ideal $\langle a_1, a_2, \ldots, a_n \rangle$. You even conclude that it possible to write a greatest common divisor in the form $r_1 a_1 + r_2 a_2 + \cdots + r_n a_n$. The assumption that all ideals are principal is extremely powerful.

With the hard work of the last couple of sections in hand, we can now carry over much of our later work in $\mathbb{Z}$ which dealt with relatively prime integers and primes. The next definition and ensuing two propositions directly generalize corresponding results about $\mathbb{Z}$.

Definition 11.4.5. *Let R be a PID. Two elements $a, b \in R$ are* relatively prime *if 1 is a greatest common divisor of a and b.*

Proposition 11.4.6. *Let R be a PID and let $a, b, c \in R$. If $a \mid bc$ and a and b are relatively prime, then $a \mid c$.*

Proof. Fix $d \in R$ with $bc = ad$. Fix $r, s \in R$ with $ra + sb = 1$. Multiplying this last equation through by c, we conclude that $rac + sbc = c$, so

$$\begin{aligned} c &= rac + s(bc) \\ &= rac + s(ad) \\ &= a(rc + sd) \end{aligned}$$

It follows that $a \mid c$. □

Proposition 11.4.7. *Suppose that R is a PID. If p is irreducible, then p is prime.*

Proof. Suppose that $p \in R$ is irreducible. By definition, p is nonzero and not a unit. Suppose that $a, b \in R$ are such that $p \mid ab$. Fix a greatest common divisor d of p and a. Since $d \mid p$, we may fix $c \in R$ with $p = dc$. Now p is irreducible, so either d is a unit or c is a unit. We handle each case.

- Suppose that d is a unit. We then have that 1 is an associate of d, so 1 is also a greatest common divisor of p and a. Therefore, p and a are relatively prime, so as $p \mid ab$ we may use the previous corollary to conclude that $p \mid b$.

- Suppose that c is a unit. We then have that $pc^{-1} = d$, so $p \mid d$. Since $d \mid a$, it follows that $p \mid a$.

Therefore, either $p \mid a$ or $p \mid b$. It follows that p is prime. □

Finally, we end with a nice result about PIDs which ties together various notions of ideals with prime and irreducible elements.

Proposition 11.4.8. *Let R be a PID and let $a \in R$ be nonzero. The following are equivalent.*

1. *$\langle a \rangle$ is a maximal ideal.*

2. *$\langle a \rangle$ is a prime ideal.*

3. *a is a prime.*

4. *a is irreducible.*

Proof. We have already proved much of this, so let's recap what we know.

- $1 \to 2$ is Corollary 10.5.12.

- $2 \leftrightarrow 3$ is Proposition 11.1.10.

- $3 \to 4$ is Proposition 11.1.9.

- $4 \to 3$ is Proposition 11.4.7

To finish the equivalences, we prove that $4 \to 1$.

Suppose that $a \in R$ is irreducible, and let $M = \langle a \rangle$. Since a is not a unit, we have that $1 \notin \langle a \rangle$, so $M \neq R$. Suppose that I is an ideal with $M \subseteq I \subseteq R$. Since R is a PID, there exists $b \in R$ with $I = \langle b \rangle$. We then have that $\langle a \rangle \subseteq \langle b \rangle$, so $b \mid a$. Fix $c \in R$ with $a = bc$. Since a is irreducible, either b is a unit or c is a unit. In the former case, we have that $1 \in \langle b \rangle = I$, so $I = R$. In the latter case we have that b is an associate of a so $I = \langle b \rangle = \langle a \rangle = M$ by homework. Thus, there is no ideal I of R with $M \subsetneq I \subsetneq R$. □

You might think that the previous proposition says that in a PID every prime ideal is maximal. This is almost true, but pay careful attention to the assumption that $a \neq 0$. In a PID R, the ideal $\{0\}$ is always a prime ideal, but it is only maximal in the trivial special case of when R is a field.

You have seen an example of an integral domain which is not a PID in the homework where you showed that in $\mathbb{Z}[x]$ the ideal $\langle 2, x \rangle$ is not principal. Looking back at the above arguments, it seems natural to conjecture that 2 and x have no greatest common divisor in $\mathbb{Z}[x]$. However, this is not the case. If you examine the proof that $\langle 2, x \rangle$ is not principal (or just work it out directly), the only common divisors of 2 and x in $\mathbb{Z}[x]$ are ± 1, so both ± 1 are greatest common divisors of 2 and x in $\mathbb{Z}[x]$. Although 1 is a greatest common divisor of 2 and x in $\mathbb{Z}[x]$, notice that there does not exist $p(x), q(x) \in \mathbb{Z}[x]$ with $1 = p(x) \cdot 2 + q(x) \cdot x$. Thus, if you are not a PID, you might still have greatest common divisors, but it might be the case that you can not "reach" them using a linear combination with coefficients from R.

11.5 Unique Factorization Domains

In the last section, we showed that when R is a PID we can recover pretty much everything we did in Section 2 when working in the special case of $\mathbb{Z}$. The one thing that we have not yet proved in a general setting is the analogue of the Fundamental Theorem of Arithmetic. When thinking about how to formulate what unique factorization means in a general integral domain R, let's first go back and look at $\mathbb{Z}$. Of course, our uniqueness of prime factorizations was only up to order, but there was one issue that we were able to sweep under the rug in $\mathbb{Z}$ by working only with positive elements. For example, consider the following factorizations of 30 in $\mathbb{Z}$:

$$30 = 2 \cdot 3 \cdot 5 = (-2) \cdot (-3) \cdot 5 = (-2) \cdot 3 \cdot (-5)$$

Thus, if we move away from working only with positive primes, then we lose a bit more uniqueness. However, if we slightly loosen the requirements that any two factorizations are "the same up to order" to "the same up to order and associates", we might have a chance.

Definition 11.5.1. *A* Unique Factorization Domain*, or UFD, is an integral domain R such that:*

1. *Every nonzero nonunit is a product of irreducible elements.*
2. *If $p_1 p_2 \cdots p_n = q_1 q_2 \dots q_m$ where each p_i and q_j are irreducible, then $m = n$ and there exists a permutation $\sigma \in S_n$ such that p_i and $q_{\sigma(i)}$ are associates for every i.*

Our primary goal is to prove that every PID is a UFD. You might think that the challenge would be in proving condition 2, but actually that part requires no fundamentally new ideas. The real challenge is the first condition. We proved that every $n \geq 2$ was a product of primes by induction. Intuitively, if n is not prime, then factor it, and if those factors are not prime, then factor them, etc. The key fact which makes this "bottom out" is that the numbers are getting smaller and there you can not have an infinite descending sequence of natural numbers. However, in a general PID, it is not clear what is getting "smaller" to force this process to stop.

The general idea is as follows. Suppose we have an element $a \in R$ which is a nonzero nonunit. If a is not irreducible, then we can write $a = bc$ where neither b nor c are units. We then have $b \mid a$ but b and a are not associates (since c is not a unit), so in terms of ideals we get $\langle a \rangle \subsetneq \langle b \rangle$. It we continue on factoring b, then we get a strictly bigger ideal, and if we keep going we will get ever larger ideals. The fundamental fact which we now prove (and which forces a repeated factorization to "bottom out") is that no infinite ascending sequence of ideals can exist in a PID.

Proposition 11.5.2. *Let R be a PID. There does not exist a sequence of ideals $\{I_n\}$ with*

$$I_0 \subsetneq I_1 \subsetneq I_2 \subsetneq \dots$$

Proof. Suppose that we have a sequence of ideals $\{I_n\}$ with

$$I_0 \subsetneq I_1 \subsetneq I_2 \subsetneq \dots$$

Let

$$J = \bigcup_{n=0}^{\infty} I_n = \{a \in R : \text{There exists } n \in \mathbb{N} \text{ with } a \in I_n\}$$

We claim that J is an ideal of R. We check the requirements.

- Notice that $0 \in J$ because $0 \in I_0$.
- Let $a, b \in J$. Fix $k, m \in \mathbb{N}$ with $a \in I_k$ and $b \in I_m$. Let $n = \max\{k, m\}$. We then have that $I_k \subseteq I_n$ and $I_m \subseteq I_n$, hence $a, b \in I_n$. Since I_n is an ideal of R, it follows that $a + b \in I_n$ and thus $a + b \in J$.

- Let $a \in J$ and $r \in R$. Fix $n \in \mathbb{N}$ with $a \in I_n$. Since I_n is an ideal of R, we have $ra \in I_n$ and hence $ra \in J$.

Therefore, J is an ideal of R. Since R is a PID, we may fix $b \in R$ with $J = \langle b \rangle$. We then have that $b \in J$, so we may fix $n \in \mathbb{N}$ with $b \in I_n$. Since $J = \langle b \rangle$ and I_n is an ideal containing b, we have $J \subseteq I_n$. Now we trivially have $I_n \subseteq J$, so putting it together we conclude that $I_n = J$. We then have that $I_n \subseteq I_{n+1}$ and $I_{n+1} \subseteq J = I_n$, so $I_n = I_{n+1}$, a contradiction. □

For our current goals, the preceding proposition is exactly what we need to prove that every nonzero nonunit in a PID can be factored into irreducibles. However, the property of not having such a sequence of ideals is an extremely powerful and interesting property of a ring which can be isolated and studied on its own. Rings which have no such infinite strictly ascending sequence are called *Noetherian* rings, and play a central role in a more advanced study of ring theory. Resisting the temptation to go down that path and extol the virtues of Noetherian rings, lets move on to our primary goal.

Theorem 11.5.3. *Every PID is a UFD.*

Proof. Let R be a PID. We first prove property 1. Suppose that $a \in R$ is a nonzero nonunit which is not a product of irreducible elements. We define a sequence of elements $d_1, d_2, \ldots$ in R as follows. Start by letting $d_1 = a$. Assume inductively that d_n is a nonzero nonunit which is not a product of irreducibles. In particular, d_n is itself not irreducible, so we may write $d_n = bc$ for some choice of nonzero nonunits b and c. Now it is not possible that both b and c are products of irreducibles because otherwise d_n would be as well. Thus, we may let d_{n+1} be one of b and c, chosen so that d_{n+1} is also not a product of irreducibles. Notice that d_{n+1} is a nonzero nonunit, that $d_{n+1} \mid d_n$, and d_{n+1} is not an associate of d_n because neither b nor c are units. Therefore,

$$\langle d_1 \rangle \subsetneq \langle d_2 \rangle \subsetneq \ldots$$

contradicting the above proposition. It follows that every nonzero nonunit is a product of irreducibles.

We next prove property 2. We prove the more general form that if $p_1 p_2 \cdots p_n = u q_1 q_2 \cdots q_m$ where each p_i and q_j are irreducible and u is a unit, then $m = n$ and there exists a permutation $\sigma \in S_n$ such that p_i and $q_{\sigma(i)}$ are associates for every i. Notice first that each p_i and q_j is prime because they are each irreducible and R is assumed to be a PID. Now since $p_1 \mid u q_1 q_2 \ldots q_m$ and p_1 is prime it follows that p_1 divides some factor on the right. Now p_1 is not a unit, so $p_1 \nmid u$, and thus p_1 must divide some q_k. By reordering (applying a permutation), we may assume without loss of generality that $p_1 \mid q_1$. Now the only divisor of an irreducible elements are units and associates, so it must be the case that p_1 and q_1 are associates. Fix a unit w with $q_1 = p_1 w$. We then have

$$p_1 p_2 \cdots p_n = u(p_1 w) q_2 \cdots q_m = p_1 (uw) q_2 \cdots q_m$$

Since we are in an integral domain, we may cancel the common factor of p_1 to conclude that

$$p_2 \cdots p_n = (uw) q_2 \cdots q_m$$

Repeating this process (or using induction), we conclude that $n = m$ and that we can pair off elements p_i with associates q_j. □

11.6 Irreducible Polynomials

Let F be a field. Every polynomial in $F[x]$ of degree 1 is irreducible by a simple degree argument (you proved this on the homework as a base case when you proved that every nonconstant polynomial in $F[x]$ is a product of irreducibles). A necessary condition for a polynomial of degree at least 2 to be irreducible is given by the next proposition.

Proposition 11.6.1. *Let F be a field and let $f(x) \in F[x]$ be a nonzero polynomial with $\deg(f(x)) \geq 2$. If $f(x)$ has a root in F, then $f(x)$ is not irreducible in $F[x]$.*

Proof. If $f(x)$ has a root a, then $(x-a) \mid f(x)$. Fixing $g(x) \in F[x]$ with $f(x) = (x-a) \cdot g(x)$. We then have

$$\deg(f(x)) = \deg(x-a) + \deg(g(x)) = 1 + \deg(g(x))$$

so $\deg(g(x)) = \deg(f(x)) - 1 \geq 1$. Now the units of $F[x]$ are the nonzero constants, so we have factored $f(x)$ as the product of two nonunits, and hence $f(x)$ is not irreducible in $F[x]$. □

Unfortunately, the test for the existence of roots is not in general sufficient to guarantee that a polynomial is irreducible, but it is enough in the special case where the polynomial has degree either 2 or 3.

Proposition 11.6.2. *Let F be a field and let $f(x) \in F[x]$ be a polynomial with either $\deg(f(x)) = 2$ or $\deg(f(x)) = 3$. If $f(x)$ has no roots in $F[x]$, then $f(x)$ is irreducible in $F[x]$.*

Proof. We prove the contrapositive. Suppose conversely that $f(x) \in F[x]$ is not irreducible. Write $f(x) = g(x)h(x)$ where $g(x), h(x) \in F[x]$ are nonunits. We have

$$\deg(f(x)) = \deg(g(x)) + \deg(h(x))$$

Now $g(x)$ and $h(x)$ are not units, so they each have degree at least 1. Since $\deg(f(x)) \in \{2, 3\}$, it follows that at least one of $g(x)$ or $h(x)$ has degree equal to 1. Suppose without loss of generality that $\deg(g(x)) = 1$ and write $g(x) = ax + b$ where $a, b \in F$ with $a \neq 0$. We then have

$$\begin{aligned} f(-ba^{-1}) &= g(-ba^{-1}) \cdot h(-ba^{-1}) \\ &= (a \cdot (-ba^{-1}) + b) \cdot h(-ba^{-1}) \\ &= (-b + b) \cdot h(-ba^{-1}) \\ &= 0 \cdot h(-ba^{-1}) \\ &= 0 \end{aligned}$$

so $-ba^{-1}$ is a root of $f(x)$. □

For example, consider the polynomial $f(x) = x^3 - 2$ over $\mathbb{Q}$. We know that $f(x)$ has no roots in $\mathbb{Q}$ because $\pm\sqrt[3]{2}$ are not rational by Theorem 2.5.13. Thus, $f(x)$ is irreducible over $\mathbb{Q}$. Notice that $f(x)$ is not irreducible when viewed as an element of $\mathbb{R}[x]$ because it has a root in $\mathbb{R}$. Moreover, no polynomial in $\mathbb{R}[x]$ of odd degree is irreducible because every such polynomial has a root (this uses the Intermediate Value Theorem because as $x \to \pm\infty$, on one side you must have $f(x) \to \infty$ and on the other you must have $f(x) \to -\infty$). In fact, it turns out that every irreducible polynomial over $\mathbb{R}$ has degree either 1 or 2, though this is far from obvious at this point since there are certainly polynomials of degree 4 with no root, such as $x^4 + 1$.

We spend most of the rest of this section gaining some understanding of irreducible polynomials in $\mathbb{Q}[x]$. Notice that every element of $\mathbb{Q}[x]$ is an associate with a polynomial in $\mathbb{Z}[x]$ because we can simply multiply through by the product of all denominators (which is a unit because it is a constant polynomial). Thus, up to associates, it suffices to examine polynomials with integer coefficients. We begin with a simple but important result above potential rational roots of such a polynomial.

Theorem 11.6.3. *Suppose that $f(x) \in \mathbb{Z}[x]$ is a nonzero polynomial and write*

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

where each $a_i \in \mathbb{Z}$ and $a_n \neq 0$. Suppose that $q \in \mathbb{Q}$ is a root of $f(x)$. If $q = \frac{b}{c}$ where $b \in \mathbb{Z}$ and $c \in \mathbb{N}^+$ are relatively prime (so we write q in "lowest terms"), then $b \mid a_0$ and $c \mid a_n$.

Proof. We have

$$a_n \cdot (b/c)^n + a_{n-1} \cdot (b/c)^{n-1} + \cdots + a_1 \cdot (b/c) + a_0 = 0$$

Multiplying through by c^n we get

$$a_n b^n + a_{n-1} b^{n-1} c + \cdots + a_1 b c^{n-1} + a_0 c^n = 0$$

From this, we see that

$$a_n b^n = c \cdot [-(a_{n-1} b^{n-1} + \cdots + a_1 b c^{n-2} + a_0 c^{n-1})]$$

and hence $c \mid a_n b^n$. Using the fact that $\gcd(b, c) = 1$, it follows that $c \mid a_n$ (either by repeatedly apply Proposition 2.4.10, or use the Fundamental Theorem of Arithmetic to note that the prime factors of c do not appear in the prime factorization of b^n). On the other hand, we see that

$$a_0 c^n = b \cdot [-(a_n b^{n-1} + a_{n-1} b^{n-2} c + \cdots + a_1 c^{n-1})]$$

and hence $b \mid a_0 c^n$. Using the fact that $\gcd(b, c) = 1$, it follows as above that $b \mid a_0$. □

As an example, we show that the polynomial $f(x) = 2x^3 - x^2 + 7x - 9$ is irreducible in $\mathbb{Q}[x]$. By the previous theorem, the only possible roots of $f(x)$ in $\mathbb{Q}$ are ± 9, ± 3, ± 1, $\pm\frac{9}{2}$, $\pm\frac{3}{2}$, $\pm\frac{1}{2}$. We check the values:

- $f(9) = 1431$ and $f(-9) = -1611$
- $f(3) = 57$ and $f(-3) = -93$
- $f(1) = -1$ and $f(-1) = -19$
- $f(\frac{9}{2}) = \frac{369}{2}$ and $f(-\frac{9}{2}) = -243$
- $f(\frac{1}{2}) = -\frac{11}{2}$ and $f(-\frac{1}{2}) = -13$
- $f(\frac{3}{2}) = 6$ and $f(-\frac{3}{2}) = -\frac{57}{2}$

Thus, $f(x)$ has no roots in $\mathbb{Q}$. Since $\deg(f(x)) = 3$, it follows that $f(x)$ is irreducible in $\mathbb{Q}[x]$. Notice that $f(x)$ is *not* irreducible in $\mathbb{R}[x]$ because it has a root between in the interval $(1, \frac{3}{2})$ by the Intermediate Value Theorem.

As we mentioned above, we are focusing on polynomials in $\mathbb{Q}[x]$ which have integer coefficients. However, even if $f(x) \in \mathbb{Z}[x]$, when checking for irreducibility in $\mathbb{Q}[x]$, we have to consider the possibility that a potential factorization involves polynomials whose coefficients are fractions. For example, we have

$$x^2 = (2x) \cdot (\frac{1}{2}x)$$

Of course, in this case there also exists a factorization into smaller degree degree polynomials in $\mathbb{Z}[x]$ because we can write $x^2 = x \cdot x$. Our first task is to prove that this is always the case. We will need the following lemma.

Lemma 11.6.4. *Suppose that $g(x), h(x) \in \mathbb{Z}[x]$ and that $p \in \mathbb{Z}$ is a prime which divides all coefficients of $g(x)h(x)$. We then have that either p divides all coefficients of $g(x)$, or p divides all coefficients of $h(x)$.*

Proof. Let $g(x)$ be the polynomial $\{b_n\}$, let $h(x)$ be the polynomial $\{c_n\}$, and let $g(x)h(x)$ be the polynomial $\{a_n\}$. We are supposing that $p \mid a_n$ for all n. Suppose the $p \nmid b_n$ for some n and also that $p \nmid c_n$ for some n (possibly different). Let k be least such that $p \nmid b_k$, and let ℓ be least such that $p \nmid c_\ell$. Notice that

$$a_{k+\ell} = \sum_{i=0}^{k+\ell} b_i c_{k+\ell-i} = b_k c_\ell + (\sum_{i=0}^{k-1} b_i c_{k+\ell-i}) + (\sum_{i=k+1}^{k+\ell} b_i c_{k+\ell-i})$$

hence

$$b_k c_\ell = a_{k+\ell} - (\sum_{i=0}^{k-1} b_i c_{k+\ell-i}) - (\sum_{i=k+1}^{k+\ell} b_i c_{k+\ell-i})$$

Now if $0 \leq i \leq k-1$, then $p \mid b_i$ by choice of k, hence $p \mid b_i c_{k+\ell-i}$. Also, if $k+1 \leq i \leq k+\ell$, then $k+\ell-i < \ell$, so $p \mid c_{k+\ell-i}$ by choice of ℓ, hence $p \mid b_i c_{k+\ell-i}$. Since $p \mid a_{k+\ell}$ by assumption, it follows that p divides every summand on the right hand side. Therefore, p divides the right hand side, and thus $p \mid b_k c_\ell$. Since p is prime, it follows that either $p \mid b_k$ or $p \mid c_\ell$, but both of these are impossible by choice of k and ℓ. Therefore, it must be the case that either $p \mid b_n$ for all n, or $p \mid c_n$ for all n. □

Proposition 11.6.5 (Gauss' Lemma)**.** *Suppose that $f(x) \in \mathbb{Z}[x]$ and that $g(x), h(x) \in \mathbb{Q}[x]$ with $f(x) = g(x)h(x)$. There exist polynomials $g^*(x), h^*(x) \in \mathbb{Z}[x]$ such that $f(x) = g^*(x)h^*(x)$ and both $\deg(g^*(x)) = \deg(g(x))$ and $\deg(h^*(x)) = \deg(h(x))$. In fact, there exist nonzero $s, t \in \mathbb{Q}$ with*

- $f(x) = g^*(x)h^*(x)$
- $g^*(x) = s \cdot g(x)$
- $h^*(x) = t \cdot h(x)$

Proof. If each of the coefficients of $g(x)$ and $h(x)$ happen to be integers, then we are happy. Suppose not. Let $a \in \mathbb{Z}$ be the least common multiple of the denominators of the coefficients of g, and let $b \in \mathbb{Z}$ be the least common multiple of the denominators of the coefficients of g. Let $d = ab$. Multiply both sides of $f(x) = g(x)h(x)$ through by d to "clear denominators" gives

$$d \cdot f(x) = (a \cdot g(x)) \cdot (b \cdot h(x))$$

where each of the three factors $d \cdot f(x)$, $a \cdot g(x)$, and $b \cdot h(x)$ is a polynomial in $\mathbb{Z}[x]$. We have at least one of $a > 1$ or $b > 1$, hence $d = ab > 1$.

Fix a prime divisor p of d. We then have that p divides all coefficients of $d \cdot f(x)$, so by the previous lemma either p divides all coefficients of $a \cdot g(x)$, or p divides all coefficients of $b \cdot h(x)$. In the former case, we have

$$\frac{d}{p} \cdot f(x) = (\frac{a}{p} \cdot g(x)) \cdot (b \cdot h(x))$$

where each of the three factors is a polynomial in $\mathbb{Z}[x]$. In the latter case, we have

$$\frac{d}{p} \cdot f(x) = (a \cdot g(x)) \cdot (\frac{b}{p} \cdot h(x))$$

where each of the three factors is a polynomial in $\mathbb{Z}[x]$. Now if $\frac{d}{p} = 1$, then we are done by letting $g^*(x)$ be the first factor and letting $h^*(x)$ be the second. Otherwise, we continue the argument by dividing out another prime factor of $\frac{d}{p}$ from all coefficients of one of the two polynomials. Continue until we have handled all prime which occur in a factorization of d. Formally, you can do induction on d. □

An immediate consequence of Gauss' Lemma is the following, which greatly simplifies the check for whether a given polynomial with integer coefficients is irreducible in $\mathbb{Q}[x]$.

Corollary 11.6.6. *Let $f(x) \in \mathbb{Z}[x]$. If there does not exist nonconstant polynomials $g(x), h(x) \in \mathbb{Z}[x]$ with $f(x) = g(x) \cdot h(x)$, then $f(x)$ is irreducible in $\mathbb{Q}[x]$. Furthermore, if $f(x)$ is monic, then it suffices to show that no such monic $g(x)$ and $h(x)$ exist.*

Proof. The first part is immediate from Proposition 11.2.2 and Gauss' Lemma. Now suppose that $f(x) \in \mathbb{Z}[x]$ is monic. Suppose that $g(x), h(x) \in \mathbb{Z}$ with $f(x) = g(x)h(x)$. Notice that the leading term of $f(x)$ is the product of the leading terms of $g(x)$ and $h(x)$, so as $f(x)$ is monic and all coefficients are in $\mathbb{Z}$, either both $g(x)$ and $h(x)$ are monic or both have leading terms -1. In the latter case, we can multiply both through by -1 to get a factorization into monic polynomials in $\mathbb{Z}[x]$ of the same degree. □

As an example, consider the polynomial $f(x) = x^4 + 3x^3 + 7x^2 - 9x + 1 \in \mathbb{Q}[x]$. We claim that $f(x)$ is irreducible in $\mathbb{Q}[x]$. We first check for rational roots. We know that the only possibilities are ± 1 and we check these:

- $f(1) = 1 + 3 + 7 - 9 + 1 = 3$
- $f(-1) = 1 - 3 + 7 + 9 + 1 = 15$

Thus, $f(x)$ has no rational roots.

By the corollary, it suffices to show that $f(x)$ is not the product of two monic nonconstant polynomials in $\mathbb{Z}[x]$. Notice that $f(x)$ has no monic divisor of degree 1 because such a divisor would imply that $f(x)$ has a rational root, which we just ruled out. Thus, we need only consider the possibility that $f(x)$ is the product of two monic polynomials in $\mathbb{Z}[x]$ of degree 2. Consider a factorization:

$$f(x) = (x^2 + ax + b)(x^2 + cx + d)$$

where $a, b, c, d \in \mathbb{Z}$. We then have

$$x^4 + 3x^3 + 7x^2 - 9x + 1 = x^4 + (a + c)x^3 + (b + ac + d)x^2 + (ad + bc)x + bd$$

We therefore have the following equations

1. $a + c = 3$
2. $b + ac + d = 7$
3. $ad + bc = -9$
4. $bd = 1$

Since $bd = 1$ and $b, d \in \mathbb{Z}$, we have that $b \in \{1, -1\}$. We check the possibilities.

- Suppose that $b = 1$. By equation 4, we conclude that $d = 1$. Thus, by equation 3, we conclude that $a + c = -9$, but this contradicts equation 1.
- Suppose that $b = -1$. By equation 4, we conclude that $d = -1$. Thus, by equation 3, we conclude that $-a - c = -9$, so $a + c = 9$, but this contradicts equation 1.

In all cases, we have reached a contradiction. We conclude that $f(x)$ is irreducible over $\mathbb{Q}$.

The following theorem, when it applies, is a simple way to determine that certain polynomials in $\mathbb{Z}[x]$ are irreducible in $\mathbb{Q}[x]$. Although it has limited general use (a polynomial taken at random typically does not satisfy the hypotheses), it is surprisingly useful how often it applies to "natural" polynomials you want to verify are irreducible.

Theorem 11.6.7 (Eisenstein's Criterion)**.** *Suppose that $f(x) \in \mathbb{Z}[x]$ and write*

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

If there exists a prime $p \in \mathbb{N}^+$ such that

- $p \nmid a_n$
- $p \mid a_i$ *for* $0 \leq i \leq n-1$
- $p^2 \nmid a_0$

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

Proof. Fix such a prime p. We use Corollary 11.6.6. Suppose that $g(x), h(x) \in \mathbb{Z}[x]$ are not constant polynomials with $f(x) = g(x)h(x)$. We then have

$$n = \deg(f(x)) = \deg(g(x)) + \deg(h(x))$$

Since we are assuming that $g(x)$ and $h(x)$ are not constant, they each have degree at least 1, and so by the above equality they both have degree at most $n-1$.

Let $g(x)$ be the polynomial $\{b_n\}$ and let $h(x)$ be the polynomial $\{c_n\}$. We have $a_0 = b_0c_0$, so since $p \mid a_0$ and p is prime, either $p \mid b_0$ or $p \mid c_0$. Furthermore, since $p^2 \nmid a_0$ by assumption, we can not have both $p \mid b_0$ and $p \mid c_0$. Without loss of generality (by switching the roles of $g(x)$ and $h(x)$ if necessary), suppose that $p \mid b_0$ and $p \nmid c_0$.

We now prove that $p \mid b_k$ for $0 \leq k \leq n-1$ by (strong) induction. Suppose that we have k with $0 \leq k \leq n-1$ and we know that $p \mid b_i$ for $0 \leq i < k$. Now

$$a_k = b_kc_0 + b_{k-1}c_1 + \cdots + b_1c_{k-1} + b_0c_k$$

and hence

$$b_kc_0 = a_k - b_{k-1}c_1 - \cdots - b_1c_{k-1} - b_0c_k$$

By assumption, we have $p \mid a_k$, and by induction we have $p \mid b_i$ for $0 \leq i < k$. It follows that p divides every term on the right-hand side, so $p \mid b_kc_0$. Since p is prime and $p \nmid c_0$, it follows that $p \mid b_k$.

Thus, we have shown that $p \mid b_k$ for $0 \leq k \leq n-1$. Now we have

$$\begin{aligned} a_n &= b_nc_0 + b_{n-1}c_1 + \cdots + b_1c_{n-1} + b_0c_n \\ &= b_{n-1}c_1 + \cdots + b_1c_{n-1} + b_0c_n \end{aligned}$$

where the last line follows from the fact that $b_n = 0$ (since we are assuming $\deg(g(x)) < n$). Now we know $p \mid b_k$ for $0 \leq k \leq n-1$, so p divides every term on the right. It follows that $p \mid a_n$, contradicting our assumption. Therefore, by Corollary 11.6.6, $f(x)$ is irreducible in $\mathbb{Q}[x]$. □

For example, the polynomial $x^4+6x^3+27x^2-297x+24$ is irreducible in $\mathbb{Q}[x]$ using Eisenstein's Criterion with $p = 3$. For each $n \in \mathbb{N}^+$ the polynomial $x^n - 2$ is irreducible in $\mathbb{Q}[x]$ using Eisenstein's Criterion with $p = 2$. In particular, we have shown that $\mathbb{Q}[x]$ has irreducible polynomials of every positive degree.

11.7 Quotients of $F[x]$

We have now carried over many of the concepts and proofs that originated in the study of the integers $\mathbb{Z}$ over to the polynomial ring over a field $F[x]$ (and more generally any Euclidean Domains or even PID). One construction we have not focused on is the quotient construction. In the integers $\mathbb{Z}$, we know that every nonzero ideal has the form $n\mathbb{Z} = \langle n \rangle$ for some $n \in \mathbb{N}^+$ (because after all $\mathbb{Z}$ is a PID and we have $\langle n \rangle = \langle -n \rangle$). The quotients of $\mathbb{Z}$ by these ideals are the now very familiar rings $\mathbb{Z}/n\mathbb{Z}$.

Since we have so much in common between $\mathbb{Z}$ and $F[x]$, we would like to study the analogous quotients of $F[x]$. Now if F is a field, then $F[x]$ is a PID, so for every nonzero ideal of $F[x]$ equals $\langle p(x) \rangle$ for some $p(x) \in F[x]$. Suppose then that $p(x) \in F[x]$ and let $I = \langle p(x) \rangle$. Given $f(x) \in F[x]$, we will write $\overline{f(x)}$ for

the coset $f(x)+I$ just like we wrote $\overline{k}$ for the coset $k+n\mathbb{Z}$. As long as we do not change the ideal I (so do not change $p(x)$) in a given construction, there should be no confusion as to which quotient we are working in. When dealing with these quotients, our first task is to find unique representatives for the cosets as in $\mathbb{Z}/\mathbb{Z}$ where we saw that $\overline{0}, \overline{1}, \ldots, \overline{n-1}$ served as distinct representatives for the cosets.

Proposition 11.7.1. *Let F be a field and let $p(x) \in F[x]$ be nonzero. Let $I = \langle p(x) \rangle$ and work in $F[x]/I$. For all $f(x) \in F[x]$, there exists a unique $h(x) \in F[x]$ such that both:*

- $\overline{f(x)} = \overline{h(x)}$
- *Either $h(x) = 0$ or $\deg(h(x)) < \deg(p(x))$*

In other words, if we let

$$S = \{h(x) \in F[x] : h(x) = 0 \text{ or } \deg(h(x)) < \deg(p(x))\}$$

then the elements of S provide unique representatives for the cosets in $F[x]/I$.

Proof. We first prove existence. Let $f(x) \in F[x]$. Since $p(x) \neq 0$, we may fix $q(x), r(x)$ with

$$f(x) = q(x)p(x) + r(x)$$

and either $r(x) = 0$ or $\deg(r(x)) < \deg(p(x))$. We then have

$$p(x)q(x) = f(x) - r(x)$$

Thus $p(x) \mid (f(x) - r(x))$ and so $f(x) - r(x) \in I$. It follows from Proposition 10.4.5 that $\overline{f(x)} = \overline{r(x)}$ so we may take $h(x) = r(x)$. This proves existence.

We now prove uniqueness. Suppose that $h_1(x), h_2(x) \in S$ (so each is either 0 or has smaller degree than $p(x)$) and that $\overline{h_1(x)} = \overline{h_2(x)}$. Using Proposition 10.4.5, we then have that $h_1(x) - h_2(x) \in I$ and hence $p(x) \mid (h_1(x) - h_2(x))$. Notice that every nonzero multiple of $p(x)$ has degree greater than or equal to $\deg(p(x))$ (since the degree of a product is the sum of the degrees in $F[x]$). Now either $h_1(x) - h_2(x) = 0$ or it has degree less than $\deg(p(x))$, but we've just seen that the latter is impossible. Therefore, it must be the case that $h_1(x) - h_2(x) = 0$, and so $h_1(x) = h_2(x)$. This proves uniqueness. □

Let's look at an example. Suppose that we are working with $F = \mathbb{Q}$ and we let $p(x) = x^2 - 2x + 3$. Consider the quotient ring $R = \mathbb{Q}[x]/\langle p(x) \rangle$. From above, we know that every element in this quotient is represented uniquely by either a constant polynomial or a polynomial of degree 1. Thus, some distinct elements of R are $\overline{1}$, $\overline{3/7}$, $\overline{x}$, and $\overline{2x - 5/3}$. We add elements in the quotient ring R by adding representatives as usual, so for example we have

$$\overline{4x - 7} + \overline{2x + 8} = \overline{6x + 1}$$

Multiplication of elements of R is more interesting if we try to convert the resulting product to one of our chosen representatives. For example, we have

$$\overline{2x + 7} \cdot \overline{x - 1} = \overline{(2x + 7)(x - 1)} = \overline{2x^2 + 5x - 7}$$

which is perfectly correct, but the resulting representative isn't one of our chosen ones. If we follows the above proof, we should divide $2x^2 + 5x - 7$ by $x^2 - 2x + 3$ and use the remainder as our representative. We have

$$2x^2 + 5x - 7 = 2 \cdot (x^2 - 2x + 3) + (9x - 13)$$

so

$$(2x^2 + 5x - 7) - (9x - 13) \in \langle p(x) \rangle$$

and hence

$$\overline{2x^2 + 5x - 7} = \overline{9x - 13}$$

It follows that in the quotient we have

$$\overline{2x + 7} \cdot \overline{x - 1} = \overline{9x - 13}$$

Here's another way to determine that the product is $\overline{9x - 13}$. Notice that for any $f(x), g(x) \in F[x]$, we have

$$\overline{f(x)} + \overline{g(x)} = \overline{f(x) + g(x)} \qquad \overline{f(x)} \cdot \overline{g(x)} = \overline{f(x) \cdot g(x)}$$

by definition of multiplication in the quotient. Now $p(x) = x^2 - 2x + 3$, so in the quotient we have $\overline{x^2 - 2x + 3} = \overline{0}$. It follows that

$$\overline{x^2} + \overline{-2x + 3} = \overline{0}$$

and hence

$$\overline{x^2} = \overline{2x - 3}$$

Therefore

$$\begin{aligned}
\overline{2x + 7} \cdot \overline{x - 1} &= \overline{2x^2 + 5x - 7} \\
&= \overline{2} \cdot \overline{x^2} + \overline{5x - 7} \\
&= \overline{2} \cdot \overline{2x - 3} + \overline{5x - 7} \\
&= \overline{4x - 6} + \overline{5x - 7} \\
&= \overline{9x - 13}
\end{aligned}$$

For another example, consider the ring $F = \mathbb{Z}/2\mathbb{Z}$. Since we will be working in quotients of $F[x]$ and too many equivalence classes begins to get confusing, we will write $F = \{0, 1\}$ rather than $F = \{\overline{0}, \overline{1}\}$. Consider the polynomial $p(x) = x^2+1 \in F[x]$. We then have that the elements of $F[x]/\langle x^2+1\rangle$ are represented uniquely by either a constant polynomial or a polynomial of degree 1. Thus, the distinct elements of $F[x]/\langle x^2+1\rangle$ are given by $\overline{ax + b}$ for $a, b \in F$. Since $|F| = 2$, we have two choices for each of a and b, and hence the quotient has four elements. Here is the additional and multiplication tables for $F[x]/\langle x^2 + 1\rangle$:

$+$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{1}$	$\overline{1}$	$\overline{0}$	$\overline{x+1}$	$\overline{x}$
$\overline{x}$	$\overline{x}$	$\overline{x+1}$	$\overline{0}$	$\overline{1}$
$\overline{x+1}$	$\overline{x+1}$	$\overline{x}$	$\overline{1}$	$\overline{0}$

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{x}$	$\overline{0}$	$\overline{x}$	$\overline{1}$	$\overline{x+1}$
$\overline{x+1}$	$\overline{0}$	$\overline{x+1}$	$\overline{x+1}$	$\overline{0}$

The addition table is fairly straightforward, but some work went into constructing the multiplication table. For example, we have

$$\overline{x} \cdot \overline{x + 1} = \overline{x^2 + x}$$

To determine which of our chosen representatives gives this coset, we notice that $\overline{x^2 + 1} = \overline{0}$ (since clearly $x^2 + 1 \in \langle x^2 + 1\rangle$), so $\overline{x^2} + \overline{1} = \overline{0}$. Adding $\overline{1}$ to both sides and noting that $\overline{1} + \overline{1} = \overline{0}$, we conclude that $\overline{x^2} = \overline{1}$. Therefore

$$\begin{aligned}
\overline{x} \cdot \overline{x + 1} &= \overline{x^2 + x} \\
&= \overline{x^2} + \overline{x} \\
&= \overline{1} + \overline{x} \\
&= \overline{x + 1}
\end{aligned}$$

The other entries of the table can be found similarly. In fact, we have done all the hard work because we now know that $\overline{x^2} = \overline{1}$ and we can use that throughout.

Suppose instead that we work with the polynomial $p(x) = x^2 + x + 1$. Thus, we are considering the quotient $F[x]/\langle x^2 + x + 1\rangle$. As above, since $\deg(x^2 + x + 1) = 2$, we get unique representatives from the elements $ax + b$ for $a, b \in \{0, 1\}$. Although we have the same representatives, the cosets are different and the multiplication table changes considerably:

$+$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{1}$	$\overline{1}$	$\overline{0}$	$\overline{x+1}$	$\overline{x}$
$\overline{x}$	$\overline{x}$	$\overline{x+1}$	$\overline{0}$	$\overline{1}$
$\overline{x+1}$	$\overline{x+1}$	$\overline{x}$	$\overline{1}$	$\overline{0}$

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{x}$	$\overline{0}$	$\overline{x}$	$\overline{x+1}$	$\overline{1}$
$\overline{x+1}$	$\overline{0}$	$\overline{x+1}$	$\overline{1}$	$\overline{x}$

For example, let's determine $\overline{x} \cdot \overline{x+1}$ in this situation. Notice first that $\overline{x^2 + x + 1} = \overline{0}$, so $\overline{x^2} + \overline{x} + \overline{1} = 0$. Adding $\overline{x} + \overline{1}$ to both sides and using the fact that $\overline{1} + \overline{1} = \overline{0}$ and $\overline{x} + \overline{x} = \overline{0}$, we conclude that

$$\overline{x^2} = \overline{x} + \overline{1} = \overline{x+1}$$

Therefore, we have

$$\begin{aligned} \overline{x} \cdot \overline{x+1} &= \overline{x^2 + x} \\ &= \overline{x^2} + \overline{x} \\ &= \overline{x} + \overline{1} + \overline{x} \\ &= \overline{1} \end{aligned}$$

Notice that every nonzero element of the quotient $F[x]/\langle x^2 + x + 1\rangle$ has a multiplicative inverse, so the quotient in this case is a field. We have succeeded in constructing of field of order 4. This is our first example of a finite field which does not have prime order.

The reason why we obtained a field when taking the quotient by $\langle x^2 + x + 1\rangle$ but not when taking the quotient by $\langle x^2 + 1\rangle$ is the following. It is the analogue of the fact that $\mathbb{Z}/p\mathbb{Z}$ is a field if and only if p is prime (equivalently irreducible) in $\mathbb{Z}$.

Proposition 11.7.2. *Let F be a field and let $p(x) \in F[x]$ be nonzero. We have that $F[x]/\langle p(x)\rangle$ is a field if and only if $p(x)$ is irreducible in $F[x]$.*

Proof. Let $p(x) \in F[x]$ be nonzero. Since F is a field, we know that $F[x]$ is a PID. Using Proposition 11.4.8 and Theorem 10.5.11, we conclude that

$$\begin{aligned} F[x]/\langle p(x)\rangle \text{ is a field} &\iff \langle p(x)\rangle \text{ is a maximal ideal of } F[x] \\ &\iff p(x) \text{ is irreducible in } F[x] \end{aligned}$$

□

When $F = \{0, 1\}$ as above, we see that $x^2 + 1$ is not irreducible (since 1 is a root or $x^2 + 1 = (x+1)(x+1)$ but $x^2 + x + 1$ is irreducible (because it has degree 2 and neither 0 nor 1 is a root). Generalizing the previous constructions, we get the following.

Proposition 11.7.3. *Let F be a finite field with k elements. If $p(x) \in F[x]$ is irreducible and $\deg(p(x)) = n$, then $F[x]/\langle p(x)\rangle$ is a field with k^n elements.*

Proof. Since $p(x)$ is irreducible, we know from the previous proposition that $F[x]/\langle p(x)\rangle$ is a field. Since $\deg(p(x)) = n$, we can represent the elements of the quotient uniquely by elements of the form

$$\overline{a_{n-1}x^{n-1} + \cdots + a_1x + a_0}$$

where each $a_i \in F$. Now F has k elements, so we have k choices for each value of a_i. We can make this choice for each of the n coefficients a_i, so we have k^n many choices in total. □

It turns out that if $p \in \mathbb{N}^+$ is prime and $n \in \mathbb{N}^+$, then there exists an irreducible polynomial in $\mathbb{Z}/p\mathbb{Z}[x]$ of degree n, so there exists a field of order p^n. However, directly proving that such polynomials exist is difficult, and one often constructs them using different techniques. Furthermore, every finite field has order some prime power, and any two finite fields of the same order are isomorphic. Consult your local field theory course.

Finally, we end this section with one way to construct the complex numbers. Consider the ring $\mathbb{R}[x]$ of polynomials with real coefficients. Let $p(x) = x^2 + 1$ and notice that $p(x)$ has no roots in $\mathbb{R}$ because $a^2 + 1 \geq 1$ for all $a \in \mathbb{R}$. Since $\deg(p(x)) = 2$, it follows that $p(x) = x^2 + 1$ is irreducible in $\mathbb{R}[x]$. From above, we conclude that $\mathbb{R}[x]/\langle x^2 + 1\rangle$ is a field. Now elements of the quotient are represented uniquely by $\overline{ax + b}$ for $a, b \in \mathbb{R}$. We have $\overline{x^2 + 1} = \overline{0}$, so $\overline{x^2} + \overline{1} = \overline{0}$ and hence $\overline{x^2} = \overline{-1}$. It follows that $\overline{x}^2 = \overline{-1}$, so $\overline{x}$ can play the role of "i". Notice that for any $a, b, c, d \in \mathbb{R}$ we have

$$\overline{ax + b} + \overline{cx + d} = \overline{(a + c)x + (b + d)}$$

and

$$\begin{aligned}
\overline{ax + b} \cdot \overline{cx + d} &= \overline{acx^2 + (ad + bc)x + bd} \\
&= \overline{acx^2} + \overline{(ad + bc)x} + \overline{bd} \\
&= \overline{ac} \cdot \overline{x^2} + \overline{(ad + bc)x} + \overline{bd} \\
&= \overline{ac} \cdot \overline{-1} + \overline{(ad + bc)x} + \overline{bd} \\
&= \overline{(ad + bc)x + (bd - ac)}
\end{aligned}$$

Notice that this multiplication is the exact same as when you treat the complex numbers as having the form $ai + b$ and "formally add and multiply" using the rule that $i^2 = -1$. One advantage of our quotient construction is that we do not need to verify all of the field axioms. We get them for free from our general theory.

11.8 Field of Fractions

Let R be an integral domain. In this section, we show that R can be embedded in a field F. Furthermore, our construction gives a "smallest" such field F in a sense to be made precise below. We call this field the *field of fractions* of R and denote it $Frac(R)$. Our method for building $Frac(R)$ generalizes the construction of the rationals from the integers we outlined in Sections 3.1 and 3.2, and in particular we will carry out all of the necessary details that we omitted there. Notice that the fact that every integral domain R can be embedded in a field "explains" why we have cancellation in integral domains (because when viewed in the larger field $Frac(R)$, we can multiply both sides by the multiplicative inverse).

Definition 11.8.1. *Let R be an integral domain. Define $P = R \times (R\backslash\{0\})$. We define a relation $\sim$ on P by letting $(a, b) \sim (c, d)$ if $ad = bc$.*

Proposition 11.8.2. *$\sim$ is an equivalence relation on P.*

Proof. We check the properties:

- Reflexive: For any $a, b \in R$ with $b \neq 0$, we have $(a, b) \sim (a, b)$ because $ab = ba$, so $\sim$ is reflexive.
- Symmetric: Suppose that $a, b, c, d \in R$ with $b, d \neq 0$ and $(a, b) \sim (c, d)$. We then have $ad = bc$, hence $cb = bc = ad = da$. It follows that $(c, d) \sim (a, b)$.
- Transitive: Suppose that $a, b, c, d, e, f \in \mathbb{Z}$ with $b, d, f \neq 0$, $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. We then have $ad = bc$ and $cf = de$. Hence,

$$\begin{aligned}(af)d &= (ad)f \\ &= (bc)f && \text{(since } ad = bc\text{)} \\ &= b(cf) \\ &= b(de) && \text{(since } cf = de\text{)} \\ &= (be)d\end{aligned}$$

  Therefore, $af = be$ because R is an integral domain and $d \neq 0$.

□

We now define $Frac(R)$ to be the set of equivalence classes, i.e. $Frac(R) = P/\sim$. We need to define addition and multiplication on F to make it into a field. Mimicking addition and multiplication of rationals, we want to define

$$\overline{(a, b)} + \overline{(c, d)} = \overline{(ad + bc, bd)} \qquad \overline{(a, b)} \cdot \overline{(c, d)} = \overline{(ac, bd)}$$

Notice first that since $b, d \neq 0$, we have $bd \neq 0$ because R is an integral domain, so we have no issues there. However, we need to check that the operations are well-defined.

Proposition 11.8.3. *Let $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 \in R$ with $b_1, b_2, d_1, d_2 \neq 0$. Suppose that $(a_1, b_1) \sim (a_2, b_2)$ and $(c_1, d_1) \sim (c_2, d_2)$. We then have*

- $(a_1d_1 + b_1c_1, b_1d_1) \sim (a_2d_2 + b_2c_2, b_2d_2)$
- $(a_1c_1, b_1d_1) \sim (a_2c_2, b_2d_2)$

Thus, the above operations of $+$ and $\cdot$ on $Frac(R)$ are well-defined.

Proof. Since $(a_1, b_1) \sim (a_2, b_2)$ and $(c_1, d_1) \sim (c_2, d_2)$, it follows that $a_1b_2 = b_1a_2$ and $c_1d_2 = d_1c_2$.

We have

$$\begin{aligned}(a_1d_1 + b_1c_1)b_2d_2 &= a_1d_1b_2d_2 + b_1c_1b_2d_2 \\ &= (a_1b_2)d_1d_2 + (c_1d_2)b_1b_2 \\ &= (b_1a_2)d_1d_2 + (d_1c_2)b_1b_2 \\ &= b_1d_1a_2d_2 + b_1d_2b_2c_2 \\ &= b_1d_1(a_2d_2 + b_2c_2)\end{aligned}$$

so $(a_1d_1 + b_1c_1, b_1d_1) \sim (a_2d_2 + b_2c_2, b_2d_2)$.

Using $a_1b_2 = b_1a_2$ and $c_1d_2 = d_1c_2$, We have $(a_1b_2)(c_1d_2) = (b_1a_2)(d_1c_2)$, and hence $(a_1c_1)(b_2d_2) = (b_1d_1)(a_2c_2)$. Therefore, $(a_1c_1, b_1d_1) \sim (a_2c_2, b_2d_2)$. □

Now that we have successfully defined addition and multiplication, we are ready to prove that the resulting object is a field.

Theorem 11.8.4. *If R is an integral domain, then $Frac(R)$ is a field with the following properties.*

- *The additive identity of $Frac(R)$ is $\overline{(0,1)}$.*
- *The multiplicative identity of $Frac(R)$ is $\overline{(1,1)}$.*
- *The additive inverse of $\overline{(a,b)} \in Frac(R)$ is $\overline{(-a,b)}$.*
- *If $\overline{(a,b)} \in Frac(R)$ with $\overline{(a,b)} \neq \overline{(0,1)}$, then $a \neq 0$ and the multiplicative inverse of $\overline{(a,b)}$ is $\overline{(b,a)}$.*

Proof. We check each of the field axioms.

1. Associativity of $+$: Let $q, r, s \in Frac(R)$. Fix $a, b, c, d, e, f \in R$ with $b, d, f \neq 0$ such that $q = \overline{(a,b)}$, $r = \overline{(c,d)}$, and $s = \overline{(e,f)}$. We then have

$$\begin{aligned} q + (r + s) &= \overline{(a,b)} + (\overline{(c,d)} + \overline{(e,f)}) \\ &= \overline{(a,b)} + \overline{(cf + de, df)} \\ &= \overline{(a(df) + b(cf + de), b(df))} \\ &= \overline{(adf + bcf + bde, bdf)} \\ &= \overline{((ad + bc)f + (bd)e, (bd)f)} \\ &= \overline{(ad + bc, bd)} + \overline{(e,f)} \\ &= \overline{[(a,b)} + \overline{(c,d)]} + \overline{(e,f)} \\ &= (q + r) + s \end{aligned}$$

2. Commutativity of $+$: Let $q, r, s \in Frac(R)$. Fix $a, b, c, d \in R$ with $b, d \neq 0$ such that $q = \overline{(a,b)}$, and $r = \overline{(c,d)}$. We then have

$$\begin{aligned} q + r &= \overline{(a,b)} + \overline{(c,d)} \\ &= \overline{(ad + bc, bd)} \\ &= \overline{(cb + da, db)} \\ &= \overline{(c,d)} + \overline{(a,b)} \\ &= r + q \end{aligned}$$

3. $\overline{(0,1)}$ is an additive identity: Let $q \in Frac(R)$. Fix $a, b \in R$ with $b \neq 0$ such that $q = \overline{(a,b)}$. We then have

$$\begin{aligned} q + \overline{(0,1)} &= \overline{(a,b)} + \overline{(0,1)} \\ &= \overline{(a \cdot 1 + b \cdot 0, b \cdot 1)} \\ &= \overline{(a,b)} \\ &= q \end{aligned}$$

   Since we already proved commutativity of $+$, we conclude that $\overline{(0,1)} + q = q$ also.

4. Additive inverses: Let $q \in Frac(R)$. Fix $a, b \in R$ with $b \neq 0$ such that $q = \overline{(a,b)}$. Let $r = \overline{(-a,b)}$. We then have

$$\begin{aligned} q + r &= \overline{(a,b)} + (-\overline{(a,b)}) \\ &= \overline{(a,b)} + \overline{(-a,b)} \\ &= \overline{(ab + (-ab), bb)} \\ &= \overline{(a(b + (-b)), bb)} \\ &= \overline{(a \cdot 0, bb)} \\ &= \overline{(0, bb)} \\ &= \overline{(0,1)} \end{aligned}$$

where the last line follows from the fact that $0 \cdot 1 = 0 = 0 \cdot bb$. Since we already proved commutativity of $+$, we conclude that $r + q = \overline{(0,1)}$ also.

5. Associativity of $\cdot$: Let $q, r, s \in Frac(R)$. Fix $a, b, c, d, e, f \in R$ with $b, d, f \neq 0$ such that $q = \overline{(a,b)}$, $r = \overline{(c,d)}$, and $s = \overline{(e,f)}$. We then have

$$\begin{aligned} q \cdot (r \cdot s) &= \overline{(a,b)} \cdot (\overline{(c,d)} \cdot \overline{(e,f)}) \\ &= \overline{(a,b)} \cdot \overline{(ce, df)} \\ &= \overline{(a(ce), b(df))} \\ &= \overline{((ac)e, (bd)f)} \\ &= \overline{(ac, bd)} \cdot \overline{(e,f)} \\ &= (\overline{(a,b)} \cdot \overline{(c,d)}) \cdot \overline{(e,f)} \\ &= (q \cdot r) \cdot s \end{aligned}$$

6. Commutativity of $\cdot$: Let $q, r, s \in Frac(R)$. Fix $a, b, c, d \in R$ with $b, d \neq 0$ such that $q = \overline{(a,b)}$, and $r = \overline{(c,d)}$. We then have

$$\begin{aligned} q \cdot r &= [(a,b)] \cdot [(c,d)] \\ &= [(ac, bd)] \\ &= [(ca, db)] \\ &= [(c,d)] \cdot [(a,b)] \\ &= r \cdot q \end{aligned}$$

7. $\overline{(1,1)}$ is a multiplicative identity: Let $q \in Frac(R)$. Fix $a, b \in R$ with $b \neq 0$ such that $q = \overline{(a,b)}$. We then have

$$\begin{aligned} q \cdot \overline{(1,1)} &= \overline{(a,b)} \cdot \overline{(1,1)} \\ &= \overline{(a \cdot 1, b \cdot 1)} \\ &= \overline{(a,b)} \\ &= q \end{aligned}$$

Since we already proved commutativity of $\cdot$, we conclude that $\overline{(1,1)} \cdot q = q$ also.

8. Multiplicative inverses: Let $q \in Frac(R)$ with $q \neq \overline{(0,1)}$. Fix $a, b \in R$ with $b \neq 0$ such that $q = \overline{(a,b)}$. Since $\overline{(a,b)} \neq \overline{(0,1)}$, we know that $(a,b) \not\sim (0,1)$, so $a \cdot 1 \neq b \cdot 0$ which means that $a \neq 0$. Let $r = \overline{(b,a)}$ which makes sense because $a \neq 0$. We then have

$$\begin{aligned} q \cdot r &= \overline{(a,b)} \cdot \overline{(b,a)} \\ &= \overline{(ab, ba)} \\ &= \overline{(ab, ab)} \\ &= \overline{(1,1)} \end{aligned}$$

where the last line follows from the fact that $ab \cdot 1 = ab = ab \cdot 1$. Since we already proved commutativity of $+$, we conclude that $r \cdot q = \overline{(1,1)}$ also.

9. Distributivity: Let $q, r, s \in Frac(R)$. Fix $a, b, c, d, e, f \in R$ with $b, d, f \neq 0$ such that $q = \overline{(a,b)}$, $r = \overline{(c,d)}$, and $s = \overline{(e,f)}$. We then have

$$\begin{aligned} q \cdot (r+s) &= \overline{(a,b)} \cdot (\overline{(c,d)} + \overline{(e,f)}) \\ &= \overline{(a,b)} \cdot \overline{(cf+de, df)} \\ &= \overline{(a(cf+de), b(df))} \\ &= \overline{(acf+ade, bdf)} \\ &= \overline{(b(acf+ade), b(bdf))} \\ &= \overline{(abcf+abde, bbdf)} \\ &= \overline{((ac)(bf)+(bd)(ae), (bd)(bf))} \\ &= \overline{(ac, bd)} + \overline{(ae, bf)} \\ &= \overline{(a,b)} \cdot \overline{(c,d)} + \overline{(a,b)} \cdot \overline{(e,f)} \\ &= q \cdot r + q \cdot s \end{aligned}$$

□

Although R is certainly not a subring of $Frac(R)$ (it is not even a subset), our next proposition says that R can be embedded in $Frac(R)$.

Proposition 11.8.5. *Let R be an integral domain. Define $\varphi\colon R \to Frac(R)$ by $\theta(a) = \overline{(a,1)}$. We then have that θ is an injective ring homomorphism.*

Proof. Notice first that $\theta(1) = \overline{(1,1)}$, which is the multiplicative identity of $Frac(R)$. Now for any $a, b \in R$, we have

$$\begin{aligned} \theta(a+b) &= \overline{(a+b, 1)} \\ &= \overline{(a \cdot 1 + 1 \cdot b, 1 \cdot 1)} \\ &= \overline{(a,1)} + \overline{(b,1)} \\ &= \theta(a) + \theta(b) \end{aligned}$$

and

$$\begin{aligned} \theta(ab) &= \overline{(ab, 1)} \\ &= \overline{(a \cdot b, 1 \cdot 1)} \\ &= \overline{(a,1)} \cdot \overline{(b,1)} \\ &= \theta(a) \cdot \theta(b) \end{aligned}$$

Thus, $\theta\colon R \to Frac(R)$ is a homomorphism. Suppose now that $a, b \in R$ with $\theta(a) = \theta(b)$. We then have that $\overline{(a,1)} = \overline{(b,1)}$, so $(a,1) \sim (b,1)$. It follows that $a \cdot 1 = 1 \cdot b$, so $a = b$. Therefore, θ is injective. □

We have now completed our primary objective in showing that every integral domain can be embedded in a field. Our final result about $Frac(R)$ is that it is the "smallest" such field. We can't hope to prove that it is a subset of every field containing R because of course we can always rename elements. However, we can show that if R embeds in some field K, then you can also embed $Frac(R)$ in K. In fact, we show that there is a unique way to do it so that you "extend" the embedding of R.

Theorem 11.8.6. *Let R be an integral domain. Let $\theta\colon R \to Frac(R)$ be defined by $\theta(a) = \overline{(a,1)}$ as above. Suppose that K is a field and that $\psi\colon R \to K$ is an injective ring homomorphism. There exists a unique injective ring homomorphism $\varphi\colon Frac(R) \to K$ such that $\varphi \circ \theta = \psi$.*

Proof. First notice that if $b \in R$ with $b \neq 0$, then $\psi(b) \neq 0$ (because $\psi(0) = 0$ and ψ is assumed to be injective). Thus, if $b \in R$ with $b \neq 0$, then $\psi(b)$ has a multiplicative inverse in K. Define $\varphi\colon Frac(R) \to K$ by letting

$$\varphi(\overline{(a,b)}) = \psi(a) \cdot \psi(b)^{-1}$$

We check the following.

- φ is well-defined: Suppose that $\overline{(a,b)} = \overline{(c,d)}$. We then have $(a,b) \sim (c,d)$, so $ad = bc$. From this we conclude that $\psi(ad) = \psi(bc)$, and since ψ is a ring homomorphism it follows that $\psi(a)\cdot\psi(d) = \psi(c)\cdot\psi(b)$. We have $b, d \neq 0$, so $\psi(b) \neq 0$ and $\psi(d) \neq 0$ by our initial comment. Multiplying both sides by $\psi(b)^{-1} \cdot \psi(d)^{-1}$, we conclude that $\psi(a) \cdot \psi(b)^{-1} = \psi(a) \cdot \psi(d)^{-1}$. Therefore, $\varphi(\overline{(a,b)} = \varphi(\overline{(c,d)}$.

- $\varphi(1_{Frac(R)}) = 1_K$: We have

$$\varphi(\overline{(1,1)}) = \psi(1) \cdot \psi(1)^{-1} = 1 \cdot 1^{-1} = 1$$

- φ preserves addition: Let $a, b, c, d \in R$ with $b, d \neq 0$. We have

$$\begin{aligned}
\varphi(\overline{(a,b)} + \overline{(c,d)}) &= \varphi(\overline{(ad + bc, bd)}) \\
&= \psi(ad + bc) \cdot \psi(bd)^{-1} \\
&= (\psi(ad) + \psi(bc)) \cdot (\psi(b) \cdot \psi(d))^{-1} \\
&= (\psi(a) \cdot \psi(d) + \psi(b) \cdot \psi(c)) \cdot \psi(b)^{-1} \cdot \psi(d)^{-1} \\
&= \psi(a) \cdot \psi(b)^{-1} + \psi(c) \cdot \psi(d)^{-1} \\
&= \varphi(\overline{(a,b)}) + \varphi(\overline{(c,d)})
\end{aligned}$$

- φ preserves multiplication: Let $a, b, c, d \in R$ with $b, d \neq 0$. We have

$$\begin{aligned}
\varphi(\overline{(a,b)} \cdot \overline{(c,d)}) &= \varphi(\overline{(ac, bd)}) \\
&= \psi(ac) \cdot \psi(bd)^{-1} \\
&= \psi(a) \cdot \psi(c) \cdot (\psi(b) \cdot \psi(d))^{-1} \\
&= \psi(a) \cdot \psi(c) \cdot \psi(b)^{-1} \cdot \psi(d)^{-1} \\
&= \psi(a) \cdot \psi(b)^{-1} \cdot \psi(c) \cdot \psi(d)^{-1} \\
&= \varphi(\overline{(a,b)}) \cdot \varphi(\overline{(c,d)})
\end{aligned}$$

- φ is injective: Let $a, b, c, d \in R$ with $b, d \neq 0$ and suppose that $\varphi(\overline{(a,b)}) = \varphi(\overline{(c,d)})$. We then have that $\psi(a) \cdot \psi(b)^{-1} = \psi(c) \cdot \psi(d)^{-1}$. Multiplying both sides by $\psi(b) \cdot \psi(d)$, we conclude that $\psi(a) \cdot \psi(d) = \psi(b) \cdot \psi(c)$. Since ψ is a ring homomorphism, it follows that $\psi(ad) = \psi(bc)$. Now ψ is injective, so we conclude that $ad = bc$. Thus, we have $(a, b) \sim (c, d)$ and so $\overline{(a,b)} = \overline{(c,d)}$.
- $\varphi \circ \theta = \psi$: For any $a \in R$, we have

$$\begin{aligned}(\varphi \circ \theta)(a) &= \varphi(\theta(a)) \\ &= \varphi(\overline{(a,1)}) \\ &= \psi(a) \cdot \psi(1)^{-1} \\ &= \psi(a) \cdot 1^{-1} \\ &= \psi(a)\end{aligned}$$

  It follows that $(\varphi \circ \theta)(a) = \psi(a)$ for all $a \in R$.

We finally prove uniqueness. Suppose that $\phi\colon Frac(R) \to K$ is a ring homomorphism with $\phi \circ \theta = \psi$. For any $a \in R$, we have

$$\phi(\overline{(a,1)}) = \phi(\theta(a)) = \psi(a)$$

Now for any $b \in R$ with $b \neq 0$, we have

$$\begin{aligned}\psi(b) \cdot \phi(\overline{(1,b)}) &= \phi(\overline{(b,1)}) \cdot \phi(\overline{(1,b)}) \\ &= \phi(\overline{(b,1)} \cdot \overline{(1,b)}) \\ &= \phi(\overline{(b \cdot 1, 1 \cdot b)} \\ &= \phi(\overline{(b,b)}) \\ &= \phi(\overline{(1,1)}) \\ &= 1\end{aligned}$$

where we used the fact that $(b, b) \sim (1, 1)$ and that ϕ is a ring homomorphism so sends the multiplicative identity to $1 \in K$. Thus, for every $b \in R$ with $b \neq 0$, we have

$$\phi(\overline{(1,b)}) = \psi(b)^{-1}$$

Now for any $a, b \in R$ with $b \neq 0$, we have

$$\begin{aligned}\phi(\overline{(a,b)}) &= \phi(\overline{(a,1)} \cdot \overline{(1,b)}) \\ &= \phi(\overline{(a,1)}) \cdot \phi(\overline{(1,b)}) \\ &= \psi(a) \cdot \psi(b)^{-1} && \text{(from above)} \\ &= \varphi(\overline{(a,b)})\end{aligned}$$

Therefore, $\phi = \varphi$. □

Corollary 11.8.7. *Suppose that R is an integral domain which is a subring of a field K. If every element of K can be written as ab^{-1} for some $a, b \in R$ with $b \neq 0$, then $K \cong Frac(R)$.*

Proof. Let $\psi\colon R \to K$ be the trivial map $\psi(r) = r$ and notice that ψ is an injective ring homomorphism. By the proof of the previous result, the function $\varphi\colon Frac(R) \to K$ defined by

$$\varphi(\overline{(a,b)}) = \psi(a) \cdot \psi(b)^{-1} = ab^{-1}$$

is an injective ring homomorphism. By assumption, φ is surjective, so φ is an isomorphism. Therefore, $K\colon Frac(R)$. □