An Abstract Algebra Primer

Dennis Chen

February 11, 2025

Table of Contents

Chapter 1

Introduction

This is meant to be, as the title suggests, a primer into abstract algebra. It is not meant to be a traditional text, so I will expound a little on the goals of the text and the style it will adopt.

This primer is not particularly concerned with thoroughness. For instance, we do not mention that the simplicity of N and $\frac{G}{N}$ implies the simplicity of G, besides as a throwaway comment in module theory. Nor do we prove the Jordan-Holder theorem, though a thorough writeup is linked near the theorem. Furthermore, the proofs are not always completely rigorous. Especially when it comes to routine inductions, we will be a little lazy and neglect to perform the actual induction.

The primary goal is to **develop a general intuition** for your first genuine study of algebra. This is not a text meant to be studied at depth. It is meant to be skimmed for an overview of introductory algebra.

The style is **highly conversational**, in stark contrast with other texts (including, to some extent, my own). It is more akin to my blogging style. The aim is not for you to take your time to develop a deep understanding of algebra. Rather, it is so you can quickly take a first pass of algebra and develop a base of intuition.[1] A highly conversational style, I think, will work best here.

The structure of the content loosely models the Spring 2024 rendition of graduate algebra (21-610) taught by Professor James Cummings at Carnegie Mellon University. However, I have taken some liberties. I have added introductory content in groups and rings, and I make fewer assumptions on the properties a ring satisfies (at least initially). Furthermore the presentation of module theory is significantly different.

1.1 Prerequisites

I could claim that this text has no prerequisites besides a basic understanding of middle school math. This would technically be true, but there are some things it would be helpful to know to have a good intuition about algebra. So I will explicitly assume you know everything that I am about to list, although you do not necessarily have to understand everything to benefit.

First and foremost, you should roughly know what a proof looks like. This is not in spite of our relaxed standards for rigor, it is because of it. If you understand how to read and write a proof, you will have a good feel for how our intuitions are converted into proofs.

[1] Of course, "quickly" here is only relative to more extensive texts. Learning all of this content will likely require a few months of active study.

Conversely, if you do not, then you are just reading a list of facts with loosey-goosey justifications.

You need to have a good intuitive understanding of what a set is and how injective/surjective functions behave. You do not need to have any formal knowledge of set theory (no one is expecting you to recite the axioms of ZFC in your sleep), or know that a function happens to be a subset of the Cartesian product of the domain and codomain.[2] But working with sets should feel natural to you.

You should know some rudimentary number theory. Know what it means to take something "modulo p", and for groups, know Fermat's Little Theorem/Euler's Totient Function. (It is very helpful if you also know the proofs!) For rings, you will also want to understand the Euclidean Algorithm and know the Fundamental Theorem of Arithmetic (you do not need to know the proof).

You will also want a good understanding of combinatorial arguments. Even if we do not use some of them directly, if you do not know how many subsets $\{1, 2, ..., n\}$ has, it is unlikely that the many combinatorial arguments in this text will make sense to you.

Some background in proof-based linear algebra would be helpful. You do not need to know what an eigenvector is, but know all the background that comes before. Bases, linear independence, Rank-Nullity, etc. I will assume you are deeply familiar with all of this. Maybe also know what a matrix is, though this is not at all necessary.

1.2 Errata

A list of errata for past versions of the book lives at https://dennisc.net/writing/blog/algebra-primer.

I wrote this book over the course of a few months. As such, it is highly probable that significant typos or errors exist. If that is the case please let me know, even if they are trivial typos. Thank you.

[2] Only in the proof of Theorem 4.42 do I expect you to know some basic undergraduate set theory. Even then it is not very crucial.

Chapter 2

Groups

The first algebraic structure we study is the group. Historically, this is backwards: the seeds of group theory were developed by Galois in order to answer questions about fields. Groups have many connections and applications outside of Galois Theory, so this is fine. But keep this in the back of your head, especially when we explicitly study Galois Theory. Some of the seemingly random definitions we create now (e.g. solvability) are directly related to it.

2.1 The Basics

Take any positive n and consider the residues modulo n. We denote this set of residues as $\mathbb{Z}/n\mathbb{Z}$. Notice the following facts:

- there are n distinct residues,
- there is an element 0 such that $0 + g \equiv g \pmod{n}$ for any $g \in \mathbb{Z}/n\mathbb{Z}$,
- for every element g, there is an element $-g$ such that $(-g) + g \equiv 0 \pmod{n}$ for any g.

There are two additional things to note.

- First note that $\underbrace{g + \ldots + g}_{n \text{ copies}} \equiv 0 \pmod{n}$ for all $g \in \mathbb{Z}/n\mathbb{Z}$.
- Similarly note that for all $g \in \mathbb{Z}/n\mathbb{Z}$, n is divisible by the smallest positive k where $\underbrace{g + \ldots + g}_{k \text{ copies}} \equiv 0 \pmod{n}$. For example, if $n = 6$, the smallest k for $g = 2$ would be 3 as $2 + 2 + 2 \equiv 0 \pmod{6}$. And $3 \mid 6$.

Very explicitly, consider $(\mathbb{Z}/n\mathbb{Z}, +)$. And note that $\mathbb{Z}/n\mathbb{Z}$ is a set, whereas $+$ is a function of the form $+ : \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$. The pair $(\mathbb{Z}/n\mathbb{Z}, +)$ is an example of a **group**.

> **Definition 2.1 (Group)**: A **group** is a pair $(G, \cdot)$ where $\cdot$ is an **associative** binary function of the form $\cdot : G \times G \to G$ that satisfies the following properties:
>
> - There is an identity element id such that for all $g \in G$, $\mathrm{id} \cdot g = g$.
> - For every $g \in G$, there is some element $g^{-1} \in G$ such that $g^{-1} \cdot g = \mathrm{id}$.[1]

Some basic consequences:

[1] As our axioms do not specify uniqueness of id, strictly we should say that $(g^{-1} \cdot g) \cdot h = h$ for all $h \in G$, i.e. $g^{-1} \cdot g$ "behaves like an identity element". But we choose to be a little imprecise for purposes of clarity here, especially because the identity element does turn out to be unique.

- The inverse is two-sided: if $g^{-1} \cdot g = \text{id}$ then $g \cdot g^{-1} = \text{id}$ too. Put another way, every element commutes with its inverse to give identity.
- The identity is two-sided: $g \cdot \text{id} = g$ too. Put another way, multiplying by the identity element does nothing, no matter where you put it.
- The inverse is unique.
- $\text{id}^{-1} = \text{id}$.

The proofs are fairly straightforward.

Definition 2.2 (Auxillary Notation for Groups):

- We will often write gh instead of $g \cdot h$. Also, when $\cdot$ is clear or does not need to be explicitly written, we will refer to the group $(G, \cdot)$ simply as G.
- We can repeatedly apply the group operation with many copies of an element: we denote $\underbrace{g \cdot \ldots \cdot g}_{k \text{ copies}}$ as g^k.
- We write $|G|$ to denote the number of elements inside the group G and refer to it as the **order** of G. For example, $|\mathbb{Z}/n\mathbb{Z}| = n$ since $\mathbb{Z}/n\mathbb{Z}$ has n elements.

  When the order is finite we call G a "finite group" and write $|G| < \infty$. And when it is infinite we call G an "infinite group" and write $|G| = \infty$.
- Similarly we write $|g|$ to denote the smallest positive integer such that $g^{|g|} = \text{id}$. We call this the **order** of g. We may classify the order of an element as finite or infinite, just as with a group.

Implicitly, groups are "closed" under their binary operation. This is by definition, but it is important to explicitly keep this in mind so you understand $\{-1, 0, 1\} \subset \mathbb{Z}$ is not a group under addition.

Here are some examples of groups:

- $(\mathbb{Z}/n\mathbb{Z}, +)$, of course.
- $(\mathbb{Z}, +)$
- $(\mathbb{Q}, +)$
- $(\mathbb{R}, +)$
- $(\mathbb{C}, +)$
- $G \times H$, i.e. the Cartesian product of groups G and H where the group operation is defined componentwise. (So for example, $\mathbb{Z} \times \mathbb{Q}$ is a group.)

All the groups we have listed so far are abelian (i.e. every pair of elements commute). By the way, it is a fact (not one we will explore in any depth) that every finite abelian group is equivalent[2] to the Cartesian product of groups of the form $\mathbb{Z}/p^k\mathbb{Z}$ (where p is prime). For example, $\mathbb{Z}/12\mathbb{Z} \cong \mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$.

[2] By "equivalent" we mean isomorphic. We will explore isomorphisms shortly.

2.1.1 Non-commutative Groups

There is one big difference between a general group and $\mathbb{Z}/n\mathbb{Z}$: the group $\mathbb{Z}/n\mathbb{Z}$ is commutative whereas a general group is not. There are lots of non-commutative groups, here are two examples:

- The group of $n \times n$ matrices with non-zero determinant, which we refer to as GL_n.
- The dihedral group D_{2n} of an n-gon, i.e. the group formed by considering the rotations the interchange vertices and the reflections about the $2n$ lines of symmetry. The dihedral group is a good source of intuition and I recommend you read about it, e.g. on https://en.wikipedia.org/wiki/Dihedral_group
- **Most importantly**, the group of bijective functions $f : X \to X$, where the group operation is function composition. We call bijections from a function to itself **permutations**[3] and it is a fact that every group is identical to some permutation group. (We will soon make this idea much more precise.)

  We denote this permutation group as S_X.

2.1.2 Subgroups and Cosets

Suppose $|G| = n$. Analogously to $\mathbb{Z}/n\mathbb{Z}$, it is a fact that for general groups G, $g^n = \text{id}$ and n is divisible by the smallest positive k such that $g^k = \text{id}$. We develop the notion of a subgroup and use it to prove these facts, but the theory of subgroups is much richer than these basic facts.

> **Definition 2.3 (Subgroup)**: A **subgroup** H of G is any subset of G closed under the group operation of G. We write $H \leq G$.
>
> The subgroup H is **proper** if it is not equal to G. In this case we write $H < G$.

As the name suggests, H (with the group operation of G) is a group in its own right. (This implies $\text{id} \in H$.) For example, a subgroup of $\mathbb{Z}/6\mathbb{Z}$ is the set of even residues.

Exercise 2.4: If $H, K \leq G$, show that $H \cap K \leq G$.

Exercise 2.5 (The Center of a Group): The **center** of G, which we denote as $Z(G)$, is the set of all g that commute with all $h \in G$. Formally, $Z(G) = \{g : g \cdot h = h \cdot g \text{ for all } h \in H\}$. It turns out $Z(G) \leq G$. Verify this.

Exercise 2.6: Show that for any proper subgroup H of $\mathbb{Q}$, there exists some $H < K < \mathbb{Q}$. (This shows that $\mathbb{Q}$ has no **maximal subgroups**.)

[3] Here is how this term squares with your prior intuition of a permutation of a list like $(1, 2, 3)$. Instead of considering $(2, 3, 1)$ as "a different ordering", consider it as "the function π where $\pi(1) = 2$, $\pi(2) = 3$, and $\pi(3) = 1$".

Definition 2.7 (Coset): Given $H \leq G$, we say $gH = \{gh : h \in H\}$ is a **left coset** of H. (Usually we will just refer to it as a coset.)

Here is an important theme that will persist for a while: we may consider $g \in G$ as an element, but we may also associate it with a unique permutation π_g that maps $\pi_g : h \mapsto gh$.[4]

Furthermore, we may consider gH as the image of π_g applied to H. And because we just established π_g is a permutation (i.e. a bijection), **the size of every H-coset is the same**. After noting that any two distinct H-cosets are disjoint, one may easily derive the following.

Theorem 2.8 (Lagrange's Theorem): If $|G| < \infty$ and $H \leq G$, then $|H|[G : H] = |G|$, where $[G : H]$ denotes the number of H-cosets in G.

This is because we may partition G into $[G : H]$ equally-sized H-cosets. An important consequence is that $|H|$ divides $|G|$.

Exercise 2.9: Show that distinct H-cosets are indeed disjoint. (Hint: Suppose two H-cosets are not disjoint, that is, there is some $g_1 \in g_2 H$. This means $g_1 = g_2 h$ for some $h \in H$, now show every $g \in g_1 H$ is in $g_2 H$ to show that $g_1 H \leq g_2 H$.)

By the way, distinct cosets being disjoint means

$$g_1 = g_2 h \text{ for some } h \in H \iff g_1 H = g_2 H.$$

Now here's the kicker. Given any $g \in G$, the set of elements of the form g^a forms a subgroup of G. The number of elements in this subgroup is precisely the smallest k such that $g^k = \mathrm{id}$. So this k better divide n, and as an easy consequence, $g^n = \mathrm{id}$.

Exercise 2.10: Connect Lagrange's Theorem with our favorite example $\mathbb{Z}/n\mathbb{Z}$.

Exercise 2.11: For any $g_1, g_2 \in G$, show that

$$g_1 H = g_2 H \iff g_1^{-1} g_2 \in H.$$

2.1.3 Indices on Groups of Infinite Order

We denote the number of distinct H-cosets as $[G : H]$. This is called the **index** of H in G.

When defining the index, there is no need for G or H to be finite. Obviously when G and H are finite, $[G : H] = \frac{|G|}{|H|}$. But it is possible for $|G|$ and $|H|$ to both be infinite and for there to be a finite number of H-cosets.

[4] This notation is convenient shorthand to say that π_g is the function where $\pi_g(h) = gh$.

Example 2.12: Consider $G = \mathbb{Z}$ and $H = n\mathbb{Z}$. Both G and H are infinite, and the expression $\frac{|G|}{|H|}$ is not well-defined, but $[G : H]$ is well-defined to be n. **Infinite subgroups of infinite groups can have finite index.**

Example 2.13: Infinite subgroups of infinite groups can have infinite index too. Consider $\mathbb{Q} \leq \mathbb{R}$. If there were a finite number of $\mathbb{Q}$-cosets then we could enumerate $\mathbb{R}$, and since we cannot enumerate $\mathbb{R}$, we know there are an infinite number of cosets. (Uncountably infinite, in fact.)

What's the takeaway? An index is just the cardinality of a set (the set of H-cosets when $H \leq G$). And any cardinality can be achieved. So do not make any hasty assumptions that G, H, or $[G : H]$ are finite. We often will not need them.

2.2 Homomorphisms and Isomorphisms

Very loosely, a homomorphism φ is a map between groups $\varphi : G \to H$ such that the group structures of G and H are respected by φ.

> **Definition 2.14 (Homomorphism)**: A **homomorphism** $\varphi : G \to H$ (where G and H are groups) is a function where
>
> $$\varphi(g_1 g_2) = \varphi(g_1)\varphi(g_2)$$
>
> for all $g_1, g_2 \in G$.
>
> We say a homomorphism is an **isomorphism** if it is a bijection. We say an isomorphism is an **automorphism** if $\varphi : G \to G$.

The motto is as follows:

> In a homomorphism, it doesn't matter if you apply the group operation or φ first.

Note that $g_1 g_2$ is a multiplication between two elements in G whereas $\varphi(g_1)\varphi(g_2)$ is a multiplication between two elements in H.

Exercise 2.15: Show that $\varphi(g^k) = \varphi(g)^k$ for all integers k.

Exercise 2.16: We say that G and H are **isomorphic** if there exists some isomorphism between G and H, and we write it as $G \cong H$.

Show that $\cong$ is an equivalence relation, that is, $\cong$ is

- reflexive
- symmetric
- transitive.

Definition 2.17 (Kernel and Image): Given a homomorphism $\varphi : G \to H$, $\ker \varphi = \{g \in G : \varphi(g) = \mathrm{id}\}$ and $\operatorname{im} \varphi = \{\varphi(g) : g \in G\}$.

Note $\ker \varphi \leq G$ and $\operatorname{im} \varphi \leq H$.[5]

Exercise 2.18: Consider the homomorphism $\varphi : \mathbb{Z} \to \mathbb{Z}/8\mathbb{Z}$ defined as follows: $\varphi : x \mapsto 2x$. What are $\ker \varphi$ and $\operatorname{im} \varphi$? Keep this exercise in mind: we will come back to this homomorphism.

Here is the setting for the rest of this section. We will consider a homomorphism $\varphi : G \to H$ and study $\ker \varphi$ and its cosets.

Theorem 2.19: For all $g_1, g_2 \in G$,

$$\varphi(g_1) = \varphi(g_2) \Longleftrightarrow g_1 \ker \varphi = g_2 \ker \varphi.$$

Proof of Theorem 2.19: Note that

$$\begin{aligned} \varphi(g_1) = \varphi(g_2) &\Longleftrightarrow \varphi(g_1 g_2^{-1}) = \mathrm{id} \\ &\Longleftrightarrow g_1 g_2^{-1} \in \ker \varphi \\ &\Longleftrightarrow g_1 \ker \varphi = g_2 \ker \varphi. \end{aligned}$$

□

This implies the $\ker \varphi$-cosets form a group with the following operation:

$$g_1 \ker \varphi \cdot g_2 \ker \varphi = (g_1 g_2) \ker \varphi.$$

We will call this group a **quotient group**. When we try to form a quotient group with some arbitrary $N \leq G$, there's one main sticking point: we need **representation invariance**, i.e. we need

$$g_1 N = g_3 N \text{ and } g_2 N = g_4 N \Longrightarrow (g_1 g_2) N = (g_3 g_4) N$$

for all $g_1, g_2, g_3, g_4 \in G$. This is true precisely when $N = \ker \varphi$ for some homomorphism φ (the algebraic manipulations to show this are nearly identical to the ones that show Theorem 2.19).

Call this group $\frac{G}{\ker \varphi}$ (and for any $N \leq G$ where we can perform a similar construction, define $\frac{G}{N}$ the same way). We may define a homomorphism

$$\varphi^* : \frac{G}{\ker \varphi} \to \operatorname{im} \varphi$$

where $\varphi^*(g \ker \varphi) = \varphi(g)$.

[5] We use $\leq$ specifically to denote subgroups!

Exercise 2.20: Check that φ^* satisfies **representation invariance**. (Part of the exercise is figuring out what precisely representation invariance means here.)

It turns out that φ^* is an isomorphism. It is injective as

$$\varphi^*(g_1 \ker \varphi) = \varphi^*(g_2 \ker \varphi) \iff \varphi(g_1) = \varphi(g_2) \iff g_1 \ker \varphi = g_2 \ker \varphi$$

and it is surjective by definition (that is why we chose the codomain to be $\operatorname{im} \varphi$). So $\frac{G}{\ker \varphi} \cong \operatorname{im} \varphi$. By the way, this is the **First Isomorphism Theorem**.

Exercise 2.21: Explicitly verify the First Isomorphism Theorem holds for Exercise 2.18.

We're not done though: we still want to find which $N \leq G$ admit a quotient group. It turns out only the subgroups that are the kernel of some homomorphism do, because **representation invariance is equivalent to being a kernel**.

We've already shown that kernels are representation invariant. And if N is representation invariant, then $\varphi : G \to \frac{G}{N}$ where $\varphi : g \mapsto gN$ is a homomorphism with kernel N.[6] So indeed these two conditions are equivalent.

There is another equivalent condition. It turns out that representation invariance holds precisely when $gNg^{-1} \leq N$ for all $g \in G$. Put another way, for all $g \in G$, $n \in N$, we must have $gng^{-1} \in N$:

- Representation invariance implies $N = \ker \varphi$ for some homomorphism φ, which means $\varphi(gng^{-1}) = \varphi(g)\varphi(n)\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = \text{id}$, ie. $gng^{-1} \in \ker \varphi = N$.
- Suppose $gNg^{-1} \leq N$ and $g_1N = g_3N$, $g_2N = g_4N$. Then we'd like to show $g_1g_2N = g_3g_4N$. Note

$$\begin{aligned} g_1g_2(g_3g_4)^{-1} &= g_1g_2g_4^{-1}g_3^{-1} \\ &= (g_1g_3^{-1})(g_3g_2g_4^{-1}g_3^{-1}) \end{aligned}$$

  By Exercise 2.11, $g_1g_3^{-1} \in N$, and as $gNg^{-1} \leq N$, $(g_2g_4^{-1})^{g_3} \in N$ as well. So the whole expression is in N, implying $g_1g_2N = g_3g_4N$ as desired.

Exercise 2.22: Verify that $gNg^{-1} = \{gng : n \in N\}$ is a subgroup of G for all $g \in G$.

For pedagogical purposes we will write $h^g = ghg^{-1}$. (This will become especially important when we consider **group actions** later.) We are going to treat $-^g$ as the map sending $h \mapsto ghg^{-1}$. (The $-$ indicates that the term we are "varying" or "feeding into" the map is positioned at the dash.[7]) We refer to this map as "conjugation by g".

Note: for each g in G we may define a map $-^g$. Put another way, there is a conjugation map associated with every element of G.

Exercise 2.23: Show that $-^g : G \to G$ is an automorphism for every $g \in G$.

Armed with all this knowledge, we may finally define and characterize **normal subgroups**.

[6] Note the identity of $\frac{G}{N}$ is N.

[7] This notation is more heavily used in Category Theory.

Definition 2.24 (Normal Subgroups): Given a subgroup N of G, we say N is **normal** if any of the following equivalent properties are satisfied:

1. $N = \ker \varphi$ for some homomorphism $\varphi : G \to H$. (We don't really care what H is.)
2. N is representation invariant.
3. $gNg^{-1} \leq N$ for all $g \in G$.
4. $gNg^{-1} = N$ for all $g \in G$.

We write $N \trianglelefteq G$ if N is normal.

We've already established that (2) $\Leftrightarrow$ (3) and that (1) $\iff$ (2). So all that is left is to make an argument that (3) $\iff$ (4).[8]

Suppose $gNg^{-1} \leq N$ for all $g \in G$. We'd like to show that for all $n \in N$, $n \in gNg^{-1}$ as well.[9] Note $g^{-1}ng^1 \in N$ so $g(g^{-1}ng)g^{-1} = n \in gNg^{-1}$.

An important note: just because "$gNg^{-1} \leq N$ for **all** $g \in G$" is equivalent to "$gNg^{-1} = N$ for **all** $g \in G$" does not mean that $gNg^{-1} \leq N$ for some **particular** $g \in G$ implies $gNg^{-1} = N$. Here is a counterexample from https://math.stackexchange.com/a/107874.

Example 2.25: Consider $S_{\mathbb{Z}}$. Define $\sigma \in S_{\mathbb{Z}}$ as $\sigma : x \mapsto x + 1$ and $H \leq S_{\mathbb{Z}}$ as $H = \{f \in H : f(x) = x \text{ for all } x \leq 0\}$.

Evidently $\sigma H \sigma^{-1} \leq H$ but every permutation in $\sigma H \sigma^{-1}$ fixes 1, meaning $\sigma H \sigma^{-1}$ does not contain H.

We've spent all this time establishing some characterizations of **normal subgroups**. Let's end it off by finding a subgroup that is not normal. We cannot look in **abelian groups** because we quickly see that every subgroup in an abelian group is normal. (Check it yourself!) So look at the non-abelian S_3, the group of permutations with 3 elements.

Exercise 2.26: There is a permutation $\sigma \in S_3$ that interchanges the first and second element. Note $|\sigma| = 2$, so $\{1, \sigma\}$ is a subgroup. Find some $\pi \in S_3$ such that $\pi \sigma \pi^{-1}$ explicitly disproves normality.

One final note: normality is a fairly non-inheritable condition. Suppose $N \trianglelefteq G$. If $H \leq N$, H is not necessarily normal in N or G. And if $H \trianglelefteq N$, H still isn't necessarily normal in G.

Example 2.27: As a stupid counterexample, $1 \trianglelefteq G$ and $G \trianglelefteq G$, but obviously we cannot conclude that any intermediate subgroup H between 1 and G (which includes pretty much every subgroup) satisfies $H \trianglelefteq G$.

Exercise 2.28: Prove that $N \leq H \leq G$ and $N \trianglelefteq G$ implies $N \trianglelefteq H$. (Hint: Use the conjugation characterization of normality.)

[8] You should convince yourself that we need $-^g$ to be an automorphism if we want (3) $\iff$ (4), particularly when G is finite.

[9] This will give us $gNg^{-1} \geq N$.

Exercise 2.29: Suppose $N \trianglelefteq G$ and $H \leq G$. Show that $N \cap H \trianglelefteq H$.

Why is it important to have a good intuition for when normality is and is not preserved? Because this allows us to discern **when we can take quotient groups**. (Remember, representation invariance is one of the characterizations of normality, and it determines when we can take quotients.)

2.3 The Ancillary Isomorphism Theorems

The First Isomorphism Theorem has some very enlightening applications that are useful in their own right. They are the Second, Third, and Fourth Isomorphism Theorems.

First we must develop some facts about subsets of the form

$$HK = \{hk : h \in H, k \in K\}.$$

(As you would expect, here we are defining $H, K \leq G$.)

> **Definition 2.30 (Normalizer)**: The **normalizer** of $H \leq G$ is the subgroup $N_G(H)$ that satisfies any of three equivalent characterizations:
>
> - $N_G(H) = \{g : gHg^{-1} = H\}$.
> - $N_G(H)$ is the largest subgroup satisfying $H \trianglelefteq N_G(H) \leq G$.
> - $N_G(H)$ is the set of all elements commuting with H.

We will not prove any of these characterizations are equivalent. The one we will use most often is the first.

If $K \leq N_G(H)$ then $HK \leq G$ (here the sticking point is closure of HK) and $H \trianglelefteq HK$. Here is a sketch of the proof:

- Show that $HK = KH \iff HK \leq G$.
- Show that $K \trianglelefteq G \implies HK = KH$.
- Because $K \leq N_G(H)$ and obviously $H \leq N_G(H)$, we may conclude that $KH \leq N_G(H)$ and deduce that $H \trianglelefteq HK$.

The flavor of algebraic manipulations required is once again similar to those in Theorem 2.19.

> **Theorem 2.31 (Second Isomorphism Theorem)**: Given $H \leq G$ and $N \leq N_G(H)$, $\frac{NH}{H} \cong \frac{N}{N \cap H}$.

The Second Isomorphism Theorem is also known as the **Parallelogram Law**. The diagram below is a handy tip for remembering it.

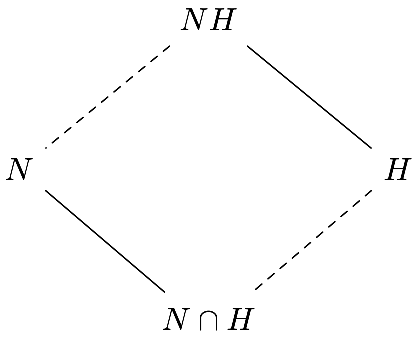


Figure 1: Solid lines represent normal subgroups.

Proof of Theorem 2.31: Define $\varphi : N \to \frac{NH}{H}$ as $\varphi : n \mapsto nH$. Note $\ker \varphi = N \cap H$ and φ is surjective. Applying the First Isomorphism Theorem on φ finishes. □

Note $N \trianglelefteq G$ is enough to guarantee $N \leq N_G(H)$. Most of the time, we will be using the Second Isomorphism Theorem when one of the subgroups is normal.

Theorem 2.32 (Third Isomorphism Theorem): Suppose $K \trianglelefteq H \trianglelefteq G$. Then

$$\frac{\frac{G}{K}}{\frac{H}{K}} \cong \frac{G}{H}.$$

Here is the analogy. If $a, b, c \in \mathbb{Q}^*$, then $\frac{a}{b} = \frac{\frac{a}{c}}{\frac{b}{c}}$. At least symbolically, groups behave the exact same way.

Proof of Theorem 2.32: Note $\frac{H}{K} \trianglelefteq \frac{G}{K}$, so the map

$$\varphi : \frac{G}{K} \to \frac{G}{H}$$
$$\varphi : gK \mapsto gH$$

is well-defined.

Obviously φ is surjective (i.e. $\operatorname{im} \varphi = \frac{G}{H}$), and $\ker \varphi = \frac{H}{K}$ as the only K-cosets mapped to the H-coset $\operatorname{id} H$ are those contained within H. Then First Isomorphism gives us exactly what we want. □

Now this last isomorphism theorem is the most important. Given $N \trianglelefteq G$, it helps us to understand the subgroups H where $N \leq H \leq G$. In fact, it tells us they are essentially the subgroups of $\frac{G}{N}$.

> **Theorem 2.33 (Fourth Isomorphism Theorem)**: Take $N \trianglelefteq G$. There is an obvious inclusion-preserving bijection between the H where $N \leq H \leq G$ and the subgroups of $\frac{G}{N}$. It is
>
> $$H \mapsto \frac{H}{N}.$$
>
> This bijection also preserves normality, as
>
> $$H \trianglelefteq G \Longleftrightarrow \frac{H}{N} \trianglelefteq \frac{G}{N}.$$

The proof of Theorem 2.33 is not hard; it is a direct application of the ideas we have developed throughout this section.

Why does this matter? Because we can now "factor" groups into a series of subgroups. If I have $N \trianglelefteq G$, then it is often enough to understand the structure of N and the structure of $\frac{G}{N}$ (i.e. their subgroups). Yes, considering only these subgroups loses us every subgroup H that isn't either contained in N or containing N, so we are not really considering the full picture. But selecting an appropriate N (or even an appropriate series of normal subgroups to successively quotient G by) can give us a lot of information. We will see this when we later study solvability, nilpotency, central series, and composition series.

2.4 Group Actions

Remember our theme about associating π_g with each $g \in G$? Here we take it to the extreme. The motto is as follows:

> Groups are collections of **permutations** on some set S closed under composition and inverse. The group operation is merely the composition of two permutations.

First to build some intuition, we will prove that every group is isomorphic to a group of permutations. Here is how: consider a group G. As stated, each $g \in G$ can be associated with the permutation $\pi_g : h \mapsto gh$ of G. There are two important observations about the mapping $g \mapsto \pi_g$:

- this mapping is injective: if $\pi_g = \pi_h$ then $g = h$,
- it is a homomorphism: note $\pi_{gh}(a) = gha$ and $(\pi_g \circ \pi_h)(a) = (\pi_g)(ha) = gha$ for all $a \in G$.

So we have an injective homomorphism $G \to S_G$.[10] Any function can be made into a surjection (just take the codomain to be the image), so voila: we have an isomorphism from G to a subgroup of S_G.

[10] Recall S_G is the group of permutations of the **set** G.

What shift in perspective have we made? Instead of performing a binary operation on G, we have G, whose members are thought of as **permutations on the set** G, acting on the set G, **whose members are just plain old elements**.

By the way, this is called **Cayley's Theorem**: "any group G is isomorphic to a subgroup of S_G."

The official definition of a group action is as follows.

> **Definition 2.34 (Group Action)**: A **group action** is defined with a group G and a set X, along with a binary operation $\cdot : G \times X \to X$ such that
>
> - $\mathrm{id} \cdot x = x$,
> - $(g_1 g_2) \cdot x = g_1 \cdot (g_2 \cdot x)$.

It turns out that each $g \in G$ can be associated with a permutation in S_X. To be explicit, for any particular g, the function $x \mapsto g \cdot x$ is a permutation since its two-sided inverse is $x \mapsto g^{-1} \cdot x$.

Be careful: unlike Cayley's Theorem, this association is **not injective**. As a counterexample, consider the very boring group action where $g \cdot x = x$ for all $g \in G$. Just because g is injective does not mean

$$g \mapsto (x \mapsto g \cdot x)$$

is injective. It usually is not.

So we can actually think of a group action another way. We can think of G as a subset of S_X that acts on X. Actually, G is a **subgroup** of S_X, not just a subset. (Why?) So a group action can also be defined just with a subset of S_X. This should feel familiar coming from Cayley's Theorem.

If you are familiar with the computer science term "currying", this is a great example of it. Because functions of type $G \times X \to X$ biject with functions of type $G \to (X \to X)$, we can think of $\cdot$ as a function of type

$$\cdot : G \to (X \to X).$$

2.4.1 Orbits and Stabilizers

Suppose we start with some $x \in X$. Which values in X can we get to by applying the group action on x?

> **Definition 2.35 (Orbit)**: The **orbit** of $x \in X$ is the set $\{g \cdot x : g \in G\}$, which we will denote as $\mathrm{Orb}(x)$ from now on.

Importantly, the orbits of X partition X. Here's why: suppose we define a relation $\sim$ on X, where for $x, y \in X$,

$$x \sim y \iff \text{there exists some } g \in G \text{ where } x \cdot g = y.$$

Exercise 2.36: Check that $\sim$ is an **equivalence relation** on X. Further note that

$$\operatorname{Orb}(x) = \operatorname{Orb}(y) \iff x \sim y.$$

Conclude that the orbits of X indeed partition X.

Furthermore, which g fix x under the group action?

> **Definition 2.37 (Stabilizer)**: The **stabilizer** of $x \in X$ is the set $\{g : g \cdot x = x\}$, which we will denote as $\operatorname{Stab}(x)$ from now on.

Exercise 2.38: Show that $\operatorname{Stab}(x) \leq G$.[11]

For typechecking purposes, remind yourself that $\operatorname{Orb}(x) \subseteq X$ and $\operatorname{Stab}(x) \leq G$.[12]

There is an important relation between the sizes of the orbit and the stabilizer that comes from bijecting the $\operatorname{Stab}(x)$-cosets of G with the elements in $\operatorname{Orb}(x)$.

> **Theorem 2.39 (Orbit-Stabilizer)**: Suppose G acts on X. For every $x \in X$,
>
> $$|\operatorname{Orb}(x)| = [G : \operatorname{Stab}(x)].$$

Proof of Theorem 2.39: Because $\operatorname{Stab}(x) \leq G$, we may look at the $\operatorname{Stab}(x)$-cosets of G.

Define a map $\varphi : \frac{G}{\operatorname{Stab}(x)} \to \operatorname{Orb}(x)$ where $\varphi : g \operatorname{Stab}(x) \mapsto g \cdot x$. This is obviously a surjection, and this is an injection as

$$\begin{aligned} g_1 \cdot x = g_2 \cdot x &\implies \left(g_1^{-1} g_2\right) \cdot x = x \\ &\implies g_1^{-1} g_2 \in \operatorname{Stab}(x) \\ &\implies g_1 \operatorname{Stab}(x) = g_2 \operatorname{Stab}(x). \end{aligned}$$

□

In particular, what happens when we consider G acting on itself with the action of conjugation? More precisely, we define the action $g \cdot x = gxg^{-1}$ where $g, x \in G$.[13] We call the orbits the **conjugacy classes**. In other words, a conjugacy class of an element $x \in G$ is "the stuff in G we can get by conjugating x".

[11] It is also true that any subgroup of G is the stabilizer of some action of G on some set X.

[12] It is true by definition that $\operatorname{Stab}(x) \subseteq G$. It is also not too hard to show that $\operatorname{Stab}(x)$ is actually a subgroup of G.

[13] To be even more explicit, the group and the set considered in the group action are both G.

Exercise 2.40: Take a group G and a subset S[14] of G. We denote the set of all elements in G that commute with every element in S as $C_G(S)$. Use Theorem 2.39 to show that for all $g \in G$, $[G : C_G(g)]$ is equal to the size of the conjugacy class of g.

What happens when we study the sizes of the conjugacy classes of G? This yields the **Class Equation**.

Theorem 2.41 (Class Equation): Consider a finite group G with center $Z(G)$ (see Exercise 2.5) and conjugacy classes $c_1, ..., c_n$ comprising all the elements not in $Z(G)$. Let g_i be an element in each c_i (it does not matter which). Furthemore let $C_G(g_i)$ denote the set of all elements that commute with g_i. Then

$$|G| = |Z(G)| + \sum_{i=1}^{n} |G : C_G(g_i)|.$$

Proof of Theorem 2.41: Note $C_G(g_i)$ is the stabilizer of g_i under the action G, so $|G : C_G(g_i)| = |\mathrm{Orb}(g_i)|$. Obviously the orbit of each center element is just itself (it commutes with everything after all). So

$$|Z(G)| + \sum_{i=1}^{n} |G : C_G(g_i)| = \text{the sum of the sizes of the orbits of } G,$$

which is obviously equal to $|G|$. □

Similarly, it is generally true that for some G acting on some finite X, where

- $|\mathrm{Fix}(X)|$ is the set of all $x \in X$ fixed under all $g \in G$,
- and the orbits $\mathrm{Orb}(x_1), ..., \mathrm{Orb}(x_n)$ partition the rest of X not in $\mathrm{Fix}(X)$,

$$|X| = |\mathrm{Fix}(X)| + \sum_{i=1}^{n} [G : \mathrm{Stab}(x_i)].$$

The proof is identical to that of Theorem 2.41.

Exercise 2.42 (Burnside's Lemma): Suppose finite group G acts on set X. Show that the number of orbits of X under G is equal to

$$\frac{1}{|G|} \sum_{g \in G} |\mathrm{Fix}(g)|.$$

Hint: Show that

$$\sum_{\{g \in G\}} |\mathrm{Fix}(g)| = \sum_{\{x \in X\}} |\mathrm{Stab}(x)|.$$

Then finish with Theorem 2.39.

[14] Said subset is not necessarily a subgroup!

Now we will explore two applications of the class equation that are important in their own right. (Meaning that you will be expected to remember these facts.)

Example 2.43: The center of a group with order p^k is non-trivial.

Solution to Example 2.43: Note $|G : C_G(g_i)|$ is divisible by p as $C_G(g_i) \neq G$. Since $|G|$ is also divisible by p, the class equation

$$|G| = |Z(G)| + \sum_{i=1}^{n} |G : C_G(g_i)|$$

implies $|Z(G)|$ must also be divisible by p. So $|Z(G)| \neq 1$. □

Example 2.44 (Cauchy's Theorem): If G is a finite group whose order is divisible by k, then it has a subgroup of order p.

Actually, it is enough to find an element $g \in G$ with order p. For then the cyclic subgroup $\{\text{id}, g, g^2, ...g^{p-1}\}$ has order p.

Solution to Example 2.44: Define the set

$$X = \{(g_1, ..., g_p) : g_i \in G, g_1 \cdot ... \cdot g_n = \text{id}\},$$

i.e. the set of p-tuples in G where the product of the entries is id, and define a group action that cycles the entries of S by 1 slot.[15] Note $|X| = g^{p-1}$: the first $p - 1$ entries of an element in S fix the last element (because it must be the inverse of the product of the first $p - 1$ elements). Since p divides g, p also divides $|X|$.

Note the fixed points under this group action are the elements of the form $(g, ..., g)$ and every other orbit has size p. Since p divides X and

$$|X| = |\text{Fix}(X)| + \sum_{i=1}^{n} [G : \text{Stab}(x_i)],$$

we must have that p divides $|\text{Fix}(X)|$. Since $(\text{id}, ..., \text{id}) \in \text{Fix}(X)$, we must have some other element $(g, ..., g) \in \text{Fix}(X)$. This corresponds to some $g \neq \text{id}$ such that $g^p = \text{id}$, as desired. □

2.5 Sylow’s Theorems

Here we extend Cauchy’s Theorem. If G is a finite group whose order is divisible by p^k and not p^{k+1}, we want to study the subgroups of G with order p^k. Our study of these subgroups will yield facts known as **Sylow’s Theorems**.

First we would like to establish that these subgroups of order p^k exist. We will call the set of all such subgroups the **Sylow p-subgroups of G** and denote it as $\text{Syl}_p(G)$. Here is how we prove there is a Sylow p-subgroup.

- We induct on $|G|$ to show the existence of a Sylow p-subgroup.

[15] You can make the underlying group have two elements, the details of the group do not matter.

- If $p \mid |Z(G)|$ then by Cauchy's, there is some $N \leq Z(G)$ with order p. Then there is some $\frac{P}{N} \leq \frac{G}{N}$ with order p^{k-1} by the inductive hypothesis. Thus $|P| = p^k$.
- Otherwise use the class equation

  $$|G| = |Z(G)| + \sum_{i=1}^{n} |G : C_G(g_i)|$$

  to conclude that there exists some $C_G(g_i)$ such that $|G : C_G(g_i)|$ is not divisible by p. (This is because $|Z(G)|$ is not divisible by p where $|G|$ is.)

  Then $|C_G(g_i)|$ is divisible p^k and we are done by the inductive hypothesis.

Now we have a Sylow p-subgroup P. What happens when we conjugate P? Suppose all the conjugates are $\{P_1, ..., P_n\}$. Now take any subgroup $H \leq G$. We may define a group action with H and $\{P_1, ..., P_n\}$. The elements of H act on $\{P_1, ..., P_n\}$ via conjugation, and this actions forms orbits $\mathcal{O}_1 \cup ... \cup \mathcal{O}_a$.

Without loss of generality suppose $P_i \in \mathcal{O}_i$ for all $1 \leq i \leq a$, meaning P_i represents the orbit $\mathcal{O}_i$. By Theorem 2.41,

$$n = |\mathcal{O}_1| + ... + |\mathcal{O}_a|,$$

and by Theorem 2.39 $|\mathcal{O}_i| = |H : N_H(P_i)|$. By definition $N_H(P_i) = N_G(P_i) \cap H$.

Now we take for granted the fact that for any $P \in \mathrm{Syl}_p(G)$ and $H \leq G$ where $|H| = p^a$, $H \cap N_G(P) = H \cap P$.[16] This yields $|\mathcal{O}_i| = [H : P_i \cap H]$, and so the class equation becomes

$$n = [H : P_1 \cap H] + ... + [H : P_a \cap H].$$

Now we will take full advantage of the fact that H is arbitrary and take some H that reveal interesting facts about the Sylow p-subgroups. Because H is arbitrary taking $H = P_1$ implies

$$n = 1 + [P_1 : P_2 \cap P_1] + ... + [P_1 : P_a \cap P_1],$$

and as none of the other P_i are equivalent to P_1, we may conclude that $[P_1 : P_i \cap P_1]$ is divisible by p, as $P_i \cap P_1$ is a proper subgroup of P_1. So $n \equiv 1 \pmod{p}$.

By the way, if H is a Sylow p-subgroup of G and H is not equal to any of the P_i, then

$$n = [H : P_1 \cap H] + ... + [H : P_a \cap H]$$

and each of the $[H : P_i \cap H]$ is divisible by p, as $P_i \cap H$ is a proper subgroup of H. This contradicts the fact that $n \equiv 1 \pmod{p}$. So actually, every Sylow p-subgroup must be one of the P_i. In other words, every Sylow p-subgroup is conjugate to every other Sylow p-subgroup. There are two important implications here.

- The number of Sylow p-subgroups is equal to n, which is congruent to 1 modulo p.

[16] The proof relies on the fact that $|HK| = \frac{|H||K|}{|H \cap K|}$ for any $H, K \leq G$, which we have neglected to prove for purposes of streamlining this text.

- Exercise 2.40 implies that $n = [G : N_G(P)]$. Notably, as $[G : N_G(P)]$ divides $[G : P]$, we must have that n divides $[G : P]$.

There is one final question we can ask. We know there is a subgroup of order p and order p^k, but what about the powers of p in between? It turns out that for any $a \leq k$, there is a subgroup of G with order p^a. In fact, we can find a subgroup of any Sylow p-subgroup with order p^a. The proof is nearly identical to that of the existence of a Sylow p-subgroup. Here is how it goes.

- Induct on k.
- Cauchy's says there is some $N \trianglelefteq G$ with order p, now $\left|\frac{G}{N}\right| = p^{k-1}$. By the inductive hypothesis there is some $\frac{H}{N} \trianglelefteq \frac{G}{N}$ with order p^{k-2}, so $|H| = p^{k-1}$. To achieve all smaller a, just look at H.

Summarized, here are all of Sylow's Theorems. We will be fairly exhaustive when listing them out.

> **Theorem 2.45 (Sylow's Theorems)**: Given a finite group G whose order is divisible by p^k and not p^{k+1}, denote $\mathrm{Syl}_p(G)$ as the set of subgroups of G with order p^k. Then
>
> - $\left|\mathrm{Syl}_p(G)\right| \equiv 1 \pmod{p}$.[17]
> - $\left|\mathrm{Syl}_p(G)\right|$ divides $[G : P]$.
> - For every $P \in \mathrm{Syl}_p(G)$, $\left|\mathrm{Syl}_p(G)\right| = [G : N_G(P)]$.
> - All the Sylow p-subgroups are conjugate to each other, that is, if $P, H \in \mathrm{Syl}_p(G)$, then there is some $g \in G$ such that $gPg^{-1} = H$.
> - Every Sylow p-subgroup has a normal subgroup of order p^a for all $a \leq k$.

2.6 Series of Subgroups

2.6.1 Solvability and Nilpotence

This section marks a transition in our treatment of group theory. In the introduction, we have stated that rigor and full proofs will not be as emphasized in this primer. So far we have actually been fairly rigorous. This is because the proofs so far have been relatively short. Now the proofs get longer, and so we will skip many of them.

We have often asked, for subgroups $H, K \leq G$, what does $\{kHk^{-1} : k \in K\}$ look like? In other words, what is the orbit of conjugation on H via K? To put it even more abstractly, how "normal" or even "abelian" is H in relation to K?

This is all well and good, but we are missing a certain aspect of symmetry. Instead, it may make sense to ask what the set $\{hkh^{-1}k^{-1} : h \in H, k \in K\}$ looks like.

[17] This implies there must exist a Sylow p-subgroup, for 0 is not congruent to 1 modulo p.

> **Definition 2.46 (Commutator)**: The **commutator** of elements $h, k \in G$ is $[h, k] = hkh^{-1}k^{-1}$. The **commutator subgroup** of subgroups $H, K \leq G$ is the smallest subgroup $[H, K]$ containing the set of commutators between H and K, i.e. $\{hkh^{-1}k^{-1} : h \in H, k \in K\}$.

Pay close attention to the fact that the commutator subgroup is "the smallest subgroup containing the set of commutators". This is because the set of commutators is not necessarily closed under multiplication.

On that note, we have never described the smallest subgroup containing a subset $S \subseteq G$. It turns out to be what you'd expect: the set of all elements you get by repeatedly multiplying and taking inverses of the elements in S. In symbols, this subgroup is

$$\{s_1 ... s_n : s_i \in S \text{ or } s_i^{-1} \in S, n \in \mathbb{N}\}.$$

It may be fruitful to prove this is true.[18]

The commutator subgroup is in some sense a measure of how commutative H and K are. If they are commutative, then every commutator will turn out to be 1 and the commutator subgroup is thus trivial. And if they are not commutative, and H and K are sufficiently big, then it is possible for $[H, K] = G$. And results in between, where the commutator subgroup is not trivial nor the entirety of G, are also possible.

The commutator subgroup is also a **symmetric** measure. More precisely, it is the case that $[H, K] = [K, H]$. (This is worth proving!)

A natural question to ask is how commutative G itself is. And a measure of that is $[G, G]$. In fact, there is a notion in which it is the "best" measure of how abelian G is.

> **Theorem 2.47**: Given a group G, $[G, G]$ is the smallest group such that $\frac{G}{[G,G]}$ is abelian.
> - If $H \trianglelefteq G$ and $\frac{G}{H}$ is abelian, then $[G, G] \leq H$.
> - If $H \trianglelefteq G$ and $[G, G] \leq H$, then $\frac{G}{H}$ is abelian.

Proof of Theorem 2.47: First we ought to show $\frac{G}{[G,G]}$ is well-defined, i.e. $[G, G] \trianglelefteq G$. This is easy once you recall $-^g$ is an automorphism (see Exercise 2.23), as for all g_1, g_2, $g_3 \in G$,

$$\begin{aligned} [g_1, g_2]^{g_3} &= \left(g_1 g_2 g_1^{-1} g_2^{-1}\right)^{g_3} \\ &= g_1^{g_3} g_2^{g_3} \left(g_1^{-1}\right)^{g_3} \left(g_2^{-1}\right)^{g_3} \\ &= \left[g_1^{g_3}, g_2^{g_3}\right]. \end{aligned}$$

[18] This is the sort of fact covered in the start of a proper abstract algebra text, whereas here we omit the proof because this fact feels true enough.

Strictly we have only shown that $-^g$ maps generators of $[G, G]$ to other generators of $[G, G]$, but as $-^g$ is an automorphism, that is enough.

Now for all $g_1, g_2 \in G$,

- $(g_1 H)(g_2 H)(g_1^{-1} H)(g_2^{-1} H) = (g_1 g_1^{-1} g_2 g_2^{-1})H = H$ as $\frac{G}{H}$ is abelian, implying that $g_1 g_2 g_1^{-1} g_2^{-1} \in H$.
- Note $[G, G] \trianglelefteq G$ and $H \leq G$ yields $[G, G] \trianglelefteq H$ (see Exercise 2.28). So Theorem 2.32 (Third Isomorphism) yields
$$\frac{G}{H} = \frac{\frac{G}{[G,G]}}{\frac{H}{[G,G]}}.$$
But since $\frac{G}{[G,G]}$ is abelian, we know that $\frac{G}{H}$ is.[19]

□

The specific fact that $[G, G] \trianglelefteq G$ will become very useful soon.

Another related question will also arise quite naturally: "what if I kept taking commutator subgroups on G?" There are two ways in which we can "keep taking commutator subgroups":

- Define $G^0 = G$ and $G^{(n+1)} = \left[G^{(n)}, G^{(n)}\right]$. This yields the **derived series**.
- Define $G^0 = G$ and $G^{n+1} = [G^n, G]$. This yields the **lower central series**.

In both the derived and lower central series, the subgroups G_n decrease. More precisely, for all n, $G^{(n+1)} \leq G^{(n)}$ and $G^{n+1} \leq G^n$. This is obvious for the derived series (closure of groups) but not so obvious for the lower central series. It turns out the latter decreases because $G_n \trianglelefteq G$ for each n. This can easily be proven by induction once you show the following fact.

Exercise 2.48: Suppose $H \trianglelefteq G$. Show that $[H, G] \trianglelefteq G$.

It is also the case that $H \trianglelefteq G \iff [H, G] \leq H$. Simply use the fact that normality means $ghg^{-1} \in H$ for all $h \in H$. Combining this fact with with Exercise 2.48 easily yields that $G^{n+1} \leq G^n$ for all $n \in \mathbb{N}$.

So a natural question to ask is: given these sequences decrease, how long does it take for these sequences to reach id (the trivial subgroup)? More pertinently, for which groups do these sequences reach id?

- If the derived series of G reaches id (i.e. there is some n such that $G^{(n)} = \text{id}$), then we say G is **solvable**. And the smallest n such that $G^{(n)} = \text{id}$ is the **derived length**.
- If the lower central series of G reaches id (i.e. there is some n such that $G_n = \text{id}$), then we say G is **nilpotent**. And the smallest n such that $G^n = \text{id}$ is the **nilpotence class**.

For obvious reasons, $G^{(n)} \leq G^n$ for all $n \in \mathbb{N}$. So every nilpotent group is solvable.

Example 2.49: But not every solvable group is nilpotent. For example, S_3 (i.e. the group of permutations of a set with 3 elements) is not nilpotent. But it is solvable.

[19] It is trivial to show in general that if G is abelian then $\frac{G}{H}$ is abelian for any $H \leq G$.

Why does the derived length matter? Because it is the shortest length of any series that "look like the derived series". (Same for the nilpotence class.)

> **Theorem 2.50**: A group G is **solvable** if and only if there is a series
>
> $$G = G_0 \trianglerighteq G_1 \trianglerighteq \ldots \trianglerighteq G_k = \mathrm{id}$$
>
> where G_i / G_{i+1} is abelian for all i.
>
> Furthermore, if G is solvable, then the length of this series is at least the length of the derived series.

Proof of Theorem 2.50: If G is solvable then its derived series is such a series and there is nothing to show.

If such a series exists, because $G^{(n+1)}$ is the smallest normal subgroup of $G^{(n)}$ where $G^n \trianglerighteq G^{(n+1)}$, we may easily show by induction that $G_i \geq G^{(i)}$ for all i. Thus G is solvable and the length of this series is at least the length of its derived series. □

Now we turn to other series of normal subgroups.

> **Definition 2.51 (Subnormal Series)**: A **subnormal series** is a series $\mathrm{id} = G_0 \trianglelefteq \ldots \trianglelefteq G_n = G$.

The derived series is a subnormal series.

> **Definition 2.52 (Central Series)**: A **central series** is a subnormal series
>
> $$\mathrm{id} = G_0 \trianglelefteq \ldots \trianglelefteq G_n = G$$
>
> where $\frac{G_{i+1}}{G_i} \leq Z\left(\frac{G}{G_i}\right)$, or equivalently, $[G, G_{i+1}] \leq G_i$.

A priori it is not obvious the two conditions are equivalent, but for purposes of conciseness we omit the proof. We will mostly be using the first characterization.

By the way, when we turn the inequality into an equality, i.e. construct a series where $\frac{G_{i+1}}{G_i} = Z\left(\frac{G}{G_i}\right)$, we get the **upper central series**. It is also a fact that the length of the upper central series is equal to the length of the lower central series.[20]

It turns out that

- a central series exists if and only if G is nilpotent,
- and supposing a central series exists, n is at least the nilpotence class of G.

[20] You can prove this via induction on the nilpotence class.

There is a somewhat surprising logarithmic relation between the derived length and nilpotence class. Suppose l and c are the derived length and nilpotence class of some group G, respectively. Then $l \leq \log_2(c) + 1$. We will not prove this, but I did want to mention it since this is an interesting fact.

2.6.2 Composition Series

Composition series can be used to "factor" a group into a chain of normal groups.

> **Definition 2.53 (Simple Group)**: A group G is **simple** if and only if the only normal subgroups of G are id and G.

> **Definition 2.54 (Composition Series)**: A **composition series** is a subnormal series
>
> $$\mathrm{id} = G_0 \trianglelefteq \ldots \trianglelefteq G_n = G$$
>
> where each $\frac{G_{i+1}}{G_i}$ is simple.

Every finite group has a composition series. Here is how you would go about finding one.

- Look at the group G. Does it have any interesting normal subgroups? If not, we are done.
- So suppose $N \trianglelefteq G$ where $N \neq \mathrm{id}, G$. Then repeat this process for N and $\frac{G}{N}$, and use Theorem 2.33 to get a series $N = G_k \trianglelefteq \ldots \trianglelefteq G_n = G$ from our series $\frac{N}{N} = H_0 \trianglelefteq \ldots \trianglelefteq H_{n-k} = \frac{G}{N}$.
- Then we may use our series $\mathrm{id} = G_0 \trianglelefteq \ldots \trianglelefteq G_k = N$, and then we may adjoin our series $N = G_k \trianglelefteq \ldots \trianglelefteq G_n = G$ to create a series for G.

To be explicit, in each of the chains we have constructed, successive quotients are **simple**.

We know this process will stop because the order of G is finite. In fact, this can easily be rephrased as a proof by induction which shows the existence of a composition series for every finite group.

For instance, the only composition series of Z_6 are

$$\mathrm{id} \trianglelefteq Z_2 \trianglelefteq Z_6 \text{ and } \mathrm{id} \trianglelefteq Z_3 \trianglelefteq Z_6.$$

What do you notice about the quotients of successive subgroups? It turns out they are the same. Because we may "recursively" construct composition series by "splitting up" the group into N and $\frac{G}{N}$, we have reason to suspect that any two composition series are essentially identical. It turns out they are.

Theorem 2.55 (Jordan-Holder): If a group G has composition series

$$\mathrm{id} = G_0 \trianglelefteq \dots \trianglelefteq G_n = G \text{ and } \mathrm{id} = H_0 \trianglelefteq \dots \trianglelefteq H_m$$

then $n = m$, and we may pair quotients of the form G_{i+1}/G_i with quotients of the form H_{j+1}/H_j such that $G_{i+1}/G_i \cong H_{j+1}/H_j$. More formally, there is a bijection π such that $G_{\pi(i)}/G_{\pi(i)-1} \cong H_i/H_{i-1}$.

The proof is painful. You may read one at https://dennisc.net/jordan-holder.pdf.

There is a natural analogy between groups uniquely factoring into simple quotients and positive integers uniquely factoring into primes. At the same time, every prime factorization is uniquely associated with a positive number. So it is natural to ask whether a factorization of simple quotients is uniquely associated with a group.

The answer is no. For instance,

$$\mathrm{id} \trianglelefteq Z_p \trianglelefteq Z_{p^2} \text{ and } \mathrm{id} \trianglelefteq Z_p \trianglelefteq Z_p \times Z_p$$

are both composition series with the same quotients, but clearly Z_{p^2} and $Z_p \times Z_p$ are not isomorphic.

2.6.3 Why should you care?

Now that we have spent a good chunk of time analyzing groups and their properties — subgroups, normality, actions, solvability, and more — it is natural to ask, “who cares?” And this is a genuinely good question. Up until now, you have been given no reason (perhaps besides Exercise 2.42) to care about groups whatsoever. We have developed this complicated machinery yet have not covered any but the most trivial of uses for group theory outside of more group theory.

The most elementary use of these concepts, normality and solvability in particular, is in Galois Theory. Suppose that we have fields $F \leq K$ and would like to find the fields $F \leq L \leq K$. Galois Theory not only tells us how many of these fields there are, but how they are structured. Among other things, this can be used to prove the Fundamental Theorem of Algebra. Furthermore, you can determine whether a polynomial like $x^5 - x - 1$ has algebraic roots or not. (It does not.) And all this is done through the machinery of groups.

As for groups themselves, they have many interesting real-world applications, none of which I am qualified to speak on. Suffice it to say that reality informs us we should care about groups, much the same way it informs us we should care about real analysis.

These series of subgroups, particularly the composition series as discussed earlier, are powerful tools that can be used to characterize groups and understand their structure. In fact, that is what we are about to do next.

2.7 Characterizing Groups of Finite Order

So far we have spent a lot of time discussing what groups do: they commute, act, normalize, solve, etc. We now give a brief overview of what they look like. (Because this section is fairly non-central to the rest of our discussion of algebra, we will choose to omit the vast majority of proofs.)

We may use Jordan-Holder to gain some degree of understanding about individual groups. Furthermore, we may factor finite abelian groups and finite nilpotent groups into direct products.

Now what if we wanted to characterize all the groups of order n up to isomorphism for some fixed positive integer n? Suppose G is a group with order n.

For purposes of having a concrete example we will set $n = 21$. Here is the overall strategy. We use Sylow's Theorem to determine that G must have a normal subgroup N of order 7, as

- the number of Sylow 7-subgroups must divide 3,
- and this number must also be congruent to 1 modulo 7,

so there exists a unique subgroup of order 7. (Meaning it is normal as gNg^{-1} is also a subgroup of order 7, so gNg^{-1} better be the same as N.)

Cauchy's also states that we must have some subgroup H of order 3 (H need not be normal). Notably, $N \cap H = \text{id}$ and $NH = G$. Importantly, every element in G can be **uniquely** represented as a product nh with $n \in N$ and $h \in H$. Now bear with me as we define the **semidirect product** seemingly at random. You will see why it matters very soon.

> **Definition 2.56 (Semidirect Product)**: Suppose N and H are groups and $\alpha : N \to \text{Aut}(H)$ is a homomorphism from N to the automorphisms of H. Then we define the group $N \rtimes_\alpha H$ as follows: the elements are ordered pairs (n, h) with $n \in N$ and $h \in H$, and the group operation is defined as
>
> $$(n_1, h_1)(n_2, h_2) = (n_1\alpha(h_1)(n_2), h_1h_2).$$

Note that $\alpha(n_1)$ returns a **function** — in particular, an automorphism in H — which is why we can then apply it to h_2.

It is not immediately obvious that the group operation defined is associative and admits an inverse, but performing the verifications is straightforward.

Now, supposing that $N \cap H = \text{id}$ and $NH = G$, if we set $\alpha(n) = nhn^{-1}$, we get that G is isomorphic to $N \rtimes_\alpha H$. The isomorphism is the obvious one: define $\varphi : G \to N \rtimes_\alpha H$ as $\varphi : nh \mapsto (n, h)$. I highly suggest you check this isomorphism is actually 1) well-defined and 2) a homomorphism (bijectivity is obvious).

So **if** G is a group of order 21, **then** it must be isomorphic to some $N \rtimes_\alpha H$, where N is a group of order 7 and H is a group of order 3. Furthermore, **every** valid triple $N, H, \rtimes_\alpha$

admits a group of order 21 as the semidirect product always forms a group. So finding these groups of order 21 reduces to finding the semidirect products that arise from these triples.

Now it is important to note that **different α may yield isomorphic groups**. In fact, supposing we fix the structure of N and H, if there is some automorphism $\beta : H \to H$, then $h \mapsto \alpha(\beta(h))$ is a homomorphism from H to $\operatorname{Aut} N$, and furthermore, $N \rtimes_{\alpha} H \cong N \rtimes_{\alpha \circ \beta} N$. The automorphism is

$$(n, h)_{\alpha} \mapsto (n, \beta^{-1}(h))_{\alpha \circ \beta}.$$

I highly recommend you verify this.

Now we finish our analysis on the groups of order 21. The structure of N and H are fixed as there is only one group of order 3 and one group of order 7, the cyclic groups. Say H is generated by the element h, i.e. $h \neq \mathrm{id} \in H$. Since $\operatorname{Aut} N$ is cyclic with order 6 (you should verify this yourself), we may conclude that the only homomorphisms from H to $\operatorname{Aut} N$ are the α_i where

$$\alpha_1(h) = \mathrm{id}, \alpha_2(h) = g \mapsto g^2, \text{and } \alpha_3(h) = g \mapsto g^4.$$

(Note $\alpha_i(h)$ uniquely determines α_i as H is cyclic.) Obviously α_1 yields the group $\mathbb{Z}/7\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$, which is abelian. And obviously α_2 and α_3 do not yield abelian groups (it is not hard to find an explicit counterexample).

Now here's the kicker: α_2 and α_3 yield isomorphic groups. For $\beta : h \mapsto h^2$ is an automorphism in H, and $\alpha_3 = \alpha_2 \circ \beta$. So we actually have two distinct groups of order 21, one abelian and the other not abelian.

2.8 Free Groups and Presentations

A word of warning: this section is fairly advanced. We will be introducing categorical theoretic ideas, ideas which are quite difficult to grasp at first. There is no shame in skipping this section.

We will work backwards in this section: first we will present an informal idea of a **group presentation**. Then we will formally define a **free group**, and finally we will use it to formally define a group presentation.

2.8.1 Presentations, informally

Recall the dihedral group D_{2n}. It is informally defined geometrically: we may compose rotations with reflections. In fact, we have never defined it formally, so we may as well make an attempt now.

Suppose we refer to the group element associated with a rotation shifting each vertex by 1 space as r.[21] Furthermore select any arbitrary reflection and denote the corresponding group element as s. Notice that

- r and s generate D_{2n},

[21] It does not matter what direction the rotation is in.

- $r^n = s^2 = (rs)^2 = \text{id}$.

We say the **group presentation** of this group is

$$\langle r, s \mid r^n, s^2, (rs)^2 \rangle.$$

On the left side are the generators of the group. On the right side are the relations the generators obey. To elaborate, each relation on the right side is a group element that is defined to be equivalent to id

Intuitively, it feels like $r^n = s^2 = (rs)^2$ is the correct amount of specificity to fully describe the group, no more or less. We want the relations to be equal to id, but nothing "extra". Furthermore we want the presentation to be distinct from a "looser" set of relations, such as

$$\langle r, s \mid s^2, (rs)^2 \rangle.$$

We would like to formalize these ideas. The looser set of relations is satisfied by (our intuitive notion of) D_{2n}. Perhaps more scarily, we can set $r = s = \text{id}$ and declare that these presentations really describe the trivial group. And since we have no formal definition of D_{2n} to fall back on, who's to say this is wrong?

This is why we need a more precise idea of a group presentation.

2.8.2 Free groups and presentations, formally

Define an **alphabet** Σ as a set of **characters**. Note that there is nothing stopping Σ from being infinite! For convenience we will say $\Sigma = \{a, b\}$.

Formally, a **word** is a finite sequence in $\Sigma \times \mathbb{Z}$. Informally, it is something like aba^2b^{-3}. We will write words in this fashion rather than saying the word is the sequence $(a, 1), (b, 1), (a, 2), (b, -3)$.

Something like $aa^{-1}b$ is a valid word. But as the exponents suggest, you ought to "simplify" it down to b, because a and a^{-1} ought to "multiply and cancel out". Your intuition here would be correct.

> **Definition 2.57 (Reduced Word)**: Formally, a **reduced word** is a word where
>
> - no two consecutive elements of the sequence have the same character,
> - no element of the sequence has an exponent of 0.

We may reduce words according to the following algorithm: while either of the following still remain,

- Combine a pair of consecutive elements with the same character, summing up their exponents. For example, replace a^2a^3 with a^5.
- Remove any element with an exponent of 0. For example, remove a^0 from the sequence.

As an example,

$$aa^{-1}b \longrightarrow a^0b \longrightarrow b$$

is the correct series of reductions, and we say b is the **reduced form** of $aa^{-1}b$.

On the face of it, this algorithm is not generally deterministic: we may select any pair of consecutive elements with the same character to combine, and we may select any element with an exponent of 0 to eliminate. But it turns out that regardless of which order you reduce in, you end up with the same reduced word.

We use **reduced words** to define the **free group** on an alphabet Σ.

> **Definition 2.58 (Free Group)**: The **free group** on an alphabet Σ is the group where
>
> - the underlying set is the set of reduced words in Σ,
> - the group operation is "concatenate two words, then reduce the result".
>
> We denote this free group as $\text{Free}(\Sigma)$.

For example, $ab^2 \cdot b^{-2}ab = a^2b$.

It turns out that $\text{Free}(S)$ is uniquely determined by the cardinality of the set S. Obviously if two sets A and B have the same cardinality, then $\text{Free}(A) \cong \text{Free}(B)$. It is not so obvious that if $|A| \neq |B|$ then $\text{Free}(A)$ and $\text{Free}(B)$ are not isomorphic. There are clever ways to prove this, but we will outline a more straightforward proof instead.

Exercise 2.59: Fill in the details for the following proof that $|A| \neq |B|$ implies $\text{Free}(A)$ and $\text{Free}(B)$ are not isomorphic:

1. Argue that $|\text{Free}(A)| = |A|$ when A is uncountable.
2. Suppose A is finite and $|A| \leq |B|$. Consider a homomorphism $\varphi : \text{Free}(A) \to \text{Free}(B)$. View $\text{Free}(A)$ as a $|A|$-dimensional vector space over $\mathbb{Z}$ where the ath entry is the sum of the exponents of all the a-terms. Do the same for B. Now φ can be viewed as a linear map between the vector spaces $\text{Free}(A)$ and $\text{Free}(B)$. Conclude that φ cannot be surjective.

Thus if λ is a cardinal, we frequently just write $\text{Free}(\lambda)$.

> **Definition 2.60 (Presentation)**: A **presentation** of a group with a generating set Σ and a relation set R is the quotient group $\frac{\text{Free}(\Sigma)}{N}$, where N is the smallest normal subgroup containing R.

Note that the **smallest normal subgroup** containing R is the smallest subgroup containing R and all of its conjugates in $\text{Free}(\Sigma)$.

2.8.3 The Universal Mapping Property

We first define some notation.

Definition 2.61 (Restrictions): Consider a function $f : A \to B$. Let $S \subseteq A$. Then we define the function $f \upharpoonright S : S \to B$ to be the unique function where for all $s \in S$, $(f \upharpoonright S)(s) = f(s)$. This is the **restriction** of f under S.

Essentially, $f \upharpoonright S$ is identical to f, but it only takes in values from S.

Free groups are interesting because they turn out to satisfy a **universal property**.

Theorem 2.62 (Universal Property of the Free Group): Given a set Σ, a group G, and a function $\varphi : \Sigma \to G$, there is a unique homomorphism $\psi : \text{Free}(\Sigma) \to G$ such that $\psi \upharpoonright \Sigma = \varphi$.

$$\begin{tikzcd}
\Sigma \arrow[r, "\text{id}"] \arrow[dr, "\varphi"'] & \text{Free}(\Sigma) \arrow[d, "\psi"'] \\
 & G
\end{tikzcd}$$

Figure 2: Equivalently, there is a unique ψ such that the given diagram commutes.

We are abusing notation a little here. The identity function on Σ technically maps to Σ rather than $\text{Free}(\Sigma)$, but I trust you know what I mean when I write id.

Proof of Theorem 2.62: To prove this theorem we need to show two things:

1. that there exists some satisfactory ψ,
2. and that any two satisfactory ψ_1 and ψ_2 must be one and the same.

Both parts are straightforward:

1. we may easily check that $\psi : g_1^{n_1}...g_k^{n_k} \mapsto \varphi(g_1)^{n_1}...\varphi(g_k)^{n_k}$ suffices,
2. and in general, the behavior of a homomorphism on a group is determined by its behavior on the generators, which means ψ_1 and ψ_2 are the same as they behave identically on generating set Σ.

□

The First Isomorphism Theorem with Theorem 2.62 implies any group can be represented with a presentation.

$$\begin{tikzcd}
G_{\text{Set}} \arrow[r, "\text{id}"] \arrow[dr, "\text{id}"] & \text{Free}(G_{\text{Set}}) \arrow[d, "\psi"] \\
& G_{\text{Group}}
\end{tikzcd}$$

Figure 3: There is a unique ψ such that $\text{id} \circ \psi = \text{id}$. By the First Isomorphism Theorem, $\frac{\text{Free}(G_{\text{Set}})}{\ker \psi} \cong \operatorname{im} \psi = G_{\text{Group}}$. Since $\ker \psi$ is obviously normal, $\langle G_{\text{Set}} \mid \ker \psi \rangle$ is a presentation of G.

Furthermore, it means that we can describe a homomorphism $\psi : \text{Free}(\Sigma) \to G$ merely by describing a function $\varphi : \Sigma \to G$. This is why presentations are even remotely viable: they can describe any arbitrary group, and they can be used to specify the behavior of homomorphisms.

Universal properties are a fundamental idea in category theory. Because the concept of a universal property will appear many more times (notably when we construct the tensor product), we will very quickly develop the relevant category theoretic concepts to appreciate universal properties in their full generality.

2.8.4 Category theory, briefly

Very often, we study some class of objects and functions between them, such as groups and homomorphisms, or topological spaces and continuous functions. And very often, patterns will emerge through wildly different classes of objects and functions. Thus in category theory, we abstract to the level of studying objects of an arbitrary type and functions between them.

Before we formally define a category, let us first give an informal picture of how they should behave. In mathematics there is a general idea of studying **structured sets** and **functions that preserve the structure**. For example, we might study groups and homomorphisms. Or topologies and homeomorphisms. There are concepts and connections that are universal between different structured sets/functions. Category theory is all about studying these universal concepts.

In category theory, we study **categories** which consist of **objects** and **arrows** between objects. Objects correspond to structured sets. Arrows correspond to structure-preserving functions. And arrow composition follows much the same rules as function composition.

For example, there is a category called **Group** whose objects are groups and arrows are homomorphisms between groups. And arrow composition is simply homomorphism composition.

Now let us concretely define a category.

Definition 2.63 (Category): A category $\mathcal{C}$ consists of a **class** of objects, a **class** of arrows, and a partial binary operation $\circ$ known as **composition** that takes arrows to arrows, where

- each arrow f is associated with a **domain** $\operatorname{dom} f$ and **codomain** $\operatorname{cod} f$, both of which are objects in $\mathcal{C}$. We will write $f : a \to b$ to denote that f has domain a and codomain b;
- for any two arrows f and g where $\operatorname{cod} f = \operatorname{dom} g$, there exists an arrow $g \circ f$ with $\operatorname{dom} (g \circ f) = \operatorname{dom} f$ and $\operatorname{cod} (g \circ f) = \operatorname{cod} g$. (If $\operatorname{cod} f \neq \operatorname{dom} g$, then $g \circ f$ **is not a defined object** and it makes no sense to say $g \circ f$.)
- The associative law is satisfied by $\circ$; that is, $h \circ (g \circ f) = (h \circ g) \circ f$, supposing that all compositions are well-defined;
- for every object a, there exists an arrow $\mathrm{id}_a : a \to a$ such that for every arrow $f : x \to a$, $\mathrm{id}_a \circ f = f$, and for every arrow $g : a \to x$, $g \circ \mathrm{id}_a = g$.

I am very deliberately using the word **class** in lieu of **set**. Very briefly, in mathematics we are interested in collections of objects. Only some collections of objects may be constructed as sets. For instance, the collection of all sets is not a set (for no set may contain itself), so we must consider it as a class. For more details, you may peruse https://dennisc.net/writing/blog/sets-classes.

This distinction is important because we want to study the category of all sets whose arrows are functions between sets, which we will aptly refer to as Set from now on. This will be our prototypical example of a category, precisely because it highlights the analogy between arrows in a category and functions between sets.

$$\begin{tikzcd}
\mathbb{Z} \arrow[r, "x \mapsto |x|"] & \mathbb{N}
\end{tikzcd}$$

Figure 4: $\mathbb{Z}$ and $\mathbb{N}$ are objects of Set, and $x \mapsto x : \mathbb{Z} \to \mathbb{N}$ is an arrow with domain $\mathbb{Z}$ and codomain $\mathbb{N}$.

Two important observations:

- Two distinct arrows may have the same domain and codomain, such as the functions $\mathrm{id} : \mathbb{Z} \to \mathbb{Z}$ and $x \mapsto -x : \mathbb{Z} \to \mathbb{Z}$.
- The domain/codomain requirements for composing arrows is identical to those for composing functions.

Now we define the **dual** of a category and the **Cartesian product** of two categories. (The Cartesian product, in particular, is exactly what you think it is.) These definitions will come in handy when we study **hom-sets**, but for now they serve as examples of categorical constructions.

Definition 2.64 (Dual Category): The **dual** of category $\mathcal{C}$ is the category $\mathcal{C}^{\text{op}}$ where every arrow has its domain and codomain switched.

Definition 2.65 (Product Category): The **product** of two categories $\mathcal{C}$ and $\mathcal{D}$ is the category $\mathcal{C} \times \mathcal{D}$ where

- objects are of the form (c, d), where c is an object in $\mathcal{C}$ and d is an object in $\mathcal{D}$,
- arrows are of the form (f, g), where f is an arrow in $\mathcal{C}$ and g is an arrow in $\mathcal{D}$,
- composition is defined pairwise, that is, $(f_2, g_2) \circ (f_1, g_1) = (f_2 \circ f_1, g_2 \circ g_1)$,
- and the identity arrow associated with (c, d) is $(\text{id}_c, \text{id}_d)$.

Example 2.66: We may define the category of sets $\mathcal{C}$ where there is a unique arrow $X \to Y$ if and only if $X \subset Y$. Then there is a unique arrow $Y \to X$ in $\mathcal{C}^{\text{op}}$ precisely when $Y \supset X$.

The reason category theory is interesting is because we can draw comparisons between categories. Just as a group homomorphism is a structure-preserving map between groups, a **functor** is a structure-preserving map between categories.

Definition 2.67 (Functor): More precisely, a **functor** $F : \mathcal{C} \to \mathcal{D}$ is a map that

- sends objects in $\mathcal{C}$ to objects in $\mathcal{D}$,
- sends arrows in $\mathcal{C}$ to arrows in $\mathcal{D}$,
- satisfies $F(\text{id}_a) = \text{id}_{F(a)}$,
- and satisfies $F(g \circ_{\mathcal{C}} f) = F(g) \circ_{\mathcal{D}} F(f)$ for all arrows f, g in $\mathcal{C}$ where $g \circ_{\mathcal{C}} f$ is well-defined. In particular, this definition forces $F(g) \circ_{\mathcal{D}} F(f)$ to be well-defined.

Here is a trivial example of a functor, the **forgetful functor on groups**. Consider the category of groups and the category of sets. There is a trivial functor $F : \text{Group} \to \text{Set}$ where $F : G \mapsto G$ and $F : \varphi \mapsto \varphi$. In other words, F acts on objects by mapping groups to their underlying set, and F acts on arrows by mapping homomorphisms to themselves, where these homomorphisms are just considered as functions in Set.

This is called the **forgetful functor** because we are literally forgetting the underlying structure of Group when we take it to Set. We forget the group structure, treating groups as just their underlying sets and homomorphisms as regular old functions.

The forgetful functor is important for our purposes because the **free functor** is loosely defined as the **adjoint** to a forgetful functor. (We will shortly define what an adjoint pair of functors is.) And as the name suggests, the functor $S \mapsto \text{Free}(S) : \text{Set} \to \text{Group}$ is such a free functor.

I have not yet specified how the free functor $S \mapsto \mathrm{Free}(S)$ maps arrows in Set to arrows in Group. There is a sensible definition which relies on Theorem 2.62: we send a function $f : X \to Y$ in Set to the unique homomorphism $\varphi : \mathrm{Free}(X) \to \mathrm{Free}(Y)$ in Group such that $\varphi \upharpoonright X = f$.[22] This homomorphism uniquely exists because we can easily extend $f : X \to Y$ to $f : X \to \mathrm{Free}(Y)$ as $Y \subset \mathrm{Free}(Y)$.[23]

$$\begin{tikzcd}
X \arrow[r, "\mathrm{id}"] \arrow[dr, "f"'] & \mathrm{Free}(X) \arrow[d, "\varphi"'] \\
& \mathrm{Free}(Y)
\end{tikzcd}$$

Figure 5: Applying Theorem 2.62 shows that φ is unique.

Exercise 2.68: Verify that the free functor F is indeed a functor. More precisely, verify that $F(g \circ f) = F(g) \circ F(f)$.

(Technically you also need to verify that $F(\mathrm{id}_S) = \mathrm{id}_{\mathrm{Free}(S)}$, but this is trivial.)

Now, as promised, we will formally introduce adjoint functors.

Definition 2.69 (Adjoint Functors): Functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$ **adjoint** if for each object c in $\mathcal{C}$, we may associate c with an arrow $\eta_c : c \to G(F(c))$ such that for each object d in $\mathcal{D}$ and arrow $f : c \to G(d)$, there exists a unique arrow $g : F(c) \to d$ with $G(g) \circ \eta_c = f$.

$$\begin{tikzcd}
c \arrow[r, "\eta_c"] \arrow[dr, "f"] & G(F(c)) \arrow[d, "G(g)"] \\
& G(d)
\end{tikzcd}$$

Figure 6: This diagram commutes when G is right adjoint to F.

Note that the statements “F and G are adjoint” and “G and F are adjoint” are different. In other words, order matters. To that end, we will often refer to F as the **left adjoint** and G as the **right adjoint**.

[22] Technically this is an abuse of notation. Really, we are identifying an element $f(x) \in Y$ with the word $y^1 \in \mathrm{Free}(Y)$. They are technically not the same thing, but you and I know they are basically equivalent.

[23] Again, this is not exactly true. But it is close enough to true.

Exercise 2.70: Convince yourself the free functor $S \mapsto \mathrm{Free}(S) : \mathrm{Set} \to \mathrm{Group}$ and forgetful functor $G \mapsto G : \mathrm{Group} \to \mathrm{Set}$ are adjoint. (Hint: compare Figure 2 with Figure 6.)

This definition of adjunction is woefully incomplete. There are two other equivalent characterizations: the **hom-set** and **counit** definitions. But first we must develop the concepts of **natural transformations** and **hom-sets**.

Definition 2.71 (Natural Transformation): Given functors $F, G : \mathcal{C} \to \mathcal{D}$, a **natural transformation** $\eta : F \Rightarrow G$ assigns each object a in $\mathcal{C}$ an arrow $\eta_a : F(a) \to G(a)$ such that for each arrow $f : a_1 \to a_2$ in $\mathcal{C}$,

$$\eta_{a_2} \circ F(f) = G(f) \circ \eta_{a_1}.$$

$$\begin{tikzcd}
a_1 \arrow[d, "f"'] & F(a_1) \arrow[r, "\eta_{a_1}"] \arrow[d, dashed, "F(f)"'] & G(a_1) \arrow[d, "G(f)"'] \\
a_2 & F(a_2) \arrow[r, dashed, "\eta_{a_2}"] & G(a_2)
\end{tikzcd}$$

Figure 7: Equivalently, this diagram commutes. The two different paths from $F(a_1)$ to $G(a_2)$ are marked with stylistically distinct arrows.

Let's write out the domain and codomain of each arrow involved in the definition.

- We compose $F(f) : F(a_1) \to F(a_2)$ and $\eta_{a_2} : F(a_2) \to G(a_2)$ to get an arrow $\eta_{a_2} \circ F(f) : F(a_1) \to G(a_2)$.
- We compose $\eta_{a_1} : F(a_1) \to G(a_1)$ and $G(f) : G(a_1) \to G(a_2)$ to get an arrow $G(f) \circ \eta_{a_1} : F(a_1) \to G(a_2)$.

Note every arrow involved in this condition is in $\mathcal{D}$!

Just as a functor is not a category, a natural transformation is not a functor. Just as a functor is a map between two categories, a natural transformation is a map between two functors.

Furthermore, we may compose natural transformations.

Definition 2.72 (Composition of Natural Transformations): Given functors $F, G, H : \mathcal{C} \to \mathcal{D}$ and natural transformations $\eta : F \Rightarrow G$ and $\varepsilon : G \Rightarrow H$, we define $\varepsilon \circ \eta$ as the natural transformation where for each object a in $\mathcal{C}$,

$$(\varepsilon \circ \eta)_a = \varepsilon_a \circ \eta_a.$$

$$\begin{tikzcd}
a_1 \arrow[d, "f"'] & F(a_1) \arrow[r, "\eta_{a_1}"] \arrow[d, "F(f)"'] & G(a_1) \arrow[r, "\varepsilon_{a_1}"] \arrow[d, "G(f)"'] & H(a_1) \arrow[d, "H(f)"'] \\
a_2 & F(a_2) \arrow[r, "\eta_{a_2}"] & G(a_2) \arrow[r, "\varepsilon_{a_2}"] & H(a_2)
\end{tikzcd}$$

Figure 8: $\varepsilon \circ \eta$ is a natural transformation as this diagram commutes for every arrow $f : a_1 \to a_2$ in $\mathcal{C}$.

Of course, the analogy is not perfect. A functor maps things in $\mathcal{C}$ to things in $\mathcal{D}$. But there are no “things” in a functor F to map to another functor G. A natural transformation really sends objects in $\mathcal{C}$ to arrows in $\mathcal{D}$, in a manner that respects the structures of functors F and G.

Definition 2.69 was a slightly tortured definition. What was going on with η_c? It turns out to be the arrow associated with c is a natural transformation from F to G. Now we may rewrite it using the language of natural transformations to get a more natural definition.

Definition 2.73 (Adjoint Functors): Functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$ are **adjoint** if there is a natural transformation $\eta : \mathrm{id}_{\mathcal{C}} \Rightarrow G \circ F$ such that for each object d in $\mathcal{D}$ and arrow $f : c \to G(d)$, there exists a unique arrow $g : F(c) \to d$ with $G(g) \circ \eta_c = f$.

Note both functors in the natural transformation η are $\mathcal{C} \to \mathcal{C}$.

A trivial example of a natural transformation is $\mathrm{id}_F : F \Rightarrow F$, where each object c in $\mathcal{C}$ is assigned to the arrow $\mathrm{id}_{F(c)}$ in $\mathcal{D}$. In fact, it is the one that shows the forgetful functor on groups is right adjoint to the free functor taking $S \mapsto \mathrm{Free}(S)$.

Exercise 2.74: As you might expect, id_F is the **identity** when it comes to composition of natural transformations. Verify that given functors $F, G : \mathcal{C} \to \mathcal{D}$ and natural transformation $\eta : F \Rightarrow G$,

$$\eta \circ \mathrm{id}_F = \eta = \mathrm{id}_G \circ \eta.$$

> **Definition 2.75 (Hom-Set)**: In a category $\mathcal{C}$ with objects c_1 and c_2, the **hom-set** $\mathrm{Hom}_{\mathcal{C}}(c_1, c_2)$ is the class of arrows in $\mathcal{C}$ from c_1 to c_2.

The collection of arrows from c to d is not necessarily a set. However, to simplify matters we will assume this collection is a set from now on. In other words, whenever we say the phrase **hom-set**, we only consider categories where the collection of arrows between any two objects is a set.

We may also consider the act of taking a hom-set as a functor. More precisely, we fix c_2 and vary c_1.

> **Definition 2.76 (Hom-functor)**: Given a category $\mathcal{C}$ with object c_1, the **hom-functor** $\mathrm{Hom}_{\mathcal{C}}(c_1, -) : \mathcal{C} \to \mathrm{Set}$ is the functor that sends objects c_2 to $\mathrm{Hom}_{\mathcal{C}}(c_1, c_2)$ and arrows $f : x \to y$ to the function $g \mapsto g \circ f$.
>
> Similarly, the **hom-functor** $\mathrm{Hom}_{\mathcal{C}}(-, c_2)$ sends objects c_1 to $\mathrm{Hom}_{\mathcal{C}}(c_1, c_2)$ and arrows $h : x \to y$ to the function $h \mapsto h \circ g$.

Convince yourself that $g \mapsto g \circ f$ is indeed an arrow between $\mathrm{Hom}_{\mathcal{C}}(c_1, x)$ and $\mathrm{Hom}_{\mathcal{C}}(c_1, y)$. This is because every arrow $g \in \mathrm{Hom}_{\mathcal{C}}(c_1, x)$, i.e. every arrow $g : c_1 \to x$, is sent to $g \circ f : c_1 \to y$, which by definition is in $\mathrm{Hom}_{\mathcal{C}}(c_1, y)$.

Similarly convince yourself that $h \mapsto h \circ g$ is an arrow between $\mathrm{Hom}_{\mathcal{C}}(x, c_2)$ and $\mathrm{Hom}_{\mathcal{C}}(y, c_2)$.

Looking at the hom-functor another way, given an arrow $f : x \to y$, we want to use f to send an arbitrary arrow $g : c_1 \to x$ to another arrow of type $c_1 \to y$. There is only one formulaic way to do this: take $g \circ f$.

But what if we varied both c_1 and c_2 at once? What would the functor $\mathrm{Hom}_{\mathcal{C}}(-, -)$ look like? It is very obvious what such a hypothetical functor would do to an object (c_1, c_2). But what would it do to an arrow

$$(f : c_1 \to {c_1}', h : c_2 \to {c_2}')?$$

We want $\mathrm{Hom}_{\mathcal{C}}(-, -)$ to take (f, h) to a function that sends arrows $c_1 \to c_2$ to arrows ${c_1}' \to {c_2}'$. Now look at the following commutative diagram.

$$\begin{tikzcd}
c_1 \arrow[r, "g"] \arrow[d, "f"'] & c_2 \arrow[d, "h"'] \\
{c_1}' \arrow[r] & {c_2}'
\end{tikzcd}$$

Figure 9: Given arrows $f : c_1 \to {c_1}'$ and $h : c_2 \to {c_2}'$, for every arrow $g : c_1 \to c_2$, this diagram commutes.

It would be great if we could somehow reverse the direction of f to get $f^{\text{op}} : {c_1}' \to c_1$. Then the arrow $h \circ g \circ f^{\text{op}} : {c_1}' \to {c_2}'$ is how we formulaically construct an arrow of type ${c_1}' \to {c_2}'$ for **any** $g : c_1 \to c_2$. But we can do exactly this by considering f^{op} as an arrow in $\mathcal{C}^{\text{op}}$. Hopefully it will now make sense when the **hom-bifunctor** is defined to be of type $\mathcal{C}^{\text{op}} \times \mathcal{C} \to \text{Set}$.

Definition 2.77 (Hom-Bifunctor): The **hom-bifunctor** $\text{Hom}_{\mathcal{C}}(-,-) : \mathcal{C}^{\text{op}} \times \mathcal{C} \to \text{Set}$ maps arrows (f^{op}, h) to the function $g \mapsto h \circ g \circ f^{\text{op}}$.

$$\begin{tikzcd}
\text{Hom}_{\mathcal{C}}(c_1, c_2) \arrow[r, "\text{Hom}_{\mathcal{C}}(h, c_2)"] \arrow[d, "\text{Hom}_{\mathcal{C}}(c_1, f^{\text{op}})"'] & \text{Hom}_{\mathcal{C}}({c_1}', c_2) \arrow[d, "\text{Hom}_{\mathcal{C}}({c_1}', f^{\text{op}})"'] \\
\text{Hom}_{\mathcal{C}}(c_1, {c_2}') \arrow[r, "\text{Hom}_{\mathcal{C}}(h, {c_2}')"] & \text{Hom}_{\mathcal{C}}({c_1}', {c_2}')
\end{tikzcd}$$

Figure 10: Given objects c_1, c_2, ${c_1}'$, ${c_2}'$ in $\mathcal{C}$ and functions $f^{\text{op}} : {c_2}' \to c_2$, $h : {c_1}' \to c_1$, this diagram of hom-functors commutes. Both paths send $g : c_1 \to c_2$ to $f \circ g \circ h : {c_1}' \to {c_2}'$.

Now we are ready to provide a more complete characterization of adjoint functors.

Theorem 2.78 (Characterizations of Adjoint Functors): Functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$ are adjoint if they satisfy any of the following equivalent conditions:

1. Definition 2.73;
2. for any objects c in $\mathcal{C}$ and d in $\mathcal{D}$, there is an isomorphism
$$\varphi : \mathrm{Hom}_{\mathcal{D}}(F(c), d) \cong \mathrm{Hom}_{\mathcal{C}}(c, G(d))$$
that is natural in both c and d;
3. there exist natural transformations $\eta : F \circ G \Rightarrow \mathrm{id}_{\mathcal{C}}$ and $\varepsilon : \mathrm{id}_{\mathcal{D}} \Rightarrow G \circ F$ such that for each object c in $\mathcal{C}$ and d in $\mathcal{D}$,
$$\mathrm{id}_{F(c)} = \eta_{F(c)} \circ F(\varepsilon_c)$$
$$\mathrm{id}_{G(d)} = G(\eta_d) \circ \varepsilon_{G(d)}.$$

We omit the proof of Theorem 2.78 because it is somewhat technical. Furthermore, we only mention the final definition, also known as the **counit-unit** definition, for the sake of completeness.

Let's look at the second condition a little more closely. Where are the functors in our natural transformation? What does "natural in both c and d" mean? There is a lot of hidden complexity in that statement.

Here we consider $\mathrm{Hom}_{\mathcal{C}}(-, G(-))$ as a functor from $\mathcal{C}^{\mathrm{op}} \times \mathcal{D}$ to Set, and likewise we consider $\mathrm{Hom}_{\mathcal{D}}(F(-), -)$ as a functor with the same source and target categories.[24] These functors behave similarly to the hom-bifunctor.

$$\begin{tikzcd}
\mathrm{Hom}_{\mathcal{D}}(F(c), d) \arrow[r, "\varphi(f)"] \arrow[d, "\mathrm{Hom}_{\mathcal{D}}(F(f^{\mathrm{op}}), d)"'] & \mathrm{Hom}_{\mathcal{C}}(c, G(d)) \arrow[d, "\mathrm{Hom}_{\mathcal{C}}(f^{\mathrm{op}}, G(d))"'] \\
\mathrm{Hom}_{\mathcal{D}}(F(c'), d) \arrow[r, "\varphi(f)"] & \mathrm{Hom}_{\mathcal{C}}(c', G(d))
\end{tikzcd}$$

Figure 11: Naturality in c means that for any $f^{\mathrm{op}} : c' \to c$ and $h : F(c) \to d$, this diagram commutes.[25] Naturality in d is similar. (Compare this diagram with Figure 10.)

[24] A functor sends objects or arrows from the source category to the target category.

[25] For clarity about typechecking, through $\mathrm{Hom}_{\mathcal{D}}(-, d)$ the arrow $F(f^{\mathrm{op}}) : F(c') \to F(c)$ sends each arrow $g : F(c) \to d$ to the arrow $g \circ F(f^{\mathrm{op}})$ of type $F(c') \to d$.

> **Exercise 2.79**: Show that Definition 2.73 implies the hom-set characterization in Theorem 2.78 by taking $\varphi(g) = G(g) \circ \eta_c$ and showing this φ makes Figure 11 commute.

We have spent a lot of effort to characterize adjoint functors, all of which begs the question: what really is a pair of adjoint functors? I will shamelessly copy the example in https://mathoverflow.net/a/51659.

> Consider the category $\mathbb{Q}$ with a single arrow $x \to y$ when $x \leq y$, the subcategory $\mathbb{Z}$, and the inclusion functor $\mathrm{id} : \mathbb{Z} \to \mathbb{Q}$. Then consider the floor functor $\lfloor - \rfloor : \mathbb{Q} \to \mathbb{Z}$ which takes an arrow $p \to q$ in $\mathbb{R}$ to the arrow $\lfloor p \rfloor \to \lfloor q \rfloor$ in $\mathbb{Z}$, and similarly define the ceiling functor $\lceil - \rceil$.
>
> Then
>
> - $\lceil - \rceil$ and id are adjoint;
> - id and $\lfloor - \rfloor$ are adjoint.

Clearly $\lceil - \rceil$ and $\lfloor - \rfloor$ are approximations of rational numbers. So in some sense, adjoints can be considered as a "best available approximation". For instance, the forgetful functor from Group to Set (which you may recall is adjoint with the free functor) is similarly the best approximation of a group as a set.

Finally, we tie adjoint functors back to where we started from: universal properties. Note that in Definition 2.69, for each c in $\mathcal{C}$ we can find some η_c such that the universal property is satisfied. (Here we use "universal property" somewhat informally.) Being able to find such an η_c for each c, in other words, being able to satisfy the universal property for each c, is precisely what it means for two functors to be adjoint.

2.8.5 On the feasibility of using presentations

Group presentations are not a catch-all solution for describing every group. In an arbitrary presentation, the computational problem of determining whether two group elements are equivalent turns out to be generally undecidable.

So what use do group presentations have, besides just being a convenient way to describe certain groups? Why bother inventing all this complicated machinery?[26] I confess I do not know the answer, so I have asked Professor Cummings. Here are his insights:

> They are a basic idea in geometric group theory where you take a generating set and cook up an object called the Cayley Graph on which the group will then act in a natural (and very revealing) way. For example, the growth rate counts the rate at which the number of elements writable as a product of n generators goes up with n and is a very useful invariant of the group.

[26] Free groups and presentations are fairly complex, even if you completely ignore the category theoretic connections. I have left out a **lot** of the details.

Chapter 3

Rings

We know $(\mathbb{Z}/n\mathbb{Z}, +)$ forms a group. But many times we will want to study $(\mathbb{Z}/n\mathbb{Z}, +, \times)$, especially in the realm of cryptography. We can generalize the study of $\mathbb{Z}/n\mathbb{Z}$ to the study of **rings**. In particular, we will focus on **unital commutative rings**, which $\mathbb{Z}/n\mathbb{Z}$ is the perfect example of. Furthermore, given a field F, we may study the **ring of polynomials** $F[x]$ — the polynomials whose coefficients are in F — and this study will give us valuable information for the study of field theory.

> **Definition 3.1 (Ring)**: A **ring** is a triple $(R, +, \times)$ where $+$ and $\times$ are associative binary functions of the form $R \times R \to R$ that satisfy the following properties:
>
> - $(R, +)$ forms an abelian group.
> - Addition distributes over multiplication: $(a + b) \times c = (a \times c) + (b \times c)$.
> - Multiplication distributes over addition: $a \times (b + c) = (a \times b) + (a \times c)$.

Notes about notation: we will usually drop the × operator and implicitly perform × before +, just as in regular arithmetic. Also, we will denote the identity of the group $(R, +)$ as 0 instead of id from now on. We usually refer to the ring just as R when it is clear what the additive and multiplicative operator are.

We say R is **commutative** if × is commutative, and we say R is **unital** if there is some element $1 \in R$ such that $1 \times a = a \times 1 = a$ for all $a \in R$ — in other words, if there is a multiplicative identity.

Furthermore, we say that an element u of a unital ring R is a **unit** if and only if it is invertible. In other words, u is a unit precisely when there exists some $v \in R$ such that $uv = 1$.

Exercise 3.2: If ring R has an identity element 1, show that it is unique.

Exercise 3.3: If $uv = 1$, show that $vu = 1$ as well.

Exercise 3.4: Show that the units of a ring R form a group under ring multiplication.

Here are some examples of rings:

- If you know anything about fields, note that a field is a unital commutative ring where every element has a multiplicative inverse.
- $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. In particular note that $\mathbb{Z}$ is not a field.
- $\mathbb{Z}/n\mathbb{Z}$.
- $2\mathbb{Z}$, the set of even integers, which is commutative but not unital.

- GL_n, the set of $n \times n$ matrices with non-zero determinant. This is a unital ring — in fact, every matrix has a multiplicative inverse — but it is not commutative.

We may also define a subring the same way we defined a subgroup.

> **Definition 3.5 (Subring)**: A **subring** S of R is any subset of R closed under $+$ and $\times$. We write $S \leq R$.

Similarly, S (with $+$ and $\times$) can be considered a ring in its own right. We may also define S-cosets under the group $(R, +)$, and each coset can be denoted as $r + S$ for some $r \in R$.

Subgroups preserve their identity: the identity element in $H \leq G$ is the same as the identity element in G. So the additive identity of subrings is obviously preserved as well. But the multiplicative identity does not behave so nicely.

Example 3.6 (Subrings do not preserve multiplicative identity): The multiplicative identity of $\mathbb{Z}^2$ is $(1, 1)$. Yet the multiplicative identity of the subring $\{(z, 0) \mid z \in \mathbb{Z}\}$ is $(1, 0)$.

It is an important fact in number theory that if $a \times b = 0$ for integers a and b, then one of a and b must be 0. Furthermore $\mathbb{Z}$ is a unital commutative ring. This motivates the following definition.

> **Definition 3.7 (Integral Domain)**: A unital commutative ring R is an **integral domain** if $ab = 0$ implies one of a, b are 0. In other words, if $a, b \neq 0$, then $ab \neq 0$.

Note finite integral domains are fields as $x \mapsto a \cdot x$ is an injective function, so there must be some a^{-1} that is sent to 1 under this function, proving the existence of an inverse for a.

3.1 Homomorphisms and Isomorphisms

In every ring there is an underlying group. So the concepts of a homomorphism and quotient ring can be naturally defined, and furthermore, natural equivalents of the Isomorphism Theorems hold.

> **Definition 3.8 (Homomorphism)**: A **homomorphism** $\varphi : R \to S$ (where R and S are rings) is a function where
>
> $$\varphi(r_1 + r_2) = \varphi(r_1) + \varphi(r_2) \text{ and } \varphi(r_1 r_2) = \varphi(r_1)\varphi(r_2)$$
>
> for all $r_1, r_2 \in R$.

A ring isomorphism is exactly what you think it is. And ring isomorphism is also an equivalence relation.

There is a small caveat. **When $\boldsymbol{R}$ and $\boldsymbol{S}$ are unital, other texts usually require that $\boldsymbol{\varphi(1_R) = 1_S}$.** However, we do not require this because it frankly makes no difference to either of our lives. A few discussions are simplified by dropping this requirement.

Exercise 3.9: And importantly, if R and S are unital and φ is an isomorphism, then we must have $\varphi(1_R) = 1_S$. Prove this.

Exercise 3.10: If $\varphi : R \to S$ is a homomorphism, prove that $\ker \varphi \leq R$ and $\operatorname{im} \varphi \leq S$. (In particular, I am asking you to show that $\ker \varphi$ and $\operatorname{im} \varphi$ are subrings.) Furthermore, note that for every $a \in \ker \varphi$ and $r \in R$, $ar, ra \in \ker \varphi$, i.e. $\ker \varphi$ is closed under multiplication by R.

In group theory, we've studied groups of the form $\frac{G}{\ker \varphi}$. We are going to do the same here: study rings of the form $\frac{R}{\ker \varphi}$. Now here is the kicker: we have representation invariance if we define $\ker \varphi$-coset addition and multiplication as

- $(r_1 + \ker \varphi) + (r_2 + \ker \varphi) = (r_1 + r_2) + \ker \varphi$
- $(r_1 + \ker \varphi)(r_2 + \ker \varphi) = (r_1 r_2) + \ker \varphi$,

respectively. Addition is representation invariant as $\ker \varphi$ is a normal subgroup of $(R, +)$. In fact, addition is representation invariant under any subring as any subring is a normal subgroup of the abelian group $(R, +)$.

The real sticking point is multiplication. Multiplication is representation invariant as for all $a, b \in \ker \varphi$,

$$(r_1 + a)(r_2 + b) + \ker \varphi = r_1 r_2 + a r_1 + b r_2 + ab + \ker \varphi = r_1 r_2 \ker \varphi,$$

as ar_1, br_2, and ab are all in $\ker \varphi$. (Recall for the first two that $\ker \varphi$ is closed under multiplication by R.)

So now we can in good faith consider the **quotient ring** $\frac{R}{\ker \varphi}$. The question now is, for which subrings I can we consider this quotient ring? Precisely the ones with representation invariance. It turns out that it is exactly when I is closed under multiplication by R.

Definition 3.11 (Ideal): Given a subring I of R, we say I is an **ideal** if any of the following equivalent properties are satisfied:

1. $I = \ker \varphi$ for some homomorphism $R \to S$. (We don't really care what S is.)
2. I is representation invariant under multiplication.
3. I is closed under multiplication by R, that is, for each $a \in I$ and $r \in R$, we have $ar, ra \in I$ as well.

We have conversationally shown that $(1) \Longrightarrow (3) \Longrightarrow (2)$ and $(1) \Longrightarrow (3)$. To see that $(2) \Longrightarrow (3)$, note that representation invariance implies

$$0r + I = ar + I$$

for all $r \in R$ and $a \in I$, which means $ar \in I$ a desired.[1]

And to see that $(3) \Longrightarrow (1)$, simply take the homomorphism $\varphi : R \to \frac{R}{I}$ where $\varphi : r \mapsto r + I$. We know $\frac{R}{I}$ is well-defined as $(3) \Longrightarrow (2)$. So these conditions really are equivalent.

The isomorphism theorems are precisely what you think they are:

1. If $\varphi : R \to S$ is a homomorphism, then $\frac{R}{\ker \varphi} \cong \operatorname{im} \varphi$.
2. If $I, S \leq R$ and I is an ideal of R, then $\frac{I+S}{I} \cong \frac{S}{I \cap S}$.
3. If I and J are ideals of R with $I \leq J$, then $\frac{J}{I}$ is an ideal of $\frac{R}{I}$ and
$$\frac{\frac{R}{I}}{\frac{J}{I}} \cong \frac{R}{J}.$$
4. If I is an ideal of R, then there is an inclusion-preserving bijection between the A where $I \leq A \leq R$ and the subrings of $\frac{R}{I}$. Furthermore, if A and $\frac{B}{I}$ are associated with each other, then
$$A \text{ is an ideal of } R \Longleftrightarrow \frac{B}{I} \text{ is an ideal of } \frac{R}{I}.$$

The proofs are nearly identical to those of the Isomorphism Theorems with groups, only with a little bit more work to take care of multiplication.

3.1.1 Special Ideals

Just as we may generate the smallest subgroup containing a set by repeatedly applying the group operation, we may generate the smallest ideal containing a set by repeatedly applying $+$ (to elements in the ideal) and $\times$ (to an element in the ideal and an element of R). For any $X \subseteq R$, we denote the smallest ideal containing X as (X).

There is a particular type of ideal we care about: those generated by a single element, i.e. those of the form (a) for some $a \in R$. We call these ideals **principal**.

Exercise 3.12: Consider principal ideals (a) and (b). Show $(a) \leq (b)$ if and only if there exists some $r \in R$ such that $a = rb$.

Now we begin focusing mostly on unital commutative rings. An ideal I is **maximal** if $I \neq R$ and there is no ideal J between I and R, i.e. $I < J < R$. Maximal ideals are important because (among other reasons), every field is isomorphic to some quotient ring $\frac{R}{I}$ where R is commutative and I is maximal.

Theorem 3.13: Suppose R is unital and commutative and I is an ideal of R. Then I is maximal if and only if $\frac{R}{I}$ is a field.

[1] To be explicit, we are using the alternative representation of 0 as a in our proof.

Proof of Theorem 3.13: Suppose I is maximal. Then $\frac{R}{I}$ has no interesting ideals by Fourth Isomorphism, meaning for every non-zero $a \in \frac{R}{I}$, $(a) = \frac{R}{I}$. Since $1 \in \frac{R}{I}$, a must be a unit.

Suppose $\frac{R}{I}$ is a field. It has no interesting ideals because for all ideals J of $\frac{R}{I}$, if $a \neq 0 \in J$, then $a^{-1}a = 1 \in J$, implying $J = \frac{R}{I}$. □

So it would perfectly fair to characterize fields as "quotients of maximal ideals" for they are one and the same.

It is also worth noting that not every ring has a maximal ideal. Consider $(\mathbb{Q}, +, q_1 q_2 \mapsto 0)$, i.e. the ring on $\mathbb{Q}$ whose multiplicative operation sends everything to 0. (This is a valid ring!) Evidently all of its ideals are just the subgroups of $\mathbb{Q}$. Now we recall Exercise 2.6 and conclude this ring has no maximal ideals.

However, every unital ring has a maximal ideal. And in fact, every proper subgroup of a unital ring is contained in some maximal ideal.

Finally we discuss **prime ideals**. Given a unital commutative ring R, we say $I \neq R$ is a prime ideal if for all $a, b \in R$,

$$ab \in I \Longrightarrow \text{one of } a, b \in I.$$

This definition is natural because

$$p \text{ prime} \Longleftrightarrow p\mathbb{Z} \text{ is a prime ideal of } \mathbb{Z}.$$

(Prove this yourself! It is very easy.)

Furthermore, for any unital commutative ring R,

$$I \text{ prime} \Longleftrightarrow \frac{R}{I} \text{ is an integral domain.}$$

Finally we may note that in a unital commutative ring **every maximal ideal I is prime**. For $\frac{R}{I}$ is a field, which necessarily is an integral domain, implying I is prime.

3.1.2 Product of Ideals

There is a notion of the product of two ideals. The naive way to define the product of ideals I and J is $\{ab : a \in I, b \in J\}$. However, to ensure closure under addition, we define the product IJ as the ideal **generated** by the set $\{ab : a \in I, b \in J\}$. It turns out that this ideal is equivalent to the set generated by all finite sums of the form

$$\left\{\sum_{k=1}^{n} a_k b_k : a_k \in I, b_k \in J\right\}.$$

(This is worth verifying yourself.)

Furthermore, we may define the product of many ideals $I_1, ..., I_n$ recursively: it is equal to the product of $I_1, ..., I_{n-1}$ and I_n. Luckily, this definition of ideal multiplication ends up being associative (a fact we will not prove).

3.2 The Chinese Remainder Theorem

In this section we work with unital commutative rings.

Consider the ring $R = \mathbb{Z}/30\mathbb{Z}$. Note that $I = 2\mathbb{Z}/30\mathbb{Z}$ and $J = 3\mathbb{Z}/30\mathbb{Z}$ are both ideals of R not equal to R. Furthermore, neither I nor J are maximal. But because 2 and 3 are coprime, every element in R can be expressed in the form $a + b : a \in I, b \in J$. In other words,
$R = I + J$.

This idea of ideals that sum to R can be generalized.

> **Definition 3.14 (Coprime Ideals)**: Ideals I and J of R are coprime if and only if $I + J = R$.

This generalizes the idea that 2 and 3 are coprime in $\mathbb{Z}$.

> **Theorem 3.15 (Chinese Remainder Theorem: Two Ideals)**: If I and J are coprime ideals in R, then
> - $I \cap J = IJ$,
> - $\frac{R}{I \cap J} \cong \frac{R}{I} \times \frac{R}{J}$.

Proof of Theorem 3.15: Note by definition that $1 = a + b$ for some $a \in I$ and $b \in J$.

- For any arbitrary ring and ideals, $IJ \subseteq I \cap J$ as ideals are closed under multiplication. If we multiply $a \in I$ with $b \in J$, then the result ab is in both I and J. Now if $r \in I \cap J$, note

$$r = r(a + b) = ra + rb \in IJ$$

  as $ra \in IJ$ (because $a \in I$ and $r \in J$) and likewise $rb \in IJ$.
- Consider the ring homomorphism $\varphi : R \longrightarrow \frac{R}{I} \times \frac{R}{J}$ where $\varphi : r \mapsto (r + I, r + J)$. Note $\ker \varphi = I \cap J$ and $\operatorname{im} \varphi = \frac{R}{I} \times \frac{R}{J}$ as

$$\varphi(ra + sb) = (rb + I, sa + J) = (ra + rb + I, sa + sb + J) = (r + I, s + J).$$

  Applying the First Isomorphism Theorem finishes.

□

It may be worth reasoning through the Chinese Remainder Theorem with the example of $\mathbb{Z}/30\mathbb{Z}$, $2\mathbb{Z}/30\mathbb{Z}$, and $3\mathbb{Z}/30\mathbb{Z}$ we started with.

Theorem 3.16 (Chinese Remainder Theorem): If $I_1, ..., I_n$ are coprime ideals in R, then

- $I_1 \cap ... \cap I_n = I_1 ... I_n$,
- $\frac{R}{I_1 \cap ... \cap I_n} \cong \frac{R}{I_1} \times ... \times \frac{R}{I_n}$.

Proof of Theorem 3.16: We've already put in the legwork to prove this theorem.

- Trivial by induction and the first part of Theorem 3.15.
- Identical to the second part of Theorem 3.15.

□

In particular, this implies the familiar Chinese Remainder Theorem in $\mathbb{Z}$: if $a_1, ..., a_n$ are pairwise coprime positive integers, then determining some integer a modulo $a_1, ..., a_n$ uniquely determines a modulo $a_1 ... a_n$, and vice versa.

3.3 Domains

Every ring we work with here is an integral domain. That is, they are unital commutative rings, and if $a, b \neq 0$, then $ab \neq 0$.

There are certain properties of rings that do not hold for all rings but do hold for many rings of interest. As an example, in the ring of integers $\mathbb{Z}$ and the ring of complex polynomials $\mathbb{C}[x]$, we may apply the Euclidean Algorithm. Furthermore we can uniquely factorize integers into primes and complex polynomials into irreducible (i.e. linear) factors. Instead of concretely studying these rings, however, we will be studying these properties in the abstract.

The three classes of rings of interest are **Euclidean Domains**, **Principal Ideal Domains**, and **Unique Factorization Domains**. We will define what they are shortly, but one of the main takeaways of this section is that

$$\text{ED} \Longrightarrow \text{PID} \Longrightarrow \text{UFD},$$

or equivalently,

$$\text{ED} \subset \text{PID} \subset \text{UFD}.$$

That is, given a ring that is a Euclidean Domain, we will see it must be a Principal Ideal Domain and thus a Unique Factorization Domain.

3.3.1 Euclidean Domains

To define a Euclidean Domain we first must define a notion of a norm on a ring. This notion is quite loose.

> **Definition 3.17 (Norm)**: A **norm** $|-|$ on a ring R is a function $|-| : R \to \mathbb{N}$ where $|0| = 0$. If $|r| = 0 \Longrightarrow r = 0$ then $|-|$ is a positive norm.

Note $\mathbb{Z}$ and $F[x]$ (i.e. the ring of polynomials under any field F) are both Euclidean Domains. For $\mathbb{Z}$ the norm is simply the absolute value function, and for $F[x]$ the norm is the degree of the polynomial.[2]

> **Definition 3.18 (Euclidean Domain)**: A ring R is a Euclidean Domain if and only if there exists a norm $|-|$ where for all $a, b \in R$, there exists some $q, r \in R$ such that
> - $a = qb + r$,
> - $r = 0$ or $|r| < |b|$.

Unlike the Euclidean Algorithm on polynomials, there is no guarantee that q and r are **unique** for a general Euclidean Domain. As a stupid example, note that $10 = 3 \cdot 3 + 1$ and $10 = 4 \cdot 3 - 2$. Since $|1| < |3|$ and $|-2| < |3|$, we have two distinct quotient-remainder pairs.

3.3.2 Greatest Common Divisors

The whole point of the Euclidean Algorithm on integers is that it generates their greatest common divisor. The Euclidean Algorithm also turns out to work for arbitrary Euclidean Domains, but first we must define what divisibility and greatest common divisors are.

> **Definition 3.19 (Divisibility)**: For $a, b \in R$, we say $a \mid b$ if and only if there exists some $r \in R$ such that $b = ar$.

Note that $a \mid b \Longleftrightarrow b \in (a) \Longleftrightarrow (b) \leq (a)$.[3] So we may translate the language of divisibility into the language of ideals (and vice versa) quite easily.

> **Definition 3.20 (Greatest Common Divisor)**: A greatest common divisor of $a, b \in R$ is some d such that
> - $d \mid a$ and $d \mid b$,
> - $d' \mid a$ and $d' \mid b$ implies $d' \mid d$.

[2] This is not precisely true. The degree of a constant **non-zero** polynomial is 0, but the degree of the 0 polynomial is actually $-\infty$. But for the purposes of showing it is a Euclidean Domain, we set its norm to be 0 instead.

[3] Recall (a) denotes the principal ideal generated solely by the element a. Similarly for (b).

Let me be very clear here. **Not every pair of elements in any ring has a greatest common divisor.** Obviously the pair $(0, 0)$ does not have a greatest common divisor. And in the ring of even integers, 6 does not even have any divisors, so the pair $(6, 2n)$ never has a greatest common divisor. And there are integral domains with a non-trivial pair of elements without a greatest common divisor, an example of which you may find at Wikipedia: https://en.wikipedia.org/wiki/Greatest_common_divisor#In_commutative_rings.

But every pair of elements $(a, b) \neq (0, 0)$ in a Euclidean Domain, Principal Ideal Domain, or Unique Factorization Domain has a greatest common divisor. We will prove this for Euclidean Domains, and I encourage you to prove them yourself for Principal Ideal Domains and Unique Factorization Domains once you have learned what they are. It is not too hard.

When they do exist, greatest common divisors are almost unique in a sense. By reframing divisibility in terms of ideals, we can see that if d and d' are both greatest common divisors of a and b, then $(d) = (d')$. This implies that $d' = ud$ where u is a unit. So greatest common divisors are unique up to multiplication by units.

Conversely, if d is a greatest common divisor of a and b and u is a unit, then ud is also a greatest common divisor of a and b. This establishes a bidirectional relationship.

For the sake of completeness we will now describe the Euclidean Algorithm on a Euclidean Domain R. It is an algorithm that takes in two ring elements and returns a ring element.

- We start with two ring elements a and b.
- If $a \mid b$, then return a. Similarly, if $b \mid a$, return b. (If both $a \mid b$ and $b \mid a$, then pick whichever of a or b you want.)
- Pick whichever of a and b has the smallest norm; without loss of generality say that $|a| \geq |b|$. Then find some r such that $a = qb + r$ and $|r| < |b|$ and replace a with r. (If it is the case that $|a| < |b|$, then swap the roles of a and b.)

> **Theorem 3.21**: If a and b are elements of a Euclidean Domain R where at least one of a and b are non-zero, and the Euclidean Algorithm on a and b returns d, then
> - d is a greatest common divisor of a and b,
> - The ideal generated by d is equivalent to the ideal generated by a and b. In other words, $(d) = (a, b)$.

Of course, Theorem 3.21 implies that every pair of elements except $(0, 0)$ in a Euclidean Domain has a greatest common divisor: we have just described a process to construct one, after all!

Proof of Theorem 3.21: Completely trivial by induction on the number of steps the Euclidean Algorithm takes.

Suppose $|a| \geq |b|$ and $a = qb + r$. By the inductive hypothesis d divides r and b, so it also must divide a. Similarly, if some d' were to divide a and b, then it must divide r and b. But then it must divide d, meaning d is a **greatest** common divisor.

Further note that $(d) = (r, b)$ by the inductive hypothesis. Obviously $(r, b) = (qb + r, b)$. □

3.3.3 Principal Ideal Domains

Recall the definition of a principal ideal: an ideal that may be generated by exactly one element. In other words, a principal ideal (a) is of the form

$$\{ar : r \in R\}$$

For reasons we will soon see, it is very convenient if every ideal of a ring is principal. So this motivates defining **Principal Ideal Domains** as those such rings.

Theorem 3.22: Every **Euclidean Domain** is a **Principal Ideal Domain**.

Proof of Theorem 3.22: Suppose I is a non-zero[4] ideal of R. Then take some $b \in I$ where $|b|$ is minimal (there may be multiple such b, take any). Then for any $a \in I$, we may write

$$a = qb + r$$

where $r = 0$ or $|r| < |b|$. But since $|b|$ is minimal we cannot have $|r| < |b|$ and must have $r = 0$. Thus $b \mid a$, implying that for all $a \in I$, $a \in (b)$. □

Technically we have omitted the step where $(b) \in I$, but that is obvious.

Theorem 3.23: In a Principal Ideal Domain, every non-zero prime ideal is maximal.

Proof of Theorem 3.23: Suppose (p) is a prime ideal of a ring R and $(p) \leq (a)$; we want to show that $(p) = (a)$ or $(p) = R$. Note $p \in (a)$, meaning there exists some b such that $ab = p$. If $a \in (p)$ we are done, otherwise $b \in (p)$ implies there is some c such that $b = cp$. But then

$$ab = a(cp) = p \Longrightarrow ac = 1$$

and thus $(a) = R$ as every $r \in R$ can be expressed as $a(cr)$. □

Theorem 3.24: A ring R is a field if and only if the polynomial ring $R[x]$ is a Principal Ideal Domain.

Proof of Theorem 3.24: If R is a field then $R[x]$ is obviously a Euclidean Domain: consider $|f| = \deg f$ as the norm.

[4] The zero ideal is trivially principal, so we do not need to consider it.

If $R[x]$ is a Principal Ideal Domain, then note (x) is a prime ideal of $R[x]$, meaning it is maximal. Further note that $\frac{R[x]}{(x)} \cong R$. By Theorem 3.13, R is a field. □

As a corollary, R is a field if and only if $R[x]$ is a Euclidean Domain.

3.3.4 Unique Factorization Domains

Finally we study rings where every element can be uniquely factored (up to units). Prototypical examples include $\mathbb{Z}$ and $F[x]$ (where F is an arbitrary field). But first we ought to define a notion of prime and irreducible.

> **Definition 3.25 (Prime Ring Elements)**: An element p of ring R is **prime** if and only if (p) is a prime ideal.

As an example, notice that the prime ideals in $\mathbb{Z}$ are exactly those that can be represented as (p) (where p is a prime number). That is why it makes sense to consider the prime numbers as **prime ring elements** in $\mathbb{Z}$. (Further notice that negative primes, such as -2 and -3, are also prime ring elements.)

> **Definition 3.26 (Irreducible Elements)**: An element r of ring R is **irreducible** if and only if
> - r is not a unit,
> - r is not the product of any two non-units. In other words, if $r = ab$ for $a, b \in R$, then one of a and b must be a unit.

Irreducibles and primes are very closely related. In an integral domain, primes are always irreducible: if $ab = p$ then one of a or b is in (p). Say a is in (p). Then $a = pr$ and $prb = p$, implying $rb = 1$, i.e. b is a unit.

Similarly, in a Principal Ideal Domain irreducibles are always prime. Say p is irreducible; then any ideal containing (p) must be of the form (m), meaning $p = rm$ for some $r \in R$. The irreducibility of p implies either r or m is a unit; if r is a unit then $(p) = (m)$, and if m is a unit then $(m) = R$. Thus (p) is maximal and hence prime (as maximal ideals are prime in Principal Ideal Domains).

Polynomial rings over arbitrary fields provide the prototypical example of irreducible elements, and indeed, they are the primary use for the theory of irreducible elements. Consider $\mathbb{Z}[x]$. The only unit is 1, and the **irreducible ring elements** are precisely the irreducible polynomials in $\mathbb{Z}[x]$.

Here is the important part: **every polynomial in $\mathbb{Z}[x]$ can be uniquely factored into irreducible polynomials**, full stop. We can make a very similar claim about $\mathbb{R}[x]$, as we will soon see.

Definition 3.27 (Unique Factorization Domain): A ring R is a Unique Factorization Domain if and only if for every $r \in R$,

- r can be factored into a finite number of irreducible elements $r = p_1 p_2 ... p_n$,
- if r can be factored into irreducible elements $r = p_1 p_2 ... p_n$ and $r = q_1 q_2 ... p_m$, then there is a bijection $\pi : \{1, ..., n\} \longrightarrow \{1, ..., m\}$ such that p_i and $q_{\pi(i)}$ are identical up to multiplication by a unit. In other words, $p_i = q_{\pi(i)} u$, where u is a unit.

Informally, the second condition means all factorizations of a ring element into irreducibles are unique, up to multiplication by a unit for each irreducible factors.

Exercise 3.28: Prove that non-zero elements in a Unique Factorization Domain are prime if and only if they are irreducible.

Now we look at $\mathbb{R}[x]$. The units are scalars (i.e. members of $\mathbb{R}$, and up to scaling by constant factors (e.g. $(2x+2)(x-1) = (x+1)(2x-2)$), the factorization of a polynomial into irreducible polynomials is unique. So $\mathbb{R}[x]$ is a Unique Factorization Domain.

You may also notice that $\mathbb{R}$ is trivially a Unique Factorization Domain (as $\mathbb{R}$ is a field). In fact, it is the case for any ring R that

$$R \text{ UFD} \Longleftrightarrow R[x] \text{ UFD}.$$

Obviously $R[x]$ UFD $\Longrightarrow R$ UFD as every ring element in R needs to be factorizable in $R[x]$, and each of these factorizations must solely be in scalars. The other direction is much harder and requires developing a good bit of theory.

Definition 3.29 (Ring of Fractions): Given an integral domain R, we may define a **ring of fractions** as ordered pairs (a, b) where $b \neq 0$, which suggestively are represented as $\frac{a}{b}$. We say $\frac{a}{b} = \frac{c}{d}$ precisely when $ad = bc$, which is exactly what you'd expect when cross-multiplying.

Addition and multiplication are defined exactly the way you think they are: $\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$ and $\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}$.

Exercise 3.30: Verify that $+$ and $\times$ as defined above are representation invariant.

Note the ring of fractions is actually a **field of fractions**, for the inverse of $\frac{a}{b}$ is just $\frac{b}{a}$.

Further note R being an integral domain implies $R[x]$ is an integral domain too. This is because for the product of two polynomials $f, g \in R[x]$ to be 0, one of the constant terms of f or g must be 0. This easily lends itself to a proof by induction on $\deg f + \deg g$.

So the ring of fractions in $R[x]$ is also a field. This allows us to prove **Gauss' Lemma**.

Theorem 3.31 (Gauss' Lemma): Suppose R is a Unique Factorization Domain with field of fractions F. If a polynomial $f(x) \in R[x]$ is reducible in $F[x]$, then it is reducible in $R[x]$.

The idea, very roughly, is to clear the denominators of the factorization in $F[x]$.

Proof of Theorem 3.31: Say $f(x) = A(x)B(x)$ in $F[x]$. Each coefficient of A and B is in F, meaning that informally, we may "cross-multiply the denominators" to get $rf(x) = A'(x)B'(x)$ in $R[x]$ for some $r \in R$.

If r is a unit, then $f(x) = r^{-1}A'(x)B'(x)$ is a valid factorization in $R[x]$. Now suppose not. Because R is a Unique Factorization Domain, we can write $r = p_1...p_n$ for irreducibles $p_1, ..., p_n$. By Exercise 3.28, (p_1) is a prime ideal.

Taking the equation modulo p_1, we see that $0 = A'(x)B'(x)$ in $\frac{R}{(p_1)}[x]$. Now note $\frac{R}{(p_1)}[x]$ is an integral domain because $\frac{R}{(p_1)}$ is one due to (p_1) being prime, and

$$\frac{R}{(p_1)} \text{ integral domain} \Longrightarrow \frac{R}{(p_1)}[x] \text{ integral domain.}$$

This means that one of $A'(x)$ or $B'(x)$ is equivalent to 0 modulo p_1; suppose without loss of generality it is $A'(x)$. But then $A'(x) = pC(x)$, meaning that we may write $p_2...p_n = C(x)B'(x)$.

Now we may easily finish by inducting on the number of irreducibles that r factors into. □

Now we will finally prove for good that R UFD $\Longrightarrow$ $R[x]$ UFD.

Theorem 3.32: If R is a Unique Factorization Domain, then so is $R[x]$.

Remind yourself of Definition 3.20 and associated facts. It will be important for proving the uniqueness of factorizations in $R[x]$.

Proof of Theorem 3.32: Let $f(x)$ be a polynomial in $R[x]$. If there is some $r \in R$ that divides $f(x)$, we may factor it out without worry, so we will now suppose that no such r exists. Note $F[x]$ is a Unique Factorization Domain, so for every $f(x) \in R[x]$, there is a unique factorization $f(x) = p_1(x)...p_n(x)$, where the factors are irreducibles in $F[x]$.

Through Theorem 3.31 we may induce a factorization $f(x) = q_1(x)q_2(x)...q_n(x)$ in $R[x]$ where each $q_i(x)$ is a multiple of $p_i(x)$ in R. And since no scalar $r \in R$ divides $f(x)$, there certainly is no scalar dividing any $q_i(x)$.

So each $q_i(x)$ is irreducible, meaning we have an irreducible factorization of $f(x)$ in $R[x]$. We now need to show that it is unique. Suppose we have

$$q_1(x)...q_n(x) = f(x) = s_1(x)...s_n(x).$$

Because $F[x]$ is a Unique Factorization Domain, for each i there is some $\frac{a_i}{b_i} \in R$ such that $q_i(x) = \frac{a_i}{b_i} s_i(x)$.[5] Then $b_i q_i(x) = a_i s_i(x)$.

Because no $r \in R$ divides $q_i(x)$, 1 is a greatest common divisor of the coefficients of $q_i(x)$. Likewise for $s_i(x)$. Thus b_i is a greatest common divisor of the coefficients of $b_i q_i(x)$ and a_i is a greatest common divisor of the divisors of $a_i s_i(x)$.

Because greatest common divisors are unique up to multiplication by units, there is some unit $u_i \in R$ such that $a_i = b_i u_i$. So $q_i(x) = u_i s_i(x)$, and as u_i is obviously a unit in $R[x]$ as well, the factorizations of $f(x)$ are equivalent up to units. □

An easy corollary is that if R is a Unique Factorization Domain, then a polynomial ring in R with an arbitrary number of variables (even infinite-variable) is also a Unique Factorization Domain.[6]

A result of this fact is Eisenstein's Criterion.

> **Theorem 3.33 (Eistenstein's Criterion)**: If p is a prime in $\mathbb{Z}$ and in the polynomial
>
> $$f(x) = x^n + a_{n-1}x^{n-1} + ... + a_0,$$
>
> p divides each of $a_0, ..., a_{n-1}$ but p^2 does not divide a_0, then $f(x)$ is irreducible in $\mathbb{Q}[x]$ (implying it is irreducible in $\mathbb{Z}[x]$.

This is a special form of a more general result for prime ideals over arbitrary integral domains. For a polynomial $f(x) = a_n x^n + ... + a_0$, if there exists a prime ideal $P \subset R$ where

- $a_0, ..., a_{n-1}$ are in P,
- a_n is not in P,
- and a_0 is not in P^2,

then $f(x)$ is irreducible.

Proof of Theorem 3.33: We prove the general version. If $f(x)$ was reducible into $f(x) = a(x)b(x)$, then $x^n \equiv a(x)b(x) \pmod{P}$. Since P is prime, $\frac{R}{P}$ is an integral domain, meaning at least one of a_0 and b_0 (the constant terms of $a(x)$ and $b(x)$) must be 0 modulo P. Very informally, this means that

- we may keep noticing that one of $a(x)$ and $b(x)$ is divisible by x in $\frac{R}{P}$,
- factor out the x,
- and then repeat the process,

[5] For convenience we label $q_1, ...q_n$ and $s_1, ...s_n$ in a way that allows them to be paired like this.

[6] The polynomial ring $R[x_1, x_2]$ is defined as $R[x_1][x_2]$. Polynomial rings with more variables are defined similarly. And an infinite-variable polynomial ring $R[x_1, x_2, ...]$ is defined as the union of $R[x_1]$, $R[x_1, x_2]$, and so on.

until we conclude that $a(x)$ and $b(x)$ are of the form x^k and x^j in $\frac{R}{P}$.

But this means that $a_0 \in P$ and $b_0 \in P$, which implies $a_0 b_0 \in P^2$, contradiction. □

We finish by relating Principal Ideal Domains to Unique Factorization Domains.

Theorem 3.34: Every Principal Ideal Domain is a Unique Factorization Domain.

To prove that Principal Ideal Domains have finite factorizations into irreducibles we establish two preliminary results.

Theorem 3.35: Consider a chain of ideals

$$(a_0) \leq (a_1) \leq \ldots$$

in a Principal Ideal Domain.

Then the chain must stabilize at some point. In other words, there is some n such that for all $k \geq n$, $(a_k) = (a_n)$.

Proof of Theorem 3.35: Note $I = (a_0) \cup (a_1) \cup \ldots$ is an ideal and it can be represented as (a) for some a in the Principal Ideal Domain. Since I contains a, there must be some (a_n) that contains a. But then $(a_n) = (a)$, which implies the result. □

Theorem 3.36 (Kőnig's Lemma): Consider a tree T with an infinite number of vertices, each with finite degree. Root it at any vertex v_0. Then T has a branch of infinite length.

An alternative formation of Theorem 3.36 is that every connected graph G with infinite vertices, each with finite degree, has an infinitely long simple path.[7] It is a trivial corollary of this statement as we may remove a subset of the edges to make the remaining graph a tree. The infinite branch in that tree is still a valid simple path when we add the removed edges back to the graph.

Proof of Theorem 3.36: We start with v_0. One of its children must have infinite descendants, otherwise T would be finite as v_0 has finitely many children. Say the child with infinite descendants is v_1. Then v_1 is the root of an infinite subtree satisfying the conditions of the lemma.

Now we repeat the process on v_1. This generates a path of infinite length $v_0 v_1 \ldots$ as desired. □

[7] A simple path is one that does not ever visit the same vertex twice.

Each node of a binary tree has a finite number of neighbors (at most 3). So by Theorem 3.36, a binary tree where every branch has finite length must be finite.

Now we prove the main result.

Proof of Theorem 3.34: To produce a finite irreducible factorization, we recursively apply the following algorithm:

- If r is irreducible, we are done.
- Otherwise, we may factor $r = ab$, where neither a nor b is a unit. Then apply the algorithm onto a and b to produce irreducible factorizations $p_1 ... p_n$ and $q_1 ... q_m$; return $p_1 ... p_n q_1 ... q_m$ as the irreducible factorization of r.

This algorithm generates a binary tree: each reducible r has children a and b. This binary tree has no infinite-length branches by Theorem 3.35,[8] thus it is finite by Theorem 3.36. So the algorithm terminates in a finite number of steps and thus returns a finite irreducible factoring of r.

To show that each factorization is unique up to units, consider two factorizations

$$r = p_1 ... p_n = q_1 ... q_m \text{ where } n \geq m.$$

We show they are essentially equivalent[9] by inducting on n. Note that p_1 divides some $q_1, ..., q_m$; without loss of generality say $p_1 = q_1 u$. Then the factorizations

$$p_2 ... p_m = (u q_2) q_3 ... q_m$$

are essentially equivalent by induction. Since p_1 and q_1 are essentially equivalent, we are done. □

[8] Suppose it had an infinite branch $a_0, a_1, ...$; note $a_0 \mid a_1 \mid ...$ and $(a_n) \neq (a_{n+1})$ as a_n is the product of a_{n+1} and a non-unit. This would induce an infinite chain of ideals $(a_0) \leq (a_1) \leq ...$ which doesn't terminate, contradicting Theorem 3.35.

[9] Formally, this means we can set up the bijection required in the definition of a Unique Factorization Domain.

Chapter 4

Modules

I presume you have a basic understanding of a field and vector space from linear algebra. From now on, we work exclusively with **unital commutative rings**, which is basically a field but without multiplicative inverses. (In fact, one could consider a field to be such a ring with multiplicative inverses.)

Just as we extend fields to vector spaces, we can extend rings to **modules**. One might expect modules to be very similar to fields, but just by dropping the inverse requirement in fields, we get some very counterintuitive consequences.

> **Definition 4.1 (Module)**: Given a ring $(R, +_R, \times)$, an ***R*-module** is a triple $(M, +_M, \cdot)$, where $+$ is an associative binary function of the form $M \times M \to M$ and $\times$ is a binary function of the form $R \times M \to M$ that satisfy the following properties:
>
> - $(M, +_M)$ forms an abelian group.
> - Addition distributes over scalar multiplication: $(r +_R s) \cdot m = r \cdot m +_M s \cdot m$.
> - Scalar multiplication is associative with ring multiplication: $(r \times s) \cdot m = r \cdot (s \cdot m)$.
> - The ring multiplicative identity 1_R of R is also a scalar multiplicative identity: $1 \cdot m = m$.
>
>   We drop this requirement when R is a ring without 1 (though we will seldom ever study this case).
>
> (Note that $r, s \in R$ and $m \in M$.)

We will often use $+$ to denote addition under both R and M. Furthermore, we will usually drop the $\times$ and $\cdot$ operators and implicitly perform $\times$ and $\cdot$ before $+_R$ and $+_M$, just as in usual arithmetic.

Here are some examples of modules:

- Any vector space.
- Just as $\mathbb{R}^n$ is an $\mathbb{R}$-vector space, $\mathbb{Z}^n$ is a $\mathbb{Z}$-module.
- And just as $\mathbb{R}^{\mathbb{N}}$ is a $\mathbb{R}$-vector space, $\mathbb{Z}^{\mathbb{N}}$ is a $\mathbb{Z}$-module.

There is also a notion of a **submodule**, which is exactly what you'd expect.

> **Definition 4.2 (Submodule)**: Given an R-module M and $N \subseteq M$, we say N is a **submodule** of M if N is closed under addition and scalar multiplication. We denote this as $N \leq M$.

Note N may be considered as an R-module in its own right.

Exercise 4.3 is **really important** and you should not move on until you have answered it.

Exercise 4.3: Given any ring R, we may consider R as an R-module. (Work out the details!) What are the submodules of R?

4.1 Homomorphisms and Isomorphisms

You doubtless are familiar with the notion of **linear maps** in linear algebra. Another term for them is **vector space homomorphisms**. And **module homomorphisms** are pretty much the same.

> **Definition 4.4 (Homomorphism)**: A **homomorphism** $\varphi : M \to P$ (where M and S are both R-modules; it is important that the underlying ring is the same) is a function where
>
> $$\varphi(m_1 + m_2) = \varphi(m_1) + \varphi(m_2) \text{ and } \varphi(rm) = r\varphi(m)$$
>
> for all $m_1, m_2 \in M$ and $r \in R$.

By the way, just as you would expect, $\ker \varphi \leq M$ and $\operatorname{im} \varphi \leq P$.

In a group we have normal subgroups. In a ring we have ideals. What is the equivalent for modules? In other words, which submodules have representation invariance, and which submodules may appear as the kernel of a module homomorphism? The answer turns out to be all of them.

Note that given modules $N \leq M$, we have representation invariance if we define N-coset addition and multiplication as

- $(m_1 + N) + (m_2 + N) = (m_1 + m_2) + N$
- $r(m + N) = (rm) + N$,

respectively. Thus

- $\frac{M}{N}$ is a well-defined quotient module,
- and trivially, any $N \leq M$ is the kernel of the obvious homomorphism $\varphi : M \to \frac{M}{N}$ (so any N may appear as a kernel).

The isomorphism theorems are precisely what you think they are:

1. If $\varphi : M \to P$ is a homomorphism, then $\frac{M}{P} \cong \operatorname{im} \varphi$.

2. If $N, S \leq M$, then $\frac{N+S}{N} \cong \frac{S}{N \cap S}$.
3. If $N \leq S \leq M$, then

$$\frac{\frac{M}{N}}{\frac{S}{N}} \cong \frac{M}{S}.$$

4. If $N \leq M$, then there is an inclusion-preserving bijection between the S where $N \leq S \leq M$ and the submodules of $\frac{M}{N}$.

This exercise will be important when proving Theorem 4.48. It is also worth working through in its own right.

Exercise 4.5: Show that given modules M_1 and M_2 with submodules $N_1 \leq M_1$ and $N_2 \leq M_2$,

$$\frac{M_1 \oplus M_2}{N_1 \oplus N_2} \cong \frac{M_1}{N_1} \oplus \frac{M_2}{N_2}.$$

This will allow us to inductively deduce that given modules M_i and $N_i \leq M_i$ for $1 \leq i \leq k$,

$$\bigoplus_{i=1}^{k} M_i / \bigoplus_{i=1}^{k} N_i \cong \bigoplus_{i=1}^{k} \frac{M_i}{N_i}.$$

4.2 Finitely Generated Modules

In linear algebra we are often interested in finite-dimensional vector spaces. Module theory is the same. But there are many key differences. For one, not every finitely generated module has a basis. (Those that have a basis are called **free modules**.) Furthermore, there are finitely generated modules with submodules that cannot be finitely generated! (Modules whose submodules are all finitely generated are called **Noetherian**.)

In linear algebra we may conflate a lot of concepts, such as “finite dimensional” and “finitely generated”, because they are all equivalent. Module theory is not so kind, so we will be a lot more precise. To that end, we will be defining some familiar concepts in linear algebra from the ground up.

For each of these definitions, $m_i \in M$ and $r_i \in R$ for each sensible i.

Definition 4.6 (Independent Set): In an R-module M, an **independent set** of elements $\{m_1, ..., m_k\}$ is one where the only solution to the equation

$$r_1 m_1 + ... + r_k m_k = 0$$

is $r_1 = ... = r_k = 0$.

Definition 4.7 (Spanning Set): In an R-module M, a **spanning set** (or alternatively, **generating set**) of elements $\{m_1, ..., m_k\}$ is one where for every $m \in M$, there is a solution to the equation

$$r_1 m_1 + ... + r_k m_k = m.$$

Definition 4.8 (Finitely Generated Modules): A module is **finitely generated** if it has a finite spanning set.

Definition 4.9 (Basis and Free Module): A **basis** for a module is an **independent** and **spanning** set. A module with a basis is **free**.

In linear algebra,

- every vector space has a basis;
- every finitely generated vector space has a finite basis.

Neither of these are true for modules. There are modules without a basis, and a module being finitely generated does not even imply it has a basis, let alone a finite basis.

Example 4.10: Consider $\mathbb{Q}$ as a $\mathbb{Z}$-module. It has no basis because any generating set must contain two distinct rationals q_1 and q_2, and there is always a non-trivial solution to $z_1 q_1 + z_2 q_2 = 0$ where $z_1, z_2 \in \mathbb{Z}$.

Example 4.11: More generally, for any ring R that is not a field and a non-trivial ideal I (i.e. $I \neq 0$ and $I \neq R$), $\frac{R}{I}$ is a module with no basis. Any non-empty set $\{r_1 + I, ..., r_k + I\}$ is not independent because for any $i \in I$,

$$i(r_1 + I) + ... + i(r_k + I) = 0 + I.$$

Furthermore, $\frac{R}{I}$ is finitely generated by 1. So it is also an example of a finitely generated module without a basis.

Bases behave differently in vector spaces and modules; one might say that bases in modules are not well-behaved. But even more distressingly, finitely generated modules are not well-behaved either. To be more specific:

- In linear algebra, any subspace of a finitely generated vector space is finitely generated. (This is more familiarly said as “any subspace of a finite-dimensional vector space is finite-dimensional” because being finitely generated and having a finite basis are the same thing.)
- In module theory, there are finitely generated modules with submodules that are not finitely generated.

Example 4.12: Let me give you a concrete example of a ring R with an ideal I that is not finitely generated. Considering R as an R-module, R is obviously finitely generated. But I will be a submodule of R that is not finitely generated.

Specifically, we define R to be the ring

$$\mathbb{Z}[x_0, x_1, ...] = \bigcup_{i=0}^{\infty} \mathbb{Z}[x_0, ..., x_i],$$

the infinite-variable ring of integer polynomials where each individual polynomial has a finite number of variables.

Now take the ideal $I = (x_0, x_1, ...)$, i.e. the ideal generated by the set $\{x_0, x_1, ...\}$. Note I is the set of polynomials with a constant term of 0.

There is no finite set of polynomials $\{f_0, ..., f_n\}$ that generates I, for there is always some maximum k such that the coefficient of x_k in some f_i is non-zero. **Because every polynomial in I has a constant term of 0**, coefficient matching insinuates that no linear combination of $\{f_0, ..., f_n\}$ in $\mathbb{Z}[x_0, x_1, ...]$ yields x_k.

By the way, we could have used any arbitrary ring instead of $\mathbb{Z}$ for this example. We just picked $\mathbb{Z}$ for concreteness.

Sure, not every finitely generated module has the property that all its submodules are finitely generated. But many of the modules we care about do have that property,[1] and it is a very convenient property to have (for it makes modules more well-behaved), so we are interested in studying these modules.

> **Definition 4.13 (Noetherian Modules)**: An R-module is **Noetherian** if all of its submodules are finitely generated over R.

Obviously a Noetherian module is finitely generated over R as well, because every module is a submodule of itself.

By the way, we say a ring R is **Noetherian** if R may be considered as a Noetherian R-module. Concretely, this is equivalent to every ideal of R being finitely generated.

Since being Noetherian is such a useful property, we'd like to characterize which modules are Noetherian.

[1] The complexity of the example I gave is a testament to this fact: there were no simpler modules that lack this property.

Theorem 4.14 (Characterizations of Noetherian Modules): An R-module M is **Noetherian** if any of three equivalent conditions hold:

1. Every submodule of M is finitely generated under R.
2. Any chain of increasing submodules $N_0 \leq N_1 \leq \ldots$ of M must eventually terminate. That is, there is always some k such that for all $i \geq k$, $N_i = N_k$.
3. Any non-empty family of submodules has a maximal element. That is, given a family of submodules $\mathcal{F}$, there is some $N_{\max} \in \mathcal{F}$ such that for all $N \in \mathcal{F}$, $N_{\max}$ is not a proper subset of N.

Proof of Theorem 4.14: Here is the strategy: we will prove $(1) \Longrightarrow (2)$ and $\neg(1) \Longrightarrow \neg(2)$, establishing $(1) \Longleftrightarrow (2)$. Then we will show $(2) \Longrightarrow (3)$ and $\neg(2) \Longrightarrow \neg(3)$ to finish the proof.

- $(1) \Longrightarrow (2)$: Let $N_\infty = \bigcup_{i=0}^{\infty} N_i$. Then N_∞ is obviously a submodule of M and thus is finitely generated by some set S. Then there is some N_i such that $S \subseteq N_i$.[2] From this we may conclude $N_i = N_\infty$.[3]
- $\neg(1) \Longrightarrow \neg(2)$: Suppose there is some $N \leq M$ that is not finitely generated under R. Then we may inductively construct a sequence of elements in N as follows:
  - n_0 is just some arbitrary element of N.
  - n_{i+1} is an element of N that is not a linear combination of $\{n_0, \ldots, n_i\}$.[4]

  Now let N_i be the submodule of N generated by $\{n_0, \ldots, n_i\}$. Note

  $$N_0 < N_1 < \ldots$$

  which is our witness to $\neg(2)$.
- $(2) \Longrightarrow (3)$: Suppose $\mathcal{F}$ is a family of submodules of M. Define an increasing chain of submodules from $\mathcal{F}$ as follows:
  - N_0 is just some arbitrary submodule in $\mathcal{F}$.
  - If N_i is maximal, then $N_{i+1} = N_i$, and otherwise, N_{i+1} is some submodule such that $N_i < N_{i+1}$.

  Since the sequence must terminate, the N_i it terminates at must be a maximal submodule.
- $\neg(2) \Longrightarrow \neg(3)$: If we have a chain $N_0 < N_1 < \ldots$ then the family $\{N_0, N_1, \ldots\}$ does not have a maximal submodule.

□

[2] For each $s \in S$, if $s \in N_\infty$ then there must be some N_i such that $s \in N_i$. As S is finite we can just take the largest such i.

[3] Because $S \subseteq N_i$, the submodule of N generated by S, i.e. N_∞, is also a submodule of N_i.

[4] Such an n_{i+1} always exists because otherwise, N would be finitely generated by $\{n_0, \ldots, n_i\}$, contradiction.

As a digression, there is a notion of **Noetherian induction** that is highly related to the terminating chain condition on Noetherian modules. The idea is as follows: if we want to show a property holds for every submodule in a family of submodules $\mathcal{F}$, it suffices to show that for any $N \in \mathcal{F}$,

$$\{P \in \mathcal{F} \mid P > N\} \text{ all satisfy the property} \Longrightarrow N \text{ satisfies the property.}$$

> **Theorem 4.15**: It is a very useful fact that if $N \leq M$, then
>
> $$\frac{M}{N} \text{ and } N \text{ Noetherian} \Longleftrightarrow M \text{ Noetherian.}$$

Proof of Theorem 4.15: For the forward direction, we use the chain characterization of Noetherian modules. Given any chain $M_0 \leq M_1 \leq \ldots$ of submodules of M, note that

$$M_0 \cap N \leq M_1 \cap N \leq \ldots \text{ and } \frac{M_0 + N}{N} \leq \frac{M_1 + N}{N} \leq \ldots$$

both terminate. Suppose they both terminate by M_i and take any $k \geq i$. Then note that for any $m \in M_k$, $m \in M_k + N = M_i + N$.[5] Thus there is some $m_i \in M_i$ such that $m - m_i \in N$. Then $m - m_i \in M_k \cap N = M_i \cap N \subseteq M_i$, meaning that $m_i + (m - m_i) \in M_i$.

Thus $M_k \subseteq M_i$, meaning $M_0 \leq M_1 \leq \ldots$ terminates at M_i.

For the backward direction,

- $\frac{M}{N}$ is Noetherian by the Fourth Isomorphism Theorem, as every chain of submodules of $\frac{M}{N}$ corresponds to a chain of submodules of N containing M. The latter terminates as M is defined to be Noetherian.
- and N is Noetherian as every submodule of a Noetherian submodule is Noetherian.

□

There is a very similar fact to Theorem 4.15 concerning solvable groups. If $N \trianglelefteq G$, then

$$\frac{G}{N} \text{ and } N \text{ solvable} \Longleftrightarrow G \text{ solvable.}$$

The proof is also very similar.

Example 4.16: If R is a Noetherian ring and M is a finitely generated R-module, show that M is a Noetherian R-module.

Solution to Example 4.16: It is easy to prove R^n is a finitely generated R-module via induction. Now suppose M is generated by $\{m_1, \ldots, m_n\}$. Now define $\varphi : R^n \to M$ where

[5] The Fourth Isomorphism Theorem says that $\frac{M_k+N}{N} = \frac{M_i+N}{N} \Longleftrightarrow M_k + N = M_i + N$.

$$\varphi : (r_1, ..., r_n) \mapsto r_1 m_1 + ... + r_n m_n.$$

Clearly φ is surjective. So by the First Isomorphism Theorem, $\frac{R^n}{\ker \varphi} \cong M$. Since R^n is Noetherian, so is $\frac{R^n}{\ker \varphi}$. □

An easy corollary of Example 4.16 is that if $n \in \mathbb{N}$ and R is a Noetherian ring, R^n is a Noetherian module.

Finally, we show how we may use Noetherian rings to generate Noetherian polynomial rings.

> **Theorem 4.17 (Hilbert's Basis Theorem)**: If R is a Noetherian ring, then $R[x]$ is too.

As a trivial consequence we will get that $R[x_1, ..., x_n]$ is Noetherian too.

Proof of Theorem 4.17: We will show that any ideal I of $R[x]$ is finitely generated. For each $n \in \mathbb{N}$, we define the ideal J_n as follows:

$$J_n = \{r \in R : \exists f \in R[x] \text{ with} \deg(f) < n \text{ and } f + rx^n \in I\}.$$

Note $J_0 \leq J_1 \leq ...$ because

$$f + rx^n \in I \Longrightarrow rf + rx^{n+1} \in I.$$

(It is easy to verify each J_n is an ideal.)

Since R is Noetherian the chain $J_0 \leq J_1 \leq ...$ stabilizes at some J_k and all of these ideals are finitely generated.

For each $n \leq k$, let F_n be a finite set of polynomials of degree i whose leading coefficients generate J_n. Define $F = \bigcup_{n=0}^{k} F_n$. We show that every polynomial $f \in R[x]$ is generated by the polynomials in F. To do this, we induct on $\deg f$. Obviously $f = 0$ works, so now we can presume that $\deg f$ is a natural number.

1. $\deg(f) = n \leq k$: Let r be the leading coefficient of f. Note by definition that $r \in J_n$, meaning we can subtract some linear combination of polynomials in F_n to get a polynomial of degree less than n. Said polynomial can be generated from F by the induction hypothesis.
2. $\deg(f) = n > k$: Let r be the leading coefficient of f, meaning $r \in J_n = J_k$. We can find some linear combination g of polynomials in F_k such that the leading coefficient of g is r. Then we may multiply g by x^{n-k} and subtract that from f to get a polynomial of degree less than n. Said polynomial can be generated from F by the induction hypothesis.

□

This proof is a proof by contradiction, meaning that it does not actually give us a basis of $R[x]$, it only tells us that one exists. When F is a field, we may find a **Gröbner basis** of any ideal I of $F[x_1, ..., x_n]$. We will not cover how to do that here.

Theorem 4.17 also holds for modules in general. If M is a Noetherian R-module, then $M[x]$ is a Noetherian $R[x]$-module.

4.3 Categorical Module Constructions

Given a set of R-modules, we may either

- take the **product** of these modules,
- take the **coproduct** of these modules,
- or take the **tensor product** of two of these modules.

Furthermore given a ring R and set S we may construct the **free R-module** $\mathrm{Free}(S)$.

These constructions are all highly connected with category theory, so this is the perspective we will introduce them with.

As a preliminary definition we introduce an **isomorphism**.

> **Definition 4.18 (Isomorphism)**: In a category $\mathcal{C}$, an arrow $f : c \to d$ is an isomorphism if there exists an arrow $g : c \to d$ such that
>
> - $g \circ f = \mathrm{id}_c$
> - $f \circ g = \mathrm{id}_d$.
>
> Naturally we say g is the **inverse** of f, because composing the two arrows gives identity.

We have discussed the product of two **categories** before. There is also a notion of the product of two **objects**. Before we go into the definition, some motivation is in order. In the category Set, we would want the product object of sets S_1 and S_2 to be $S_1 \times S_2$. Similarly in the category group, we would want the product object of groups G_1 and G_2 to be $G_1 \times G_2$. And so on.

Definition 4.19: A **product** of two objects c and d in a category $\mathcal{C}$ is an object $c \times d$ associated with **projection arrows** $\pi_c : c \times d \to c$ and $\pi_d : c \times d \to d$ satisfying the universal property that, for every pair of arrows $f_c : x \to c$ and $f_d : x \to d$ coming from any object x, there exists a unique $f : x \to c \times d$ such that

$$\begin{tikzcd}
 & x \arrow[dl, "f_c"'] \arrow[d, "f"'] \arrow[dr, "f_d"] & \\
c & c \times d \arrow[l, "\pi_c"'] \arrow[r, "\pi_d"] & d
\end{tikzcd}$$

commutes.

We have already covered the product of two sets or two groups. The product of two R-modules is also exactly what you'd expect: the module $M_1 \times M_2$ with elements (m_1, m_2) and addition being defined coordinate-wise and ring multiplication distributing over each coordinate.

Not every pair of objects in a category has a product. As a trivial example, consider a category with only the identity arrows. Then for any distinct objects c and d in this category, $c \times d$ does not exist.

Importantly, a pair of objects may have more than one product. But any two products are equivalent up to **unique isomorphism**, which in category theory is as good as saying two objects are the same. For their arrows behave the same, and an object is just the behavior of the arrows going in and out of it.

Suppose that $c \times d$ and $c \times' d$ are distinct products of c and d. Let $c \times d$ project onto c and d with π_c and π_d. Similarly, let $c \times' d$ project onto c and d with σ_c and σ_d. Then the universal mapping property permits us to draw the unique dashed arrows in the commutative diagram below:

$$\begin{tikzcd}
 & c \times d \arrow[ddl, "\pi_c"'] \arrow[d, dashed] \arrow[ddr, "\pi_d"] & \\
 & c \times' d \arrow[dl, "\sigma_c"] \arrow[d, dashed] \arrow[dr, "\sigma_d"'] & \\
c & c \times d \arrow[l, "\pi_c"'] \arrow[r, "\pi_d"] & d
\end{tikzcd}$$

Figure 13: The dashed arrows form our unique isomorphism.

Furthermore there is also the notion of the product of a family of objects. For instance, if we have a family $\mathcal{F} = \{S_1, S_2, ...\}$ of objects in Set, the product $\prod \mathcal{F}$ ought to be $S_1 \times S_2 \times ....$ So we ought to generalize the notion of a product not just to any finite family of objects (a task that is trivial with recursion), but to **infinite** families of objects, perhaps even those that are uncountable.

Definition 4.20: A **product** of a family of objects $\mathcal{F}$ in a category $\mathcal{C}$ is an object $\prod \mathcal{F}$ along with **projection arrows** $\pi_c : \prod \mathcal{F} \to c$ for every $c \in \mathcal{F}$ satisfying the universal property that, for every family of arrows $\{f_c : x \to c \mid c \in \mathcal{F}\}$ coming from any object x, there exists a unique $f : x \to \prod \mathcal{F}$ such that

$$\begin{tikzcd}
 & x \arrow[dl, "f_c"'] \arrow[d, "f"] \\
c & \prod \mathcal{F} \arrow[l, "\pi_c"']
\end{tikzcd}$$

commutes for every $c \in \mathcal{F}$.

For the same reason, general products are also unique up to isomorphism (so long as they exist).

In category theory, any categorical concept has a notion of a **dual**, where the definition is kept largely the same but the direction of the arrows are swapped. We have already seen this with Definition 2.64. Now we will define the dual of a product, the **coproduct**.

Definition 4.21 (Coproduct): A **coproduct** of a family of objects $\mathcal{F}$ in a category $\mathcal{C}$ is an object $\bigoplus \mathcal{F}$ along with arrows $i_c : \bigoplus \mathcal{F} \to c$ for every $c \in \mathcal{F}$ satisfying the universal property that, for every family of arrows $\{f_c : c \to x \mid c \in \mathcal{F}\}$ pointing into any object x, there exists a unique $f : \bigoplus \mathcal{F} \to x$ such that

$$\begin{tikzcd}
& x \\
c \arrow[ru, "f_c"] \arrow[r, "i_c"] & \bigoplus \mathcal{F} \arrow[u, "f"]
\end{tikzcd}$$

commutes for every $c \in \mathcal{F}$.

As you might expect, the products in a category $\mathcal{C}$ correspond to the coproducts in $\mathcal{C}^{\mathrm{op}}$. This is because we swap the direction of the arrows both in $\mathcal{C}^{\mathrm{op}}$ and in the definition of a coproduct.

What do products and coproducts look like in the category Module_R, where the objects are R-modules for some ring R and the arrows are module homomorphisms? The product is exactly what you would expect: addition is defined component-wise, and multiplication distributes across all components. And the projection functions return the appropriate component.

Formally, we may consider the product module to consist of all elements of the form $\varphi : \mathcal{F} \to \bigcup \mathcal{F}$ where $\varphi(M) \in M$ for each $M \in \mathcal{F}$. The "coefficient" of each $M \in \mathcal{F}$ is encoded as $\varphi(M)$, and the projection $\pi_M : \prod \mathcal{F} \to M$ is merely the function $\varphi \mapsto \varphi(M)$.

The coproduct behaves exactly the same as the product when $\mathcal{F}$ is finite. But this is not the case when $\mathcal{F}$ is infinite. Instead, $\bigoplus \mathcal{F}$ is the submodule of $\prod \mathcal{F}$ with finitely many non-zero coefficients. (Formally, this means $\{M : M \in \mathcal{F} \text{ and } \varphi(M) \neq 0_M\}$ is finite.) The arrows $i_M : M \to \bigoplus \mathcal{F}$ do exactly what you expect them to: m is sent to the function φ where $\varphi(M) = m$ and $\varphi(N \neq M) = 0$.

Why is this the case? Suppose we consider this commutative diagram in Module_R (where i_M behaves as described above):

$$\begin{tikzcd}
& N \\
M \arrow[ru, "f_M"] \arrow[r, "i_M"] & \prod \mathcal{F} \arrow[u, "f"]
\end{tikzcd}$$

We need $f(i_M(m)) = f_M(m)$, and all the consequences that the linearity of f give, but nothing else is required of f. The f_M only determine the behavior of f on finite linear combinations of the form $i_M(M)$, because the linearity constraint on module homomorphisms (in this case, f) only applies to finite sums.

In other words, if we know $f(\varphi_1), f(\varphi_2), \ldots$, we can determine $f(\varphi_1 + \ldots + \varphi_n)$ in a recursive manner: we may determine $f(\varphi_1 + \ldots + \varphi_{n-1})$ and $f(\varphi_n)$. But there is nothing restricting $f(\varphi_1 + \ldots)$. Concretely, if we define $\varphi : M \mapsto 1_M$, there is nothing stopping us from having $f(\varphi) = 0_N$.

This means $\prod \mathcal{F}$ is too "big" to satisfy the universal property. We need $\bigoplus \mathcal{F}$ to only consist of the elements of $\prod \mathcal{F}$ that can be determined by linearity, and this is exactly the elements with finitely many non-zero coefficients.

> **Exercise 4.22**: Verify the products and coproducts in Module are as described above. Concretely, all you need to do is show the existence and uniqueness of f that makes the relevant diagrams commute. Because products and coproducts are unique up to isomorphism, this also shows that the **only** product and coproduct modules are of the forms described.

4.3.1 The Tensor Product

Before I tell you what the tensor product is, I will briefly tell you what it is not. Just as a matrix is not a 2-dimensional array of elements, a tensor is not an n-dimensional array of elements.

The study of tensor products is really the study of **bilinear maps**, just as the study of matrices is really the study of **linear maps**.

> **Definition 4.23 (Bilinear Mapping)**: Let M, N, and P be R-modules for some ring R. Then $\varphi : M \times N \to P$ is bilinear precisely when
> - the map $m \mapsto \varphi(m, n)$ is linear for all $n \in N$;
> - the map $n \mapsto \varphi(m, n)$ is linear for all $m \in M$.

As a warning, note that while $M \bigoplus N$ has an underlying set of $M \times N$, the linear maps $M \bigoplus N \to P$ are not the same as the bilinear maps $M \times N \to P$. As an example, if M, N, and P are all the $\mathbb{Z}$-module $\mathbb{Z}$, then

- $(x, y) \mapsto xy$ is bilinear considered as a function $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$, but not linear when considered as $\mathbb{Z} \bigoplus \mathbb{Z} \to \mathbb{Z}$.
- $(x, y) \mapsto x + y$ is not bilinear considered as a function $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$, but it is linear when considered as $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$.

Just as linearity is preserved when composing linear maps, bilinearity is preserved upon composition with linear maps too.

Theorem 4.24: If $\varphi : M \times N \to P$ is bilinear and $\gamma : P \to Q$ is linear, then $\varphi \circ \gamma : M \times N \to Q$ is bilinear.

(Here M, N, P, and Q are all R-modules for some ring R.)

Exercise 4.25: Verify Theorem 4.24.

We will really quickly establish some category theoretic notions.

Definition 4.26 (Initial Object): An **initial object** c in a category $\mathcal{C}$ is one such that, for any object d in $\mathcal{C}$, there is exactly one arrow $f : c \to d$.

Definition 4.27 (Terminal Object): A **terminal object** c in a category $\mathcal{C}$ is one such that, for any object d in $\mathcal{C}$, there is exactly one arrow $f : d \to c$.

Note that terminal objects are the dual of initial objects. More precisely, c is a terminal object in $\mathcal{C}$ if and only if c is an initial object in $\mathcal{C}^{\text{op}}$.

To define tensor products, we will construct the category of bilinear maps with domain $M \times N$. The objects are these bilinear maps, and the arrows between objects $\varphi : M \times N \to P$ and $\varphi' : M \times N \to Q$ are the functions of the form $\gamma : P \to Q$ such that $\varphi' = \gamma \circ \varphi$.

Definition 4.28 (Tensor Product): A **tensor product** $M \otimes N$ is an initial object in this category.

$$\begin{tikzcd}
M \times N \arrow[r, "\otimes"] \arrow[rd, "\varphi"] & P \arrow[d, "\exists!\gamma"] \\
 & Q
\end{tikzcd}$$

Figure 17: Concretely, this means $\otimes : M \times N \to P$ is a tensor product if and only if it satisfies the universal property that, for all bilinear $\varphi : M \times N \to Q$, there is a unique linear $\gamma : P \to Q$ such that $\varphi = \gamma \circ \otimes$.

The way to prove that a tensor product always exists is via a direct construction. Because the direct construction is infeasible to work with, we are not going to bother doing it.

Two tensor products are equivalent up to unique isomorphism. Suppose $\otimes$ and $\otimes'$ are both tensor products. Then there exist unique $\gamma : P \to Q$ and $\gamma' : Q \to P$ such that

$$\otimes' = \gamma \circ \otimes$$

$$\otimes = \gamma' \circ \otimes'$$

implying that $\otimes = \gamma' \circ \gamma \circ \otimes$. Since γ and γ' are unique, they form our unique isomorphism.

This is why we may refer to **the** tensor product, rather than **a** tensor product, for any two are essentially equivalent. We write $m \otimes n$ to denote where the bilinear map $\otimes$ sends (m, n) to.

A few examples of tensor products:

- The tensor product is obviously commutative (bilinearity and the universal property are symmetric conditions).
- For an R-module M, $R \otimes M$ is the linear map $\varphi : R \times M \to M$ where $\varphi : (r, m) \mapsto rm$.
- For R-modules R^m and R^n and vectors $u \in R^m$, $v \in R^n$, $R^m \otimes R^n$ is the linear map $(u, v) \mapsto uv^T$ (where v^T denotes the transpose of v).

Importantly, $\otimes$ is **not** generally surjective. For example, if R is a field, the codomain of $R^m \otimes R^n$ is R^{mn}. (More precisely, it is the $m \times n$ matrices whose elements are in R. But the two are isomorphic, so it does not matter.)

Now the range of $R^m \otimes R^n$ is not the entirety of R^{mn}. Informally, we only have $m + n$ degrees of freedom: m from u and n from v^T. So the rank of the range is at most dimension $m + n$, which is less than mn when m and n are sufficiently large.

To elaborate, saying that $uv^T = M$ for some $m \times n$ matrix M specifies a system of mn linear equations with $m + n$ variables (m from u and n from v^T). It is well-known that if $m + n < mn$, then some of these systems of equations have no solutions. Thus $\otimes$ cannot be surjective in this case.

> **Definition 4.29 (Tensors)**: We call the **elements** of P in the tensor product $\varphi : M \times N \to P$ **tensors**.

Note the difference between a **tensor**, an element of an R-module P, and a **tensor product**, the bilinear map $\otimes : M \times N \to P$ satisfying a universal property.

> **Definition 4.30**: We say that a **simple tensor** is one that can be expressed in the form $m \otimes n$. In other words, a **simple tensor** is in the range of the tensor product.

We have established that not every tensor is simple. In fact, most of them aren't. However, enough tensors are simple to generate the codomain of the tensor product.

Informally, here is why: look at the universal property of the tensor product. We need our unique linear γ to behave correctly on simple tensors, and as a consequence their behavior is defined on linear combinations of simple tensors. But if the codomain were to be bigger, then γ can do whatever it wants on the elements in the codomain of $\otimes$ that are not spanned, because it does not affect whether $\varphi = \gamma \circ \otimes$.

> **Theorem 4.31**: Every tensor is the linear combination of some simple tensors.

Proof of Theorem 4.31: Suppose not. That is, suppose that the tensor product is $\otimes : M \times N \to P$ and there exists some $p \in P$ that is not generated by the set $\{m \otimes n : m \in M, n \in N\}$. Further let the module generated by $\{m \otimes n : m \in M, n \in N\}$ be S.

Now define $\varphi : M \times N \to S$ as follows: $\varphi : (m, n) \mapsto m \otimes n$. In other words, φ is just $\otimes$ with a restricted codomain. They are essentially the same function.

But now any $\gamma : P \to S$ where $\gamma \upharpoonright P = \mathrm{id}_P$ has $\varphi = \gamma \circ \otimes$ has $\varphi = \gamma \circ \otimes$. Because p is linearly independent from S, we may have $\gamma(p)$ be any element of S while still having γ be linear. The non-uniqueness of γ means $\otimes$ fails to satisfy the universal property, contradiction. □

4.3.2 Free Modules

Here we will briefly consider general rings, that is, rings that are not necessarily unital or commutative. When we do so, we will make it very clear.

Every vector space has a basis. This is not the case for modules.

> **Definition 4.32 (Free Module)**: A **free module** is a module with a basis. That basis may either be finite or infinite.

Over **commutative rings** R, any two bases of a free R-module must have the same cardinality, whether this cardinality is finite or infinite.

> **Theorem 4.33**: Suppose R is a unital commutative ring. If an R-module M has bases A and B, then $|A| = |B|$.

We will be skipping a lot of details in the proof. It will be very instructive to convince yourself of the veracity of each step. In particular, identify where exactly we use the commutativity of R in this argument, because as we will soon see, it falls completely asunder when R is not commutative.

Proof of Theorem 4.33: Take a maximal ideal I of R. If we have a basis of M with size n, then $M \cong R^n$, meaning $M \otimes \frac{R}{I} \cong \left(\frac{R}{I}\right)^n$.[6]

Now note $\left(\frac{R}{I}\right)^n$ is a vector space as $\frac{R}{I}$ is a field. Since the dimension of a vector space is fixed, so must n be. □

> **Definition 4.34 (Rank of a Free Module)**: The **rank** of a free module whose bases all have the same cardinality is said cardinality.

We have just shown that two bases of a free module over a unital commutative ring always have the same size. However, this is not the case for free modules over non-commutative rings.

Example 4.35: Here is the standard example. I am referencing https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=5487&context=etd-project, starting from page 34.

Take a field F and a vector space V over F with a **countable** basis $\{e_1, ...\}$. Then let $R = \operatorname{Hom}(V, V)$, that is, let R be the ring of linear maps from V to V.

Now let n be an arbitrary positive number and take the unique linear map f_i specified by

$$f_i(e_k) = e_{\frac{k-i}{n}+1} \text{ if } n \mid k - i, \text{otherwise } 0$$

for each $1 \leq i \leq n$. It is very straightforward to verify $\{f_1, ..., f_n\}$ forms a basis of R. Informally, they are independent because each f_i deals with a different residue modulo n, and they span all of R because the f_i cover all the residues.

However, two bases of an **infinitely generated** free R-module must have the same cardinality, whether R is commutative or not.

> **Theorem 4.36**: More formally, if A is an infinite basis of a free R-module M and B is another basis of M, then $|A| = |B|$.

Proof of Theorem 4.36: First note B is not finite, otherwise

- each $b \in B$ can be represented as a **finite** linear combination in A,
- for each $b \in B$, there is some maximal k such that the coefficient of a_k is maximal,

[6] When we write $M \otimes \frac{R}{I}$, we are really referring to the codomain of the tensor product. This may be considered a slight abuse of notation.

- and taking the maximum over all such k, we see that there must exist (infinitely many) a_i with coefficient 0 in the representation of all $b \in B$,

which easily leads to a contradiction. (Check it yourself!)

Now for each $b \in B$, let A_b be the subset of A that generates b. (Note A_b is unique as A is a basis.) Now as each A_b is finite, and $\bigcup_{b \in B} A_b = A$, we have that

$$|A| = \left|\bigcup_{b \in B} A_b\right| \leq |B|.$$

Similarly $|B| \leq |A|$. So $|A| = |B|$, as desired. □

So far we have only described a free module and its cardinality. But we may also generate a free module.

> **Definition 4.37 (Free Module)**: Given a set S and ring R, the **free R-module** $\mathrm{Free}(S)$ is the module with elements $f : S \to R$ where $\{s \in S \mid f(s) \neq 0\}$ is finite.

In other words, the free module is the coproduct $\bigoplus_{s \in S} R$, where the copies of R are indexed by the set S.

> **Theorem 4.38**: If S is a basis for a module M, then $M \cong \mathrm{Free}(S)$.

This may be a little counterintuitive when we have a module M that has bases of different sizes. But it is certainly possible to have a scenario where $R \cong R^2$. This could only happen if R is not commutative, because we have established that the bases of a module over a unital commutative ring have fixed cardinality.

> **Theorem 4.39**: The free module is solely determined by the cardinality of S. Formally, if $|A| = |B|$, then $\mathrm{Free}(A) \cong \mathrm{Free}(B)$.

The isomorphism between $\mathrm{Free}(A)$ and $\mathrm{Free}(B)$ is really obvious. Furthermore, if S is a basis for the free module M, then $M \cong \mathrm{Free}(S)$.

Furthermore if $|A|$ is finite then $\mathrm{Free}(A) \cong R^{|A|}$.

Just like the free group, the free module also satisfies the universal property.

Theorem 4.40 (Universal Property of the Free Module): Given a set S, an R-module M, and a function $\varphi : S \to M$, there is a unique homomorphism $\psi : S \to M$ such that $\psi \restriction S = \varphi$.

$$\begin{tikzcd}
S \arrow[r, "\mathrm{id}"] \arrow[dr, "\varphi"] & \mathrm{Free}(S) \arrow[d, "\psi"] \\
& M
\end{tikzcd}$$

Figure 18: Equivalently, there is a unique ψ such that the given diagram commutes.

Exercise 4.41: The proof of Theorem 4.40 is spiritually identical to that of the proof of Theorem 2.62. Carry out the details and prove Theorem 4.40.

Free modules are useful for two reasons:

- Presentations of groups are defined as a quotient of a free group and the least normal subgroup generated by some relations. Presentations of modules are much the same.
- They are useful for the **structure theorem**. Roughly speaking, an n-dimensional vector space over a field F is isomorphic to F^n. Similarly, a finitely generated R-module, where R is a Principal Ideal Domain, is isomorphic to some quotient of R^n. So free modules underlie the decomposition of a finitely generated module over a Principal Ideal Domain, much as F^n underlies the decomposition of an n-dimensional vector space.

The former we will not elaborate on further. The latter will be gone over in the next section.

4.4 Modules Over a Principal Ideal Domain

Take a finitely generated vector space. Every subspace of that vector space is finitely generated. Distressingly, the same is not true of free modules. To be more precise, free modules are not necessarily even Noetherian, let alone free. (Recall Example 4.12 and note that R is obviously free as an R-module.)

However, as we may recall from Example 4.16, if R is Noetherian, then every finitely generated R-module is Noetherian.

We can take things even further when we have a free module over a Principal Ideal Domain. Here we get the highly desirable result that every submodule is free.

> **Theorem 4.42**: If R is a Principal Ideal Domain and M is a free R-module, then every submodule of M is also free. Furthermore, the rank of the submodule is less than or equal to the rank of M.

Note this does not require M to have a finite rank.

This proof requires a lot of set theoretic knowledge. It is not crucial to the rest of the text, so if you find yourself completely lost, it is fine to skip it. To learn the background required for this proof (and much more), I highly recommend Professor Ernest Schimmerling's text "A Course on Set Theory".

Proof of Theorem 4.42: Recall that for some B we have $M \cong \operatorname{Free}(B)$, where $\operatorname{Free}(B)$ is the module with elements $f : B \to R$ with $\{b \in B \mid f(b) \neq 0\}$ finite. For notational convenience, we will be directly reasoning about $\operatorname{Free}(B)$.

Suppose B is a basis of M. By the Well-Ordering Principle, we may take some well-ordering on B. For each $b \in B$, define $A_b = \{a \in B \mid a \leq b\}$ and M_b to be the submodule of M generated by the elements in A_b. (Formally, $M_b \cong \operatorname{Free}(A_b)$.)

Now for any $N \leq M$, define $N_b = N \cap M_b$ for each $b \in B$. Further consider the projection $\pi_b : N_b \to R$ where $\pi_b : f \mapsto f(b)$. Now note $\operatorname{im} \pi_b$ is an ideal of R, and as R is a Principal Ideal Domain, we may consider $\operatorname{im} \pi_b = (r_b)$. Now for each r_b, take some $n_b \in N_b$ such that $\pi_b(n_b) = r_b$, with the stipulation that if $r_b = 0$ we must have $n_b = 0$.

Now we claim the non-zero n_b form a basis of N. To show this, we use a set-theoretic method of induction known as **transfinite induction**. We may do this because every well-ordering is isomorphic to some ordinal through the Mostowski collapse.

For convenience, we are now going to take every basis of a free module to be an ordinal α. This is because the only thing that matters in the basis of a module is the cardinality; isomorphism takes care of the rest.

First we prove by transfinite induction that the non-zero n_b are linearly independent.

- $\beta = 0$: Obvious.
- $\beta \Rightarrow \beta + 1$: By the inductive hypothesis, $\{n_b \mid b \in \beta, n_b \neq 0\}$ is linearly independent. If $\{n_b \mid b \in \beta, n_b \neq 0\} \cup \{n_\beta\}$ is linearly dependent, this must be because $\pi_b(n_\beta) = 0$, otherwise there is no way to get

$$\pi_b\left(r_\beta n_\beta + \sum r_b n_b\right) = 0,$$

which is obviously required if

$$r_\beta n_\beta + \sum r_b n_b = 0.$$

But then this means $r_b = 0$, which means $n_b = 0$. So even in this case, $\{n_b \mid b \in \beta + 1, n_b \neq 0\}$ is linearly independent.

- β is a limit ordinal: By the inductive hypothesis, $\{n_a \mid a \in \alpha\}$ is a basis of N_α for each $\alpha < \beta$ as $N_\alpha \leq \mathrm{Free}(\alpha)$. Because the coproduct is only concerned with **finite** sums, the union of all these bases must be linearly independent, for any finite sum must be contained in some N_α.

Now we prove these bases $\{n_b \mid b \in \beta, n_b \neq 0\}$ generate N. Obviously whatever it generates is a submodule of N. But does it generate every $n \in N$? Again, we employ transfinite induction:

- $\beta = 0$: Obvious.
- $\beta \Rightarrow \beta + 1$: Every $f \in N$ can be expressed as $f_\beta + r_\beta n_\beta$ by definition of β, and by the inductive hypothesis, f_β is generated by $\{n_b \mid b \in \beta, n_b \neq 0\}$.
- β is a limit ordinal: Each $f \in \mathrm{Free}(\beta)$ is contained in some $\mathrm{Free}(\alpha)$ for $\alpha < \beta$ (recall what a coproduct is), and by the inductive hypothesis, $\{n_b \mid b \in \alpha, n_b \neq 0\}$ already generates α.

□

Though the proof is long, I want you to note that there was nothing particularly difficult in it. The setup requires some background in set theory, but it is all straightforward if you are familiar with it. Just by recalling the definition of a coproduct, the inductive cases where β was a limit ordinal were trivial. The step $\beta \Rightarrow \beta + 1$ is the trickiest part in that it requires a little thought. But still, it does not demand too much from you.

One last thing. We have not explicitly checked that these bases $\{n_b \mid b \in \beta, n_b \neq 0\}$ have smaller or equal cardinality to β. But this is really obvious, and the proof was running long enough as is, so we just note it now.

Exercise 4.43: If you did not understand the proof due to a lack of a set theoretical background, reconstruct it for free modules with a finite basis.

Now we turn our attention to finitely generated modules over a Principal Ideal Domain. First we start with the free modules. Because these modules have finite rank, we can say something a lot stronger than Theorem 4.42: not only can you say N is free, there exists a basis of N that is closely related to some basis of M.

Theorem 4.44 (Smith Normal Form): If R is a Principal Ideal Domain, M is a free R-module with **finite** rank k, and $N \leq M$ has rank n, then there is some basis $\{m_1, ..., m_k\}$ of M and elements

$$r_1 \mid r_2 \mid ... \mid r_n$$

in R where $r_1 m_1, ..., r_n m_n$ is a basis of N.

Furthermore, the $r_1, ..., r_n$ are unique up to multiplication by units. That is, regardless of the basis $\{m_1, ..., m_k\}$, there is no other way to construct a basis $r_1 m_1, ..., r_n m_n$ of N.

We already know from Theorem 4.42 that N has a basis. The important part is that we may use a basis of M to construct a basis of N in Theorem 4.44.

If such a basis exists, then the inclusion map $\mathrm{id} : N \to M$ can be represented as the matrix

$$\begin{pmatrix} r_1 & 0 & \dots & 0 \\ 0 & r_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

with bases $\{r_1 m_1, ... r_n m_n\}$ for N and $\{m_1, ..., m_k\}$ for M. (If the ranks of N and M are the same, then there are no extra zeroes below.) The converse is also true: if we show that **some** change of basis leads to such a representation of id, then we are done.

We will also use the fact that every pair of elements in a Principal Ideal Domain has a greatest common divisor.

Exercise 4.45: If you have not already proved this yourself, do it now by showing that d is a greatest common divisor of (a, b) if $(d) = (a, b)$ and $d \neq 0$.

Proof of Theorem 4.44: Start with any bases of N and M, and take the matrix of $\mathrm{id} : N \to M$ under these bases.

Recall that on a matrix we may

- swapping rows/columns,
- multiplying a row/column by a unit,
- adding a multiple of some row/column to another row/column,

for these are precisely the operations that are invertible over modules. Row operations represent a change of the basis in N, and column operations represent a change of the basis in M.

First we want to show that we may get the greatest common divisor of all the entries as an entry via changes in bases. It turns out all we need to do is get the greatest common divisor of two entries in the same row/column. Here is how.

For entries (a, b) and some greatest common divisor d of a and b, there are some $x, y \in R$ such that $xd = a$ and $yd = b$. Since $\gcd(x, y) = 1$, there are some $u, v \in R$ such that $ux + vy = 1$. Here is explicitly how you go about the transformation:

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ (u + v - 1)x + y \end{pmatrix} \to \begin{pmatrix} (u + v)x \\ (u + v - 1)x + y \end{pmatrix} \to \begin{pmatrix} (u + v)x \\ y - x \end{pmatrix} \to \begin{pmatrix} ux + vy \\ y - x \end{pmatrix}$$

and as $ux + vy = 1$, we concluded with the matrix $\begin{pmatrix} 1 \\ y-x \end{pmatrix}$. Multiplying every matrix in the process by d, we see how we may get the matrix $\begin{pmatrix} d \\ d(y-x) \end{pmatrix}$ from $\begin{pmatrix} a \\ b \end{pmatrix}$.

To get the greatest common divisor of the elements in a column, we "propagate upwards". Here is an example:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \to \begin{pmatrix} a \\ \gcd(b,c) \\ ? \end{pmatrix} \to \begin{pmatrix} \gcd(a,b,c) \\ ? \\ ? \end{pmatrix}$$

And here is how we get the gcd of all the elements:

$$\begin{pmatrix} a & ? & ... \\ \vdots & \vdots & \ddots \end{pmatrix} \to \begin{pmatrix} a & \gcd(...) = d & ... \\ \vdots & \vdots & \ddots \end{pmatrix} \to \begin{pmatrix} \gcd(a,d) & ... \\ \vdots & \ddots \end{pmatrix} \to \begin{pmatrix} \gcd(a,d,...) = r_1 & ... \\ \vdots & \ddots \end{pmatrix}$$

Ignoring the first column, we get the greatest common divisor d of the entries of the rest of the matrix. Now we take the greatest common divisor of a and d, and then we find the greatest common divisor of the first row. We end up getting the greatest common divisor of all the elements. (Formally, this process can be defined inductively.)

Now we may subtract the greatest common divisor r_1 from the other entries in the first row and column to get something in the form

$$\begin{pmatrix} r_1 & 0 & ... \\ 0 & ? & ... \\ \vdots & \vdots & \ddots \end{pmatrix}$$

After that, we may ignore the first row and column and find the Smith Normal Form of the rest of the matrix. (Formally, we may perform induction on the size of the matrix.) This gives us a matrix of the form

$$\begin{pmatrix} r_1 & 0 & ... \\ 0 & r_2 & ... \\ \vdots & \vdots & \ddots \end{pmatrix}$$

How do we know $r_1 \mid r_2$? Regardless of what matrix operations we perform, the greatest common divisor of the entries never changes. Thus $r_1 \mid r_2$.

Furthermore, this line of reasoning shows the Smith Normal Form is unique. As any change in basis may be achieved through row operations (or column operations), this means r_1 is fixed in any Smith Normal Form. Dropping the first element in the bases of N and M to generate a linear map between two smaller matrices, we see that r_2 is fixed for the same reason. So on and so forth.

$$\begin{pmatrix} r_1 & 0 & ... \\ 0 & r_2 & ... \\ \vdots & \vdots & \ddots \end{pmatrix} \to \begin{pmatrix} r_2 & ... \\ \vdots & \ddots \end{pmatrix} \to ...$$

□

Exercise 4.46: We have not explicitly stated the base case for Theorem 4.44 in our proof. State and prove the base case.

Exercise 4.47: It is not immediately obvious that any change of basis may be performed via row operations. Show this. More formally, given two bases $\{e_1, ..., e_n\}$

and $\{f_1, ..., f_n\}$ of a free module over a Principal Ideal Domain, show that we may go from the first basis to the second by

- multiplying elements by units,
- or adding some multiple of one element to another.

(Hint: $\gcd(e_1, ..., e_n) = \gcd(f_1, ..., f_n)$, and you should know how to get the gcd of the elements of $\{f_1, ..., f_n\}$.)

This proof also shows uniqueness of the r_i in the original theorem statement. The Smith Normal Form of any matrix is unique, and even though there are many matrices for $\mathrm{id} : N \to M$ with different bases, the Smith Normal Form of each of them must be the same. (Otherwise, we may chain together changes of bases to find distinct Smith Normal Forms, which is a contradiction.)

Why does Theorem 4.44 require M to have a finite rank? We can see the matrix manipulations in the proof require this. But a priori it is not obvious why the theorem must fail on infinite dimensional M. See https://math.stackexchange.com/questions/4967770/nice-bases-for-submodules-of-infinite-free-modules-over-a-pid for an explanation.

Every finitely generated module over a Principal Ideal Domain, even if it is not free, can be nicely decomposed into the direct sum of some very “simple” modules. This relies very heavily on Theorem 4.44, so for the same reasons, the theorem requires that the module be finitely generated.

> **Theorem 4.48 (Structure Theorem, Invariant Decomposition)**: If R is a Principal Ideal Domain and M is a finitely generated R-module, then there exist some $n \in \mathbb{N}$ and $d_1 \mid d_2 \mid ... \mid d_m$ which are not units in R where
>
> $$M \cong R^n \oplus \frac{R}{(d_1)} \oplus ... \oplus \frac{R}{(d_m)}.$$
>
> Furthermore, this decomposition is unique: n is fixed, m is fixed, and each (d_i) is fixed (meaning each d_i is fixed up to multiplication by a unit).

Proof of Theorem 4.48: Suppose that M is finitely generated by $\{m_1, ..., m_n\} \subset M$. Then define a map $\varphi : R^n \to M$ where $\varphi(r_1, ..., r_n) = r_1 m_1 + ... + r_n m_n$. Since φ is surjective, the First Isomorphism Theorem yields that

$$M \cong \frac{R^n}{\ker \varphi}.$$

Since $\ker \varphi \leq R^n$, Theorem 4.44 implies that there is some basis $\{e_1, ..., e_n\}$ of R^n and some $d_1 \mid ... \mid d_k$ in R such that $\{d_1 e_1, ..., d_k e_k\}$ is a basis of $\ker \varphi$. Thus

$$\begin{aligned}\frac{R^n}{\ker\varphi} &\cong \frac{(e_1)\oplus\ldots\oplus(e_n)}{(d_1e_1)\oplus\ldots\oplus(d_ke_k)\oplus(0)\oplus\ldots\oplus(0)}\\ &\cong \frac{(e_1)}{(d_1e_1)}\oplus\ldots\oplus\frac{(e_k)}{(d_ke_k)}\oplus(e_{k+1})\oplus\ldots\oplus(e_n)\\ &\cong \frac{R}{(d_1)}\oplus\ldots\oplus\frac{R}{(d_k)}\oplus R^{n-k},\end{aligned}$$

where we use Exercise 4.5 to decompose the quotient. This is of the desired form. □

It is easy to show the decomposition is unique. The decomposition in Theorem 4.48 can be related to the decomposition induced by Theorem 4.44,[7] and because Theorem 4.44 induces a unique decomposition, so does Theorem 4.48.

A useful corollary is the Primary Decomposition.

Theorem 4.49 (Structure Theorem, Primary Decomposition): If R is a Principal Ideal Domain and M is a finitely generated R-module, then there exist some $n \in \mathbb{N}$, prime $p_1, \ldots, p_m \in R$ and positive exponents $e_1, \ldots, e_m$ where

$$M \cong R^n \oplus \frac{R}{(p_1^{e_1})} \oplus \ldots \frac{R}{(p_m^{e_m})}.$$

Furthermore, this decomposition is unique: n is fixed, m is fixed, each e_i is fixed, and each (p_i) is fixed (meaning each p_i is fixed up to multiplication by a unit).

Proof of existence for Theorem 4.49: To show such a decomposition exists, take a decomposition

$$M \cong R^n \oplus \frac{R}{(d_1)} \oplus \ldots \oplus \frac{R}{(d_m)}$$

from Theorem 4.48. Since R being a Principal Ideal Domain implies it is also a Unique Factorization Domain, each d_i is of the form $p_1^{e_1} \ldots p_k^{e_k}$. And the Chinese Remainder Theorem gives us that

$$\frac{R}{(d_i)} \cong \frac{R}{(p_1^{e_1})} \oplus \ldots \oplus \frac{R}{(p_k^{e_k})}.$$

Apply this for each of the d_i and you will get the desired decomposition. □

Uniqueness is not hard. The gist is that there is a canonical way to reverse this decomposition, i.e. to take a decomposition from Theorem 4.49 and turn it into a decomposition in the form of Theorem 4.48. Because the invariant decompositions are unique, we will easily be able to conclude that the primary decomposition is too.

Proof of uniqueness for Theorem 4.49: To be more precise, we write

[7] This is literally how the proof works.

$$M \cong R^n \oplus \left[R/\left(p_1^{e_{1,1}}\right) \oplus ... \oplus R/\left(p_1^{e_{1,m}}\right)\right] \oplus ... \oplus [R/(p_n^{e_{n,1}}) \oplus ... \oplus R/(p_n^{e_{n,m}})]$$

where each of the p_i are distinct[8] and the exponents are written in increasing order, i.e. $j < k \Longrightarrow e_{i,j} \leq e_{i,k}$. Here we permit some of the exponents to be 0 so that the number of exponents for each prime is the same, with the added stipulation that there must be some i such that $e_{i,j} \neq 0$ for all j.[9] Say that such a decomposition is a **padded primary decomposition**.

Now for each $1 \leq j \leq m$, we define d_j to be

$$R/\left(p_i^{e_{1,j}}\right) \oplus ... \oplus R/(p_n^{e_{n,j}}).$$

It is easy to check that $d_1 \mid ... \mid d_m$, each of the d_j is not a unit, and $M \cong R^n \oplus \frac{R}{(d_1)} \oplus ... \oplus \frac{R}{(d_m)}$, so this is a valid invariant decomposition.

Now note two distinct padded primary decompositions would produce distinct invariant decompositions, and by Theorem 4.48, we know that distinct invariant decompositions cannot be isomorphic. So the padded primary decomposition must be unique. From here it is incredibly easy to conclude that the primary decomposition itself must be unique as well. □

Both decompositions are useful, but in this primer we will only show an example using the primary decomposition.

4.4.1 The Structure Theorem for Finitely Generated Abelian Groups

Note that an abelian group G can be thought of as a $\mathbb{Z}$-module as follows: the elements of the module are the elements of G, and for every $z \in \mathbb{Z}$ and $g \in G$, we have $z \cdot g = g^z$.

Applying Theorem 4.49 immediately yields that if G is a finitely generated abelian group, there exist unique $n \in \mathbb{N}$, prime numbers $p_1, ..., p_m$ and positive exponents $e_1, ..., e_m$ where

$$G \cong \mathbb{Z}^n \oplus \frac{\mathbb{Z}}{\left(p_1^{e_1}\right)} \oplus ... \oplus \frac{\mathbb{Z}}{\left(p_m^{e_m}\right)}.$$

We may also find a unique invariant decomposition of G, but this is not used as often.

4.4.2 The Jordan Canonical Form

For this section, you ought to know what an invariant subspace and block matrix are.

Consider an algebraically closed field $\mathbb{F}$ and a finite-dimensional vector space V over $\mathbb{F}$. Not every linear map $T : V \to V$ admits a basis of eigenvectors. In other words, there is not always some basis $\mathcal{B}$ where the matrix $\mathcal{M}$ of T with basis $\mathcal{B}$ is diagonal.

The closest we can get is the **Jordan Canonical Form**: there is some basis $\mathcal{B}$ such that $\mathcal{M}$ is of the form

[8] When we say primes p and q are distinct, we precisely mean that there is no unit u where $pu = q$. Equivalently, we mean that $(p) \neq (q)$.

[9] This ensures that we do not unnecessarily pad exponents of 0 to our original decomposition, which is crucial to ensure the d_j we define are not units.

$$\begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_n \end{pmatrix}$$

where each B_i is a block matrix of the form

$$\begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

where the λ_i are the eigenvalues of $\mathcal{M}$ which lie on the main diagonal, and the 1's lie directly above the main diagonal of B_i. We call these **Jordan blocks**. Furthermore, up to permutation of the B_i, the Jordan Canonical Form is unique.

Concretely, an example of a matrix in Jordan Canonical Form is

$$\begin{pmatrix} \lambda_1 & 1 & & & & \\ & \lambda_1 & 1 & & & \\ & & \lambda_1 & & & \\ & & & \lambda_2 & 1 & \\ & & & & \lambda_2 & \\ & & & & & \lambda_3 \end{pmatrix}$$

where λ_1, λ_2, and λ_3 are not necessarily distinct.

To see why such a matrix always exists, we need to make a preliminary observation.

Theorem 4.50: Given any ring R, there is a one-to-one correspondence between the $R[x]$-modules and R-modules M with a linear transformation $T : M \to M$.

Proof of Theorem 4.50: For every M and $T : M \to M$, define M to be an $R[x]$-module where for all $m \in M$, $x \cdot m = T(m)$.

(This is enough to specify how T behaves by the distributive property.) □

I leave it to you to verify this correspondence is indeed bijective.

Now we show the Jordan Canonical Form exists. We apply Theorem 4.49 and note that as an $\mathbb{F}[x]$-module,

$$V \cong \frac{\mathbb{F}[x]}{(x-\lambda_1)^{e_1}} \oplus \dots \frac{\mathbb{F}[x]}{(x-\lambda_m)^{e_m}}.^{10}$$

[10] Here we leverage the algebraicity of $\mathbb{F}$ to conclude that every prime ideal is generated by a linear polynomial. And because V is finite-dimensional, there are no copies of $\mathbb{F}[x]$ in the decomposition, for $\mathbb{F}[x]$ is an infinite-dimensional vector space over V.

By definition, when we apply T to some $v \in V$, we are multiplying each of the direct summands (which are all polynomials) by the polynomial symbol x. This means that each $\frac{\mathbb{F}[x]}{(x-\lambda_i)^{e_i}}$ is an invariant subspace, and thus we may consider the restriction of T to any of these subspaces.

For each λ_i, consider the basis

$$(x - \lambda_i)^{e_i - 1}, ..., (x - \lambda_i)^0 \in \frac{\mathbb{F}[x]}{(x - \lambda_i)^{e_i}}.$$

As $x = \lambda + (x - \lambda)$, we have that for each $0 \leq k < e_i$,

$$\begin{aligned} x \cdot (x - \lambda_i)^k &= \lambda \cdot (x - \lambda_i)^k + (x - \lambda_i) \cdot (x - \lambda_i)^k \\ &= \lambda (x - \lambda_i)^k + (x - \lambda_i)^{k+1}. \end{aligned}$$

Note that when $k = e_i - 1$, we have $(x - \lambda_i)^{k+1} = 0$ as we quotient by $(x - \lambda)_i^e$.

So the matrix of T (restricted to this subspace) with this basis is

$$\begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

which is a Jordan block, exactly as desired.

Combining the Jordan blocks gives us the Jordan Canonical Form. And because the decomposition of V as an $\mathbb{F}[x]$-module is unique, so is the Jordan Canonical Form (up to permutation of the blocks).

Chapter 5

Fields

Rather than studying fields through the lens of linear transformations, we will develop the theory of **field extensions**. A field extension is not complicated: it is just some fields F and K where $F \leq K$. We will be able to study the Galois group $\operatorname{Aut}_F(K)$, which consists of the field automorphisms on K that fix F.

We will be able to connect some group theoretic notions to field theory as well. Certain field extensions K/F are **Galois**, which induces a one-to-one correspondence between the normal extensions H/F where $F \leq H \leq K$ and the normal subgroups of $\operatorname{Aut}_F(K)$. And the solvability of a polynomial $f \in F[x]$ over a radical extension of F is directly related to the solvability of the Galois group of f.[1]

5.1 Basics of Field Extensions

First we define two preliminary notions.

> **Definition 5.1 (Characteristic of a Field)**: The **characteristic** of field F is defined to be the smallest integer p such that $\underbrace{1 + ... + 1 = 0}_{p \text{ copies}}$ if such a p exists, and 0 otherwise.

For our convenience, we denote $\underbrace{1 + ... + 1}_{n \text{ copies}}$ as n_F from now on.

As the letter p suggests, the characteristic of a field is either 0 or prime. For $(mn)_F = m_F n_F$ by the distributive property, and if $(mn)_F = 0$ then one of m_F or n_F must be 0.[2] And by definition, $1_F = 1 \neq 0$.

For example, $\mathbb{Q}$ has characteristic 0 and $\mathbb{F}_p$ has characteristic p. (We denote the field $\mathbb{Z}/p\mathbb{Z}$ as $\mathbb{F}_p$.)

Example 5.2: We are used to infinite fields like $\mathbb{R}$ having characteristic 0 and finite fields like $\mathbb{F}_{p^n}$ having prime characteristic p. These tend to be the well-behaved fields (we will see why when we discuss separability). However, there are also slightly unorthodox infinite fields with prime characteristic p.

For example, the field $\mathbb{F}_p(x)$[3] is infinite and has characteristic p.

[1] To define the Galois Group of a polynomial, we need the notion of a splitting field which we will develop later. But very informally, it is the group of permutations of the roots of f that fix the base field F.

[2] Recall that fields are integral domains.

> **Definition 5.3 (Prime Subfield)**: The **prime subfield** of a field F is the smallest subfield of F. It is also the subfield generated by 1.

The prime subfield of F is isomorphic to $\mathbb{Q}$ when the characteristic of F is 0, and it is isomorphic to $\mathbb{F}_p$ when the characteristic of F is a prime p.

> **Definition 5.4 (Field Extension)**: If a field F is contained by another field K, we say that K is an **extension** of F. We denote this relation by writing "K/F is a field extension".

Somewhat confusingly, the / does not denote a quotient. It merely instantiates fields F and K and further specifies that $F \leq K$.

> **Definition 5.5**: The **degree** of a field extension K/F is the dimension of K as a vector space over F. We denote this as $[K : F]$.

The degrees of field extensions are multiplicative. That is, if $F \leq H \leq K$, we can use any two of $[H : F]$, $[K : H]$, and $[K : F]$ to determine the third.

> **Theorem 5.6**: Given fields $F \leq H \leq K$,
>
> $$[K : F] = [K : H][H : F].$$

Note Theorem 5.6 holds whether the degrees of the extensions are finite or infinite. If the degrees of the extensions are infinite, then Theorem 5.6 is a statement about cardinal arithmetic (a quite trivial one at that).

Proof sketch of Theorem 5.6, finite case: Suppose that $[K : H] = m$ and $[H : F] = n$ for $m, n \in \mathbb{N}$. Further suppose that $\alpha_1, ..., \alpha_m$ is a basis of K over H and $\beta_1, ..., \beta_n$ is a basis of H over F. Then it is enough to check that the set

$$\{\alpha_i \beta_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$$

is a basis of K over F. □

The infinite case is practically identical to the finite case.

5.2 Algebraic and Splitting Field Extensions

[3] This is the field of fractions of the polynomial ring $\mathbb{F}_p[x]$, and its elements are of the form $\frac{f}{g}$ where f and g are polynomials in $\mathbb{F}_p$.

We define the notion of an algebraic extension and use it to develop ideas that will be useful for Galois Theory, such as algebraic closures, separable extensions, and splitting fields of polynomials.

> **Definition 5.7 (Algebraic Element)**: Consider a field extension K/F and an element $a \in K$. We say a is **algebraic** over F if there exists a polynomial $p(x) \in F[x]$ such that a is a root of $p(x)$.

If an element $a \in K$ is algebraic, we may identify it with a unique **minimal polynomial** in $F[x]$.

> **Theorem 5.8 (Minimal Polynomial)**: For algebraic $a \in K$, there is a unique irreducible monic polynomial $m_a(x) \in F[x]$ which has a as a root.
>
> Furthermore, for every $p(x) \in F[x]$ with a as a root, $\deg m_a(x) \leq \deg p(x)$. So $m_a(x)$ is minimal in the sense that no polynomial with smaller degree has a as a root.

To be clear, when I say $m_a(x)$ is irreducible, it is irreducible over F.

Proof of Theorem 5.8: Define $m_a(x)$ to be a polynomial in $F[x]$ with minimal degree.[4] It is utterly trivial to make it monic. If it were reducible in $F[x]$, then there are some polynomials $f, g \in F[x]$ with positive degree such that $fg = m_a$. Because $F[x]$ is an integral domain, we know that $f(a)g(a) = 0$ implies at least one of $f(a)$ or $g(a)$ is 0, contradicting the minimality of $\deg m_a$.

Now we just need to show $m_a(x)$ is unique. Presume not; suppose there is some irreducible monic polynomial $f \in F[x]$ that is distinct from $m_a(x)$. Since $\deg f \geq \deg m_a$ by the minimality of $\deg m_a$, the Euclidean Algorithm says that there are polynomials $g, r \in F[x]$ where

$$m_a g + r = f \text{ and} \deg r < \deg m_a,$$

and since $r(a) = f(a) - m_a(a)g(a) = 0$, this contradicts the minimality of $\deg m_a$. □

Exercise 5.9: Show that in an extension K/F, if $a \in K$ is a root of $f \in F[x]$ then m_a divides f.

There are some interesting results that the study of minimal polynomials yields. We will lay the groundwork to prove these results right now.

Denote $F(a)$ as the subfield of K generated by F and a. (Informally, $F(a)$ is the result of adding a to F, and then adding just enough elements to ensure that $F \cup \{a\}$ remains a field.)

[4] We do not yet know that m_a is unique, so we just pick any such polynomial.

Now consider the homomorphism $\varphi : \frac{F[x]}{(m_a(x))} \to F(a)$ where $\varphi : f(x) \mapsto f(a)$.[5] Because $m_a(x)$ is irreducible (and in a Principal Ideal Domain, irreducibles are primes), the quotient $\frac{F[x]}{(m_a(x))}$ is a field. So $\operatorname{im} \varphi$ is a field, and since φ maps the constant a to itself in $F(a)$, we must have $\operatorname{im} \varphi = F(a)$.

By virtue of the quotient, this map is injective (every polynomial $p \in F[x]$ with a as a root is divisible by $m_a(x)$; prove this yourself). So φ is an isomorphism as desired. In other words,

$$\frac{F[x]}{(m_a(x))} \cong F(a).$$

In particular, this implies

$$[F(a) : F] = \deg m_a(x).$$

This is a fact that will be very useful when we wish to study the intermediate fields of $F(a)$ for some given a. For example, we may wish to study the intermediate fields of $\mathbb{R}(\zeta)$, where ζ is a non-trivial fifth root of unity. We may see that as $[\mathbb{R}(\zeta) : \mathbb{R}] = 5$, there must not be any intermediate fields between $\mathbb{R}$ and $\mathbb{R}(\zeta)$ as 5 is prime. (Recall Theorem 5.6.)

To continue this discussion we will need to first define an **algebraic extension**.

> **Definition 5.10 (Algebraic Extension)**: A field extension K/F is **algebraic** if every $a \in K$ is the root of some polynomial $p(x) \in F[x]$.
>
> In other words, K/F is algebraic if every element in K is algebraic over F.

In this argument we have only used the fact that m_a is irreducible. Indeed, for any irreducible $f \in F[x]$, the quotient $F[x]/(f(x))$ is a field and $x \in F[x]/(f(x))$ is a root of f.

There is an obvious embedding $\varphi : F \to F[x]/(f(x))$, and since $F \cong \varphi(F)$, there is an extension K/F with an isomorphism $K \cong F[x]/(f(x))$ which extends φ. Since f has a root in $F[x]/(f(x))$, it does in K as well.

Suppose for some $a \in K$ that $f(a) = 0$. Then by definition f is the minimal polynomial of a in F, so $\frac{F[x]}{m_a(x)} \cong F(a)$, and $F(a)$ is an algebraic extension of F.

Exercise 5.11: Verify that $F(a)$ is indeed algebraic over F. It is enough to show that if $a, b \in K$ are algebraic, then

- $a + b$ is algebraic;
- ab is algebraic;
- $\frac{1}{a}$ is algebraic.

To do this, suppose $f, g \in F[x]$ where $f(a) = 0$ and $g(a) = 0$. Explicitly construct polynomials with $a + b$, ab, and $\frac{1}{a}$ as a root.

[5] This map is representation invariant because $m_a(x) = 0$, so it is genuinely well-defined.

We summarize this discussion with a theorem.

> **Theorem 5.12**: Suppose K/F is a field extension. If $a \in K$ is algebraic in F, then
>
> $$\frac{F[x]}{(m_a(x))} \cong F(a) \text{ which implies } [F(a) : F] = \deg m_a.$$
>
> Also, given a field F and irreducible $f \in F[x]$, there is an algebraic extension K/F such that
>
> $$\frac{F[x]}{(f(x))} \cong K \text{ which implies } [K : F] = \deg f$$
>
> and f has a root in K.

Algebraic extensions always exist: a very boring example is F/F. (Sometimes it is the only example!) And we just showed that when F contains an irreducible polynomial f with $\deg f \geq 2$, $F[x]/(f(x))$ is isomorphic to a non-trivial algebraic extension K of F. But perhaps more surprisingly, **algebraic closures** always exist too. And we will leverage Theorem 5.12 to show this.

> **Definition 5.13 (Algebraically Closed Field)**: We say F is an **algebraically closed field** if every non-constant polynomial $p \in F[x]$ is a root in F.

A field F being algebraically closed is equivalent to every polynomial $f \in F[x]$ with $\deg f \geq 2$ being reducible. Alternatively no irreducible polynomials of degree ≥ 2 exist.

Exercise 5.14: Suppose F is an algebraically closed field. Show that F/F is the only possible algebraic extension of F.

> **Definition 5.15**: An **algebraic closure** of F is an algebraic extension $\overline{F}/F$ where $\overline{F}$ is algebraically closed.

Usually we identify $\overline{F}$ as the algebraic closure of F rather than the extension $\overline{F}/F$.

> **Theorem 5.16**: Every field F has an algebraic closure, and any two algebraic closures of F are equivalent up to an isomorphism that fixes F.

This will allow us to refer to **the** algebraic closure of F, because it is essentially unique. We prove this via Zorn's Lemma, which says that there is a maximal element in every non-empty partially ordered set where every increasing chain is bounded from above.[6]

Informally, we would like to take the collection of algebraic extensions K/F endowed with the order $\subseteq$ and apply Zorn's Lemma. Once we retrieve a maximal element $\overline{F}$, then we can show that the existence of an irreducible polynomial $f \in F[x]$ with $\deg f \geq 2$ implies the existence of a strict algebraic extension of $\overline{F}$ which contradicts its maximality.

Unfortunately, this collection is a proper class (i.e. it is not a set) for an utterly idiotic reason: there are many isomorphic yet set-theoretically distinct (algebraic) extensions of F. But so long as we can remove these stupid isomorphic copies, this method will suffice. This is why we restrict the algebraic extensions to those in some set S in the proof that follows.

Proof of existence for Theorem 5.16: Construct a set $S \supset F$ where $|S| > \max(|F|, |\mathbb{N}|)$. Define

$$\mathcal{F} = \{K \subset S : K/F \text{ is an algebraic extension}\}$$

and endow it with the relation $\subseteq$ to get the partially ordered set $\{\mathcal{F}, \subseteq\}$. Note that Zorn's Lemma applies because the union of a chain of algebraic extensions is itself an algebraic extension.[7]

So take a maximal element $\overline{F}$ of $\mathcal{F}$. For the sake of contradiction, assume that $\overline{F}$ is not an algebraic closure of F. Then there is a polynomial $f \in \overline{F}[x]$ with no roots and $\overline{F}/(f(x))$ is isomorphic to a proper algebraic extension K of $\overline{F}$. We want to show there is an isomorphism from K to some field in S fixing F. To do this, note

$$|K \setminus \overline{F}| \leq |K| = \max(|F|, |N|) < |S| = |S \setminus \overline{F}|,$$

implying there is an injection $\iota : K \setminus \overline{F} \to S \setminus \overline{F}$. Now consider $\mathrm{id}_{\overline{F}} : \overline{F} \to \overline{F}$ and note $\iota \cup \mathrm{id}_{\overline{F}}$ is an injection fixing $\overline{F}$. Now endow the correct algebraic structure on $\mathrm{im}(\iota \cup \mathrm{id}_{\overline{F}}) \subset S$ and the rest is obvious. □

As the notation in the proof suggests, we will denote the algebraic closure as $\overline{F}$.

Exercise 5.17: Verify that if K/F is algebraic, then $|K| = \max(|F|, |\mathbb{N}|)$. Hint: for each $f \in \mathbb{F}$, define S_f to be the set of roots of f in K. Note that

$$K = \bigcup_{f \in F[x]} S_f$$

as K/F is algebraic. How large is each S_f, and how many $f \in F[x]$ are there?

Exercise 5.18: Show that if K/F is an algebraic extension and $\overline{K}$ is an algebraic closure of K, it is an algebraic closure of F as well. To do this, show that algebraicity is transitive, that is, if H/F and K/H are algebraic extensions, then K/F is algebraic too.

The uniqueness of the algebraic closure is not so obvious, but the proof ideas have all been seen before in the proof of existence.

[6] It is equivalent to the Axiom of Choice under ZF.

[7] To be pedantic, we ought to note $\mathcal{F}$ is non-empty as F/F is algebraic.

Proof of uniqueness for Theorem 5.16: Suppose K and L are both algebraic closures of F. Then define

$$\Phi = \{\varphi : H \to L \mid F \leq H \leq K \text{ and } \varphi \upharpoonright F = \mathrm{id}\}.$$

In other words, Φ is the family of subfield homomorphisms from K to L which fix F.

Now we define a natural partial ordering $(\Phi, \subseteq)$, where $\varphi_1 \subseteq \varphi_2$ if φ_1 is the restriction of φ_2 on a smaller subfield.[8] Note that $(\Phi, \subseteq)$ satisfies the conditions for Zorn's Lemma, because the union of any chain is also a subfield homomorphism from K to L which fixes F.[9]

So there is some maximal element $\varphi : H \to L$. If $H \neq K$, then there is some $a \in K \setminus H$, which means that the field $H(a)$ is a strict superset of H. Now take the minimal polynomial $m_a \in F[x]$ of a, and find some root $b \in L$ of m_a. Then there is an extension of φ with domain $H(a)$ where $\varphi(a) = b$, contradicting the maximality of φ. So we may conclude φ is from K to L.

We show φ is injective by showing its kernel is trivial. Suppose that $\varphi(a) = 0$ for $a \in K$; we want to show that $a = 0$. For every $f \in F[x]$, we have $\varphi(f(a)) = f(\varphi(a))$ as φ is a homomorphism which fixes F. Suppose for the sake of contradiction $a \in K \setminus F$. Then take the minimal polynomial $m_a \in F[x]$ and note

$$0 = \varphi(m_a(a)) = m_a(\varphi(a)) = m_a(0).$$

But 0 being a root of m_a contradicts that m_a is irreducible. So if $\varphi(a) = 0$ we have $a \in F$, and as φ fixes F this implies $a = 0$.

We use a similar argument to show φ is surjective. For any $b \in L$ there is some $f \in F[x]$ with b as a root. Since $\varphi(f(a)) = f(\varphi(a))$ for all $a \in K$, every root $a \in K$ of f corresponds to a root $\varphi(a) \in L$ of f. Because the degree of f doesn't change whether it is interpreted in K of L, the only roots of f in L are of the form $\varphi(a)$ where $a \in K$ is a root of f.

Now as $b \in L$ is a root of f, there is some $a \in K$ (which is a root of f) such that $f(a) = b$. □

We will show a few more results related to Theorem 5.12. First we show a natural converse to the theorem.

> **Theorem 5.19**: Suppose K/F is a field extension and $a \in K$. Then a is algebraic over F if and only if $[F(a) : F]$ is finite.

We have already shown that a algebraic implies $[F(a) : F]$ is finite so we only need to show the other direction.

[8] The symbol $\subseteq$ can be interpreted set theoretically if the functions φ_1 and φ_2 are thought of as sets.

[9] To be pedantic again, we ought to note Φ is non-empty as it contains $\mathrm{id} : F \to L$.

Proof of Theorem 5.19: If $[F(a) : F] = n$ for some $n \in \mathbb{N}$ then the set $\{0, a, ..., a^n\}$ is linearly dependent in F as it has more than n elements. Thus there is a polynomial in $F[x]$ of degree at most n with a as a root. □

Furthermore, in this case there is no polynomial of degree less than n with a as a root, because we know the degree of the minimal polynomial $m_a(x)$ is equal to $[F(a) : F]$.

As a corollary, if K/F is a field extension where $[K : F]$ is finite, then K/F is algebraic. This is because for every $a \in K$, $[F(a) : F]$ is finite.

Next we show that given a polynomial $f \in F[x]$, there is a unique (up to isomorphism) minimal extension K/F where f factors completely in K. In this case we say that K is a **splitting field extension** of F over f. Furthermore, if f factors completely in K (i.e. into linear factors), we will say that f **splits** over K.

When we say K is minimal, we precisely mean that for any $F \leq H < K$, f does not split over H. Note this is a trivial statement if f splits over F.

> **Theorem 5.20 (Splitting Fields Exist)**: Suppose $f \in F[x]$. Then there is a splitting field extension K/F for f. Furthermore, it is unique up to isomorphism.

The proof is a trivial consequence of the existence of an algebraic closure.

Proof of Theorem 5.20: Let $\overline{F}$ be the closure of F and define

$$\mathcal{F} = \left\{K \leq \overline{F} \mid f \text{ splits over } K\right\}.$$

Note f splits over $\bigcap \mathcal{F}$ and it is obviously minimal.

Uniqueness is easy too. Suppose K/F and L/F are splitting field extensions. Then $\overline{K}$ and $\overline{L}$ are closures of F and there is an isomorphism $\varphi : \overline{K} \to \overline{L}$ fixing F. Field isomorphisms preserve subfield structure and φ preserves algebraic properties of F as it fixes F.

Concretely, this means that if f splits over $H \leq K$, it splits over $\varphi(H) \leq L$ as well. Because φ is a bijection, the converse holds as well. And since K is a splitting field, it is the intersection of all the subfields of $\overline{K}$ which f splits over. Likewise for L. Thus

$$K = \bigcap\left\{H \leq \overline{K} \mid f \text{ splits over } H\right\} \cong \bigcap\left\{H \leq \overline{L} \mid f \text{ splits over } H\right\} = L.$$

□

Exercise 5.21: Use the exact same reasoning to show that if f splits over K/F then there is a unique $F \leq H \leq K$ such that H/F is a splitting field extension.

Exercise 5.22: Show that if K/F is a splitting field extension for a polynomial $f \in F[x]$, then $[K : F] \mid (\deg f)!$ Hint: by Theorem 5.6, it is enough to construct some extension L/F which f splits over where $[L : F] \mid (\deg f)!$[10]

We may define the splitting field of a family of polynomials in a similar way.

Definition 5.23 (Splitting Field Extension): A **splitting field extension** of a family of polynomials $\{f_i \mid i \in I\} \subset F[x]$ is a minimal extension K/F where every f_i splits over K.

Such an extension exists and is unique for the exact same reasons as Theorem 5.20. Similarly Exercise 5.21 still holds as well.

There is actually an algorithm to produce a splitting field extension for a finite family of polynomials. We construct a series of extensions $F = K_0 \leq K_1 \leq ... \leq K_n$ as follows:

- If the polynomials split in K_i, we are done.
- If an extension K_{i+1}/K_i splits over every factor of $f \in K_i[x]$, then it splits over f. Therefore, we factor each polynomial in the family into its irreducible components.
- Now we have a family of irreducibles $f_1, ..., f_n \in K_i[x]$ which we want to split in an extension K_{i+1}/K_i.
- Select some f and take $K_{i+1} \cong \frac{K_i[x]}{(f(x))}$.

Denote our final extension K_n as K. This is the obvious way to construct an extension K/F where f splits over K. But is it minimal? Yes.

For $S \subset K$, we denote $F(S)$ to be the subfield of K generated by $F \cup S$. Note that $F(a, b) = (F(a))(b)$ for $a, b \in K$. And since $K_{i+1} = K_i(a_i)$ for some $a_i \in K$, we may conclude that

$$K = F(a_1, ..., a_n).$$

This means K/F is a minimal extension, as the smallest subfield of K containing F and the roots $a_1, ..., a_n$ of the family of polynomials is $F(a_1, ..., a_n)$.

5.3 Separable Extensions

For a field F we develop the notion of a **separable polynomial**, and we develop the notion of **separable extensions** K/F in which every minimal polynomial is separable. Separability is one of the two conditions we will use to characterize Galois extensions, the other being normality.

Definition 5.24 (Separable Polynomial): A polynomial $f \in F[x]$ is **separable** if it has no multiple roots in a splitting field extension K/F.

Because splitting fields are isomorphic, it does not matter which splitting field extension we take.

[10] It turns out that the obvious extension you need to construct is a splitting field extension; we see this very soon. But for the purposes of this exercise, it is easier to not worry about the minimality of this extension.

Note that f having no repeated roots in K is equivalent to f having no repeated roots in $\overline{F}$. Extending K to $\overline{F}$ has no effect on the roots of f, but it makes future analysis easier to perform.

In high school calculus you likely learned that a polynomial (in $\mathbb{C}[x]$, perhaps) has multiple roots if and only if its derivative is 0. A similar result holds for separable polynomials.

Definition 5.25 (Formal Derivative): The **formal derivative** of a polynomial $a_n x^n + ... + a_0 \in F[x]$ is $n_F a_n x^{n-1} + ... + a_1$.

Alternatively, the **formal derivative** is the unique linear map $-' : F[x] \to F[x]$ where $1' = 0$ and $(x^n)' = n_F x^{n-1}$ for every $n \geq 1$.

In short, the formal derivative is exactly what you'd expect from the Power Rule.

Exercise 5.26: Verify the formal derivative satisfies the Product Rule. That is, for $f, g \in F[x]$, $(fg)' = fg' + f'g$.

Theorem 5.27: A polynomial $f \in F[x]$ is separable if and only if is relatively prime to f'.

Just to be clear, two polynomials are relatively prime if they do not share any common non-constant factors. And two polynomials $f, g \in F[x]$ are relatively prime if and only if f and g have no common roots in $\overline{F}$. For any common factor of f and g in F must have the same roots in $\overline{F}$, and if $a \in \overline{F}$ is a root of f and g, then m_a must divide f and g.

First we prove a trivial preliminary lemma.

Theorem 5.28: For $a \in F$ and $f \in F[x]$,

$$(x-a)^2 \mid f(x) \Longleftrightarrow f(a) = f'(a) = 0.$$

Proof of Theorem 5.28: If $(x-a)^2 \mid f(x)$ then $f'(a) = 0$ by the Product Rule.

If $f(a) = f'(a) = 0$ then there is some $g \in F[x]$ such that $f(x) = (x-a)g(x)$. By the Product Rule, $f'(x) = g(x) + (x-a)g'(x)$, so

$$0 = f'(a) = g(a) + (a-a)g'(a) = g(a).$$

□

Now the proof of the main theorem is trivial.

Proof of Theorem 5.27: We have established that

$$\begin{aligned} f \text{ and } f' \text{ are relatively prime} &\iff \text{there is no } a \in \overline{F} \text{ with } f(a) = f'(a) = 0 \\ &\iff \text{there is no } a \in \overline{F} \text{ with } (x-a)^2 \mid f(x) \\ &\iff f \text{ is separable.} \end{aligned}$$

□

As a corollary, there is a much simpler characterization of irreducible polynomials with the derivative.

> **Theorem 5.29**: If $f \in F[x]$ is irreducible, then f is separable if and only if $f' \neq 0$.

Proof of Theorem 5.29: If $f' = 0$ then obviously f and f' have shared roots in $\overline{F}$.

And $f' \neq 0$ implies that f' and f are relatively prime, as the irreducibility of f implies that its only non-constant factor in $F[x]$ is f. (Obviously $f' \neq f$.) □

Finally we define a separable extension.

> **Definition 5.30 (Separable Extension)**: An algebraic extension K/F is separable if for every $a \in K$, $m_a \in F[x]$ is separable.

Inseparable algebraic extensions are somewhat uncommon. For every irreducible polynomial in a field F with characteristic 0 has non-zero derivative, meaning that every extension K/F is separable.

Showing that every extension $\mathbb{F}_{p^n}/F$ of a finite field F is separable requires a little more ingenuity, but it is a short argument. Note $a^{p^n} - a = 0$ for every $a \in \mathbb{F}_{p^n}$, and the polynomial $x^{p^n} - x$ is separable as it has derivative -1, so the minimal polynomial of a is separable too.

Example 5.31: The extension $\mathbb{F}_p(x)/\mathbb{F}_p(x^p)$ is not separable because the minimal polynomial of $x \in \mathbb{F}_p(x)$ in $\mathbb{F}_p(x^p)$ is $f(t) = t^p - x^p$.[11] This polynomial is not separable because it has derivative 0, so the extension is not separable.

5.4 Normal Extensions, i.e. more on Splitting Field Extensions

In this section we define **normal extensions** and characterize several equivalent conditions for an extension to be normal.

Normal field extensions are indispensible for Galois Theory. In fact, since separability is a trivial consequence of algebraicity in fields of characteristic 0 and finite fields (the two types

[11] This is the minimal polynomial because $[\mathbb{F}_p(x) : \mathbb{F}_p(x^p)] = p$ and $\mathbb{F}_p(x) = \mathbb{F}_p(x^p)(x)$. To spell it out a little more, use Theorem 5.12 to conclude the minimal polynomial of x has degree p, and clearly $t^p - x^p$ has degree p. And if it was reducible, there would be a polynomial of smaller degree which p is a root of, contradiction.

of fields we are most interested in studying), normality is the only condition that needs to be checked for a potential Galois extension.

> **Definition 5.32**: An algebraic extension K/F is **normal** if there is a family of polynomials $\{f_i \mid i \in I\}$ such that K/F is a splitting field extension for $\{f_i \mid i \in I\}$.

Just to be extremely clear, "normal extension" is a synonym for "splitting field extension". There are two reasons we use the term "normal":

1. It is shorter to say "the extension K/F is normal" than "the extension K/F is a splitting field extension". Having a one-word adjective to describe an extension is much more concise than having a three-word adjective.
2. Far more importantly, it will turn out that in a Galois (i.e. normal and separable) extension K/F, the normal subgroups $G \leq \operatorname{Aut}_F(K)$ directly correspond to the normal extensions H/F where $F \leq H \leq K$.

It is easy to verify that an extension is normal: just find the family $\{f_i \mid i \in I\}$. But it is not immediately obvious how to show an extension is not normal, or even to generally determine whether an extension is normal. Therefore we establish some conditions for normality.

> **Theorem 5.33**: Consider an algebraic extension K/F. The following are equivalent:
>
> 1. K/F is normal.
> 2. For every $a \in K$, its minimal polynomial $m_a \in F[x]$ splits over K.
> 3. Every irreducible polynomial $f \in F[x]$ with a root in K splits over K.
> 4. For any field L and homomorphisms $\varphi, \psi : K \to L$ where $\varphi \restriction F = \psi \restriction F$, we have $\varphi(K) = \psi(K)$.
>
> Furthermore if $[K : F]$ is finite then another equivalent condition is that K/F is a splitting field extension for a single polynomial $f \in F[x]$.

To prove $(4) \implies (3)$ we must first establish a technical lemma.

> **Theorem 5.34**: Suppose $\varphi : F_1 \to F_2$ is a field isomorphism and that K_1/F_1 and K_2/F_2 are field extensions. For every $a \in K_1$ and $b \in K_2$, there is a unique extension[12] $\varphi_a : F_1(a) \to K_2$ of φ with $\varphi_a(a) = b$ if and only if b is a root of $\varphi(m_a) \in F_2[x]$.[13]

Proof of Theorem 5.34: For ease of reading define $\varphi(m_a) = f \in F_2[x]$.

[12] Precisely, $\varphi_a \restriction F_1 = \varphi$.

[13] As usual we define $m_a \in F_1[x]$ to be the minimal polynomial of $a \in K_1$.

If φ_a exists then $0 = \varphi_a(m_a(a)) = f(\varphi_a(a)) = f(b)$.

If b is a root of f then f is the minimal polynomial of b as it is irreducible.[14] Thus

$$F_1(a) \cong \frac{F_1[x]}{(m_a(x))} \cong \frac{F_2[x]}{(f(x))} \cong F_2(b) \leq K$$

and in particular, there is a natural isomorphism from $F_1[x]/(m_a(x))$ to $F_2[x]/(m_a(x))$ which extends φ. Compose the isomorphisms together to get the desired φ_a.

The uniqueness of φ_a is obvious: its behavior on $F_1 \cup \{a\}$ is fixed and this set generates $F_1(a)$. □

Proof of Theorem 5.33: First we make no presumptions on $[K : L]$ and show $(2) \Longrightarrow (1) \Longrightarrow (4) \Longrightarrow (3) \Longrightarrow (2)$.

- $(2) \Longrightarrow (1)$: Clearly K splits over the family $\{m_a \mid a \in K\}$. And the extension K/F is minimal because for each $a \in K$, m_a splitting in K guarantees a is in the minimal extension.
- $(1) \Longrightarrow (4)$: Field homomorphisms are either injective or trivial; we ignore the case where φ and ψ are trivial. Since we may embed K into L, for convenience just take $L \supseteq K$.

  Suppose K/F is a splitting field for the family $\mathcal{F} \subseteq F[x]$. Then $\varphi(F) = \psi(F) = F$, implying φ and ψ fix every polynomial in $\mathcal{F}$. Thus $\varphi(K)/F$ is a splitting field extension for $\mathcal{F}$. So is $\psi(K)/F$ (for $\mathcal{F}$). By Exercise 5.21 we have that $\varphi(K) = \psi(K)$.
- $(4) \Longrightarrow (3)$: Take any irreducible $f \in F[x]$ with root $a \in K$. (Note that f is the minimal polynomial of a.) We want to show that every root $b \in \overline{K}$ of f is contained in K.

  By Theorem 5.34 there exists $\varphi : F(a) \to \overline{K}$ such that $\varphi(a) = b$ and $\varphi \upharpoonright F = \mathrm{id}$. (To be clear, we are extending $\mathrm{id}_F : F \to F$.) Since the embedding $\mathrm{id} : F(a) \to \overline{K}$ agrees with φ on F, we have that $\varphi(F(a)) = F(a)$. This means that $b \in F(a) \subseteq K$.
- $(3) \Longrightarrow (2)$: Every minimal polynomial is irreducible.

Now suppose $[K : L]$ is finite. If K/F is a splitting field extension for f then K/F is normal by definition. And if K/F is a splitting field extension for $f_1, ..., f_n$, then it is a splitting field extension for their product $f_1 ... f_n$. □

5.5 Galois Theory

We have laid all of the necessary groundwork to begin exploring Galois theory. For each extension K/F, we may consider the **Galois group** $\mathrm{Aut}_F(K)$, the group of field automorphisms $\varphi : K \to K$ that fix F. And for a special type of field extensions — the

[14] Since $\varphi : F_1 \to F_2$ is an isomorphism, algebraic properties of polynomials are preserved.

Galois extensions — there is a correspondence between the intermediate fields $F \leq H \leq K$ and subgroups $G \leq \operatorname{Aut}_F(K)$.

We may apply these results to solve classical questions, such as “can we trisect an angle?” and “can we determine whether a polynomial is solvable?” Furthermore we can use it to show that $\mathbb{C}$ is the algebraic closure of $\mathbb{R}$, where we define $\mathbb{C} = \mathbb{R}[x]/(x^2+1)$. But all this in due time.

> **Definition 5.35 (Galois Extension)**: An algebraic extension is **Galois** if it is normal and separable.

We will now proceed to show that several characterizations of Galois extensions are equivalent. Really, I should be showing you why Galois extensions are interesting first. But this characterizations is necessary to show the interesting results, which is why we must do it first.

> **Theorem 5.36**: For a field extension K/F with finite $[K : F]$, the following are equivalent:
>
> 1. K/F is Galois.
> 2. K/F is a splitting field extension for some $f \in F[x]$ whose irreducible factors are all separable.
> 3. $|\operatorname{Aut}_F(K)| = [K : F]$.
> 4. $F = \operatorname{Fix}(\operatorname{Aut}_F(K))$.

Remember from group theory that $\operatorname{Fix}(\operatorname{Aut}_F(K))$ is defined as the subset of K fixed under every automorphism in $\operatorname{Aut}_F(K)$. By definition $F \subseteq \operatorname{Fix}(\operatorname{Aut}_F(K))$, but the other direction is the non-trivial part in this theorem.

Furthermore recall that $[K : F]$ being finite implies that K/F is algebraic.

Exercise 5.37: Verify that $\operatorname{Fix}(\operatorname{Aut}_F(K))$ is a subfield of K.

Proof of Theorem 5.36: We show that $(1) \Longrightarrow (2) \Longrightarrow (3) \Longrightarrow (4) \Longrightarrow (1)$.

- $(1) \Longrightarrow (2)$: Since $[K : F]$ is finite, there are some $a_1, ..., a_n \in K$ such that $K = F(a_1, ..., a_n)$. Now take $f = \prod_{i=1}^{n} m_{a_i} \in F[x]$.

  Because K/F is normal (i.e. a splitting field extension), each of the m_{a_i} splits over K, implying f splits over K. Minimality of this extension is obvious, so it is a normal extension.

  And the irreducible factors of f are the m_{a_i}. They are obviously separable.
- $(2) \Longrightarrow (3)$: Suppose K/F is a splitting field extension for $f \in F[x]$ whose irrreducible coefficients are all separable. Let $a_1, ..., a_n$ be the roots of f and note $K = F(a_1, ..., a_n)$.

Define $K_0 = F$ and $K_i = F(a_1, ..., a_i)$ for $1 \leq i \leq n$. Let $m_i \in K_{i-1}[x]$ be the minimal polynomial of a_i in K_{i-1}. Note by Theorem 5.6 and Theorem 5.12 that

$$\begin{aligned} [K : F] &= [K_n : K_{n-1}]...[K_1 : F] \\ &= \deg m_n ... \deg m_1. \end{aligned}$$

By Exercise 5.9 m_i is separable as $m_i \mid f$. **It is genuinely important that all the roots are distinct.**

Now note that an automorphism $\varphi : K \to K$ that fixes F may uniquely be determined by $\varphi(a_i)$ for $1 \leq i \leq n$. We claim that because of Theorem 5.34,

1. every φ where $\varphi(a_i)$ is a root of m_i may be extended to a unique automorphism on K fixing F,
2. and for every automorphism $\varphi : K \to K$ fixing F, we must have $\varphi(a_i)$ be a root of m_i.

Since each m_i has $\deg m_i$ unique roots[15], this would imply that $\mathrm{Aut}_F(K) = \deg m_n ... \deg m_1$. Now it just remains to prove these two claims are true.

1. Repeatedly apply Theorem 5.34 to get unique $\varphi_i : K_i \to K$ which culminates with $\varphi = \varphi_n : K_n \to K$. By Theorem 5.33, since φ agrees with $\mathrm{id} : K \to K$ on F, φ is surjective. Interpreting φ as a linear map on K as a F-vector space, we see that F must also be injective by Rank-Nullity. So the φ we generate is indeed an automorphism.
2. If φ is a homomorphism, $\varphi_i = \varphi \upharpoonright K_i$ better be as well. And by Theorem 5.34, φ_i may only exist if $\varphi_i(a_i)$ is a root of m_i.

- (3) $\Longrightarrow$ (4): Because $[K : F]$ is finite, there are $a_1, ..., a_n \in K$ such that $K = F(a_1, ..., a_n)$. Define K_i and m_i as before and let r_i be the number of distinct roots of each m_i in K.

  By similar reasoning as before, we have that the number of **homomorphisms** in $\varphi : K \to K$ fixing F is $r_1...r_n$. Since $r_i \leq \deg m_i$ we have $\mathrm{Aut}_F(K) = r_1...r_n \leq \deg m_1 ... \deg m_n = [K : F]$.

  Obviously $\mathrm{Aut}_F(K) = \mathrm{Aut}_{\mathrm{Fix}(F)}(K)$ so $\mathrm{Aut}_F(K) \leq [K : \mathrm{Fix}(F)]$ implying that $F = \mathrm{Fix}(F)$.
- (4) $\Longrightarrow$ (1): We prove that every irreducible $f \in F[x]$ with a root in K is separable and splits over K. Since $\mathrm{Aut}_F(K)$ consists of automorphisms, we may define an equivalence relation on the roots of f (in K) as follows:

  $$a \sim b \Longleftrightarrow \text{there exists } \varphi \in \mathrm{Aut}_F(K) \text{ where } \varphi(a) = b$$

  Take some equivalence class R of the roots and let $f = gh$ where h has no roots in R.

[15] This is where separability comes in play!

Note every $\varphi \in \mathrm{Aut}_F(K)$ permutes the roots of f because $\varphi(f) = f$, meaning that $a \in K$ is a root of f if and only if $\varphi(a)$ is a root of $\varphi(f) = f$. Furthermore, φ permutes the roots in R because φ is injective and cannot send roots in R outside of R.

The multiplicity of every root in R must be the same since given some $\varphi \in \mathrm{Aut}_F(K)$, if $\varphi(a) = b$ for $a, b \in R$, then the multiplicity of b in $\varphi(f) = f$ is the multiplicity of a in f. And by definition of R, there is some φ such that $\varphi(a) = b$ for every pair $a, b \in R$.

This means that φ fixes g, meaning that φ must fix every coefficient of g. And the only way for this to happen for **every** $\varphi \in \mathrm{Aut}_F(K)$ is for the coefficients of g to be in F, meaning $g \in F[x]$.

This means g must be equal to (a scalar multiple of) f as f is irreducible. Notable, every root of f must be in R, and f must split over K.

Now suppose the roots of f are $r_1 ... r_n$. Because φ permutes the roots of f, $h = r_1 ... r_n$ is fixed under φ, meaning its coefficients are in F. Since h divides f and f is irreducible, h must be equal to (a scalar multiple of) f. Thus f is separable.

□

In the proof we have shown that $|\mathrm{Aut}_F(K)| \leq [K : F]$ for **every** extension K/F where $[K : F]$ is finite, regardless of whether it is Galois or not. Keep this in mind because it will be important.

This may look like a lot, but the theorem is actually quite natural. The proof of $(1) \Longrightarrow (2)$ is completely trivial, the proof of $(2) \Longrightarrow (3)$ is the only possible argument that could work when using Theorem 5.12 and Theorem 5.34, and the rest of the proofs are just riffs on $(2) \Longrightarrow (3)$. None of these ideas are particularly tricky if you have a good understand of the theory of field extensions we developed earlier.

Now we can finally establish some interesting connections between field theory and group theory.

Theorem 5.38 (Galois Correspondence): Suppose K/F is a field extension with $[K : F]$ finite. We describe the **Galois correspondence** as follows:

1. There is a map $G \mapsto \text{Fix }(G)$ where $G \leq \text{Aut}_F(K)$ and $\text{Fix }(G)$ is the fixed field of K under G.
2. There is a map $H \mapsto \text{Aut}_H(K)$ where $F \leq H \leq K$ and $\text{Aut}_H(K)$ is the group of automorphisms $\varphi : K \to K$ fixing H.

Then K/F is Galois if and only if the maps are two-sided inverses of each other, which implies both maps are bijective.

Furthermore, the correspondence has several more interesting properties:

1. It is inclusion-reversing.
   a. If $G_1 \leq G_2 \leq \text{Aut}_F(K)$ then $\text{Fix }(G_1) \geq \text{Fix }(G_2)$.
   b. If $F \leq H_1 \leq H_2 \leq K$ then $\text{Aut}_{H_1}(K) \geq \text{Aut}_{H_2}(K)$.
2. **If K/F is Galois**, then the size of an automorphism subgroup is equal to the degree of the corresponding intermediate extension. If $G \leq \text{Aut}_F(K)$ then $|G| = [K : \text{Fix }(G)]$ and $[\text{Fix }(G) : F] = |\text{Aut}_F(K)|/|G|$.
3. **If K/F is Galois**, then for all $F \leq H \leq K$, K/H is Galois.
4. **If K/F is Galois**, then normality is preserved. For all $F \leq H \leq K$, H/F is normal if and only if $\text{Aut}_H(K) \trianglelefteq \text{Aut}_F(K)$. Furthermore, when this is the case, $\text{Aut}_F(K)/\text{Aut}_H(K) \cong \text{Aut}_F(H)$.

We extract the difficult part into a technical lemma.

Theorem 5.39: Suppose K is a field and G is a finite subgroup of $\text{Aut}(K)$. Then $|G| = [K : \text{Fix}(G)]$ and $G = \text{Aut}_{\text{Fix}(G)}(K)$.

Note this implies that $K/\ \text{Fix}(G)$ is Galois.

Proof of Theorem 5.39: Since $\text{Aut}_{\text{Fix}(G)}(K) \geq G$ (as every automorphism in G fixes $\text{Fix }G$ by definition), we know that

$$[K : \text{Fix}(G)] \geq \left|\text{Aut}_{\text{Fix}(G)}(K)\right| \geq |G|.$$

If we show that $[K : \text{Fix}(G) = |G|$, then this chain of inequalities collapses. In particular, this would imply that $\text{Aut}_{\text{Fix}(G)}(K) = G$ as both groups are finite.

Assume for the sake of contradiction that $[K : \text{Fix}(G)] > |G|$. Enumerate the elements of G as $\varphi_1, ..., \varphi_n$ where $\varphi_1 = \text{id}$, select some $a_1, ..., a_{n+1} \in K$ that are independent over $\text{Fix}(G)$ and define vectors $v_1, ..., v_{n+1} \in K^n$ where

$$v_i = \begin{pmatrix} \varphi_1(a_i) \\ \vdots \\ \varphi_n(a_i) \end{pmatrix}$$

for each $1 \leq i \leq n + 1$.

Obviously these vectors are dependent over K^n as a vector space over K. So let m be the minimal size of a dependent subset. Without loss of generality, suppose $v_1, ..., v_m$ are dependent. Then there are some $\lambda_1, ..., \lambda_m \in K$ such that

$$\sum_{i=1}^{m} \lambda_i v_i = 0.$$

Since m is minimal, none of the λ_i are 0. Thus we may multiply by a scalar to get $\lambda_1 = 1$.

Coordinate matching yields

$$\sum_{i=1}^{m} \lambda_i \varphi_j(a_i) = 0$$

for each $1 \leq j \leq n$. In particular, as $\varphi_1 = \mathrm{id}$, we have

$$\sum_{i=1}^{m} \lambda_i a_i = 0.$$

Because the a_i are independent over $\mathrm{Fix}(G)$, there must be some $1 \leq i \leq m$ where $\lambda_i \notin \mathrm{Fix}(G)$. Without loss of generality suppose that $\lambda_m \notin \mathrm{Fix}(G)$ and that φ_n is the witness to $\lambda_m \notin \mathrm{Fix}(G)$, i.e. $\varphi_n(\lambda_m) \neq \lambda_m$. (We must have $1 = \lambda_1 \in \mathrm{Fix}(G)$ as every automorphism fixes 1, so this truly is without loss of generality.)

Applying φ_n to each of the coordinates of $\sum_{i=1}^{m} \lambda_i v_i$, we see that

$$\sum_{i=1}^{m} \varphi_n(\lambda_i) \varphi_n \varphi_j(a_i) = 0$$

for all $1 \leq j \leq n$. But since G is a group, there is some j such that $\varphi_n \varphi_j = \mathrm{id}$, and considering this j yields that

$$\sum_{i=1}^{m} \varphi_n(\lambda_i)(a_i) = 0.$$

This implies that

$$\sum_{i=1}^{m} (\varphi_n(\lambda_i) - \lambda_i)(a_i) = 0,$$

and in particular, because $\lambda_1 = 1$, we must have $\varphi_n(\lambda_1) = \lambda_1$. Thus the coefficient of a_i in this linear combination is 0. And as $\varphi_n(\lambda_m) - \lambda_m \neq 0$ (recall by definition $\varphi_n(\lambda_m) \neq \lambda_m$), this contradicts the minimality of m. □

Now we prove the main theorem.

Proof of Theorem 5.38: The first property of the correspondence is obvious and we take it for granted from here on out.

We first show that K/F is Galois if the Galois correspondences are inverses. If the Galois correspondences are inverse to each other, then there must exist some unique G such that $\mathrm{Fix}_G(K) = F$. Because the correspondence is inclusion-reversing, we must have G be maximal, i.e. $G = \mathrm{Aut}(K/F)$.

From here on out, we presume K/F is Galois. We will be proving the properties of the correspondence out of order.

- Note that for all $F \leq H \leq K$, since K/F is a splitting field extension for separable $f \in F[x]$, H/F is a splitting field extension for separable $f \in H[x]$. Thus K/H is Galois, proving (3).
- Since H/K is Galois, we know that $H = \mathrm{Fix}(\mathrm{Aut}_H(K))$, so
$$(H \mapsto \mathrm{Aut}_H(K)) \circ (G \mapsto \mathrm{Fix}(G)) = \mathrm{id}\,.$$
And by Theorem 5.39 we know for each $G \leq \mathrm{Aut}_F(K)$, $G = \mathrm{Aut}_{\mathrm{Fix}(G)}(K)$. Thus
$$(G \mapsto \mathrm{Fix}(G)) \circ (H \mapsto \mathrm{Aut}_H(K)) = \mathrm{id}$$
as well, which shows the two Galois correspondences are two-sided inverses of each other.
- From Theorem 5.39 we have $[K : \mathrm{Fix}(G)] = |G|$ and by Theorem 5.6 we have
$$[\mathrm{Fix}(G) : F] = [K : F]/[K : \mathrm{Fix}(G)] = |\mathrm{Aut}_F(K)|/|G|,$$
proving (2).
- Suppose H/F is normal. Obviously it is separable, so it is Galois. We aim to show that for any $\varphi \in \mathrm{Aut}_F(K)$, $\varphi \restriction H \in \mathrm{Aut}_F(H)$. This will induce a natural map $\Phi : \mathrm{Aut}_F(K) \to \mathrm{Aut}_F(H)$ where $\mathrm{Aut}_F(K) : \varphi \mapsto \varphi \restriction H$, and applying the First Isomorphism Theorem yields
$$\mathrm{im}\,\Phi \cong \frac{\mathrm{Aut}_F(K)}{\ker \Phi} = \frac{\mathrm{Aut}_F(K)}{\mathrm{Aut}_H(K)}.$$
Since this quotient is well-defined, $\mathrm{Aut}_H(K) \trianglelefteq \mathrm{Aut}_F(K)$.

  To show that $\mathrm{im}\,\Phi = \mathrm{Aut}_F(H)$, note by Theorem 5.36 and Theorem 5.6 that

$$\begin{aligned}|\operatorname{im}\Phi| &\le |\operatorname{Aut}_F(H)| \\ &= [H : F] \\ &= \frac{[K : F]}{[K : H]} \\ &= \frac{|\operatorname{Aut}_F(K)|}{|\operatorname{Aut}_H(K)|} \\ &= \left|\frac{\operatorname{Aut}_F(K)}{\operatorname{Aut}_H(K)}\right|.\end{aligned}$$

Since equality holds, we must have $\operatorname{im}\Phi = \operatorname{Aut}_F(H)$ as desired.

Now we just need to check that $\varphi \restriction H \in \operatorname{Aut}_F(H)$ for every $\varphi \in \operatorname{Aut}_F(K)$. For every $a \in H$, the minimal polynomial $m_a \in F[x]$ splits over H as H/F is normal. Since φ permutes the roots of m_a in L, we may deduce that

$$m_a(a) = 0 \Longrightarrow m_a(\varphi(a)) = 0$$

and since every root of m_a lies in H, we know $\varphi(a) \in H$ as well. So $\varphi \restriction H$ may be considered as a map to H, and since $\varphi \restriction H$ is injective, viewing it as a map on H as a vector space over F and applying Rank Nullity allows us to deduce it is surjective as well.

This proves one direction of (4).

- Suppose that $\operatorname{Aut}_H(K) \trianglelefteq \operatorname{Aut}_F(K)$. It is enough to show H/F is normal as separability comes for free. Suppose $a \in H$ and $m_a \in F[x]$ is the minimal polynomial of a. We know it splits over K as K/F is normal, so it suffices to show for every root $b \in K$ of m_a, we also have $b \in H = \operatorname{Fix}(\operatorname{Aut}_H(K))$.

  We showed in the proof of $(4) \Longrightarrow (1)$ for Theorem 5.36 that there is some $\varphi \in \operatorname{Aut}_F(K)$ where $\varphi(a) = b$. Now for every $\psi \in \operatorname{Aut}_H(K)$ we have $\varphi\psi\varphi^{-1} = \psi$ as $\operatorname{Aut}_H(K) \trianglelefteq \operatorname{Aut}_F(K)$. In particular,

$$\begin{aligned}\psi(b) &= \varphi\psi\varphi^{-1}(b) \\ &= \varphi\psi(a) \\ &= \varphi(a) \\ &= b,\end{aligned}$$

  so $b \in \operatorname{Fix}(\operatorname{Aut}_H(K))$ as desired.

□

5.5.1 Fundamental Theorem of Algebra

To show $\mathbb{C}$ is algebraically closed, it suffices to show that every polynomial $f \in \mathbb{C}[x]$ splits over $\mathbb{C}$. Actually, it suffices to show that polynomials in $\mathbb{R}[x]$ split over $\mathbb{C}$, because if we know $f\overline{f} \in \mathbb{R}[x]$ splits over $\mathbb{C}$, we must have that f splits in $\mathbb{C}$.

Here is how we will proceed. If we show that the degree of every splitting field extension $F/\mathbb{R}$ for $f \in \mathbb{R}[x]$ is 1 or 2, then we are home free. (To state the obvious, $F/\mathbb{R}$ is Galois; we will use this to great effect later.)

1. If $[F : \mathbb{R}] = 1$, then obviously $F = \mathbb{R}$ and thus f splits over $\mathbb{R}$.
2. Any extension $F/\mathbb{R}$ can be written in the form $\mathbb{R}(a)$, and as $\mathbb{C} = \mathbb{R}(i)$, there is an obvious automorphism from $\mathbb{C}$ to $\mathbb{R}(a)$ where $i \mapsto a$. **To be very explicit, any two extensions of $\mathbb{R}$ with degree 2 are isomorphic.**

We now show that if an extension $F/\mathbb{R}$ is non-trivial, it must have even degree. **Note that we do not stipulate the extension is normal here.** Suppose $a \in F \setminus \mathbb{R}$, then $\mathbb{R}(a)/F$ is an intermediate extension whose degree we will show is even. Note $[\mathbb{R}(a) : \mathbb{R}] = \deg m_a$ where $m_a \in F[x]$ is the minimal polynomial of a. The degree of m_a is even because every odd degree polynomial in $\mathbb{R}[x]$ has a root in $\mathbb{R}$ and thus is not irreducible. Applying Theorem 5.6 yields that $[F : \mathbb{R}]$ is even as well.

So the degree of any splitting field extension $F/\mathbb{R}$ must be a power of 2. Because if G is a Sylow-2 subgroup of $\mathrm{Aut}_{\mathbb{R}}(F)$, we may conclude that $[\mathrm{Fix}(G) : \mathbb{R}]$ is an extension with odd order, implying it has order 1. Since $\mathrm{Fix}(G) = \mathbb{R}$, we may conclude that $G = \mathrm{Aut}_{\mathbb{R}}(F)$.

Suppose for the sake of contradiction that the degree of $F/\mathbb{R}$ is 2^n for $n \geq 2$. Then Sylow's Theorem says there is $H \leq K \leq \mathrm{Aut}_{\mathbb{R}}(F)$ where $|H| = 2^{n-2}$ and $|K| = 2^{n-1}$. And the Galois correspondence implies that $[\mathrm{Fix}(H) : \mathrm{Fix}(K)] = 2$ and $[\mathrm{Fix}(K) : \mathbb{R}] = 2$. We know that $\mathrm{Fix}(K) \cong \mathbb{C}$, so all that is left to do is show there is no extension of $\mathbb{C}$ with degree 2.

If there was one, then it would be of the form $\mathbb{C}(a)$, implying that $\deg m_a = 2$. But every quadratic in $\mathbb{C}$ has roots and is thus reducible, contradiction.

5.5.2 Solvability over Radical Extensions

It is well-known that there is a formula for finding the roots of a polynomial with degree at most 4. Our goal is to show there is no general formula for degree 5 polynomials, and furthermore, when to determine whether a particular polynomial f can be "solved".

To do this, we first define the notion of a "solvable" polynomial, find a characterization for solvable polynomials, and finally we will connect this to the notion of an explicit formula for the roots of a general degree n polynomial. Though our original motivating question came from polynomials in $\mathbb{R}[x]$, we will see that our analysis bears fruit for general fields.

Suppose we have a field F and some $f \in F[x]$. As a matter of shorthand, we define the **Galois group** of f as $\mathrm{Aut}_F(K)$ where K/F is a splitting field extension over f.

> **Definition 5.40 (Simple Radical Extension)**: An extension K/F is simple if $K = F(a)$ and there are some $b \in F$ and $n \in \mathbb{N}$ such that $a^n = b$.

An example of a simple radical extension is $\mathbb{Q}\left(\sqrt{2}\right)/\mathbb{Q}$ because $\sqrt{2}^2 = 2 \in \mathbb{Q}$. Also, it is very important to remember that $\mathbb{Q}(i)$ is a radical extension too ($i^2 = -1 \in \mathbb{Q}$, after all).

Definition 5.41 (Radical Extension): An extension K/F is **radical** if there is a series of fields

$$F = K_0 \leq \ldots \leq K_n = K$$

where each extension K_{i+1}/K_i is simple.

Note that the extension $\mathbb{Q}\left(\sqrt{\sqrt{2}+\sqrt{3}}\right)/\mathbb{Q}$ is not a simple radical extension. But it is a radical extension because each of the extensions in the chain

$$\mathbb{Q} \leq \mathbb{Q}\left(\sqrt{2}\right) \leq \mathbb{Q}\left(\sqrt{2}, \sqrt{3}\right) \leq \mathbb{Q}\left(\sqrt{\sqrt{2}+\sqrt{3}}\right)$$

is a simple radical extension.

Now we prove a few technical facts about radical extensions which will make the discussion surrounding solvability much easier.

Theorem 5.42: Suppose that F is a field with characteristic 0 and K/F is a splitting field extension of $x^n - b \in F[x]$. Then there exist $a \in K$ and $\zeta \in K$ with $a^n = b$ and $\zeta^n = 1$ such that $K = F(a, \zeta)$.

Notably, this implies $F(a, \zeta)/F$ is radical because the extensions in the chain $F \leq F(a) \leq F(a, \zeta)$ are simple radical extensions.

Proof of Theorem 5.42: Suppose the roots of $x^n - b$ are $a_1, \ldots, a_n \in K$. Now define $\zeta_i = \frac{a_i}{a_1}$ for each $1 \leq i \leq n$ and note that $\zeta_i^n = 1$. Each of the ζ_i are unique as $x^n - b$ is irreducible and thus separable (remember the characteristic of F is 0) and there are at most n roots of $x^n - 1$, so they are the nth roots of unity.

Note the multiplicative group on $\{\zeta_1, \ldots, \zeta_n\}$ has order n and is cyclic, so define ζ to be a generator of this group. Take $a = a_1$ for concreteness, though the choice of a does not matter.

We can easily see that $K = F(a, \zeta)$. Clearly $x^n - b$ is the minimal polynomial of a, so by separability all of the roots are unique. We have identified n distinct roots $a, a\zeta, \ldots, a\zeta^{n-1}$, so $x^n - b$ splits over $F(a, \zeta)$. And this extension is obviously minimal. □

Definition 5.43: We say a polynomial $f \in F[x]$ is **solvable** (alternatively **solvable by radicals**) precisely when there is a radical extension K/F which f splits over.

Let us quickly take stock of why we should care about solvability. A real number r may be written with the operations $+, \times, (-)^q$ (where q is a rational) over rationals precisely when r is contained in a radical extension of $\mathbb{Q}$.

The existence of a general-form[16] solution for the roots of a degree n polynomial implies that the roots of every rational polynomial are contained in a radical extension. (For we are inserting rationals as the coefficients and performing the operations $+, \times, (-)^q$.)

Now, if there exists even one $f \in \mathbb{Q}[x]$ with $\deg f = n$ which is not solvable, we will have shown that there is no general formula for the roots of a degree n polynomial. So we should look into whether a polynomial is solvable.

(By the way, this entire discussion works with any extension K/F and any $a \in K$.)

> **Theorem 5.44**: In a field F with characteristic 0, a polynomial $f \in F[x]$ is solvable if and only if the Galois group of f is solvable.

Proof of Theorem 5.44: Suppose f is solvable, that is, there exists a chain

$$F \leq F(a_1) \leq \ldots \leq F(a_1, \ldots, a_n)$$

of simple radical extensions where f splits over $F(a_1, \ldots, a_n)$. But we may extend $F(a_1)$ to $F(a_1, \zeta_1)$ and so on to get the chain

$$F \leq F(a_1, \zeta_1) \leq \ldots \leq F(a_1, \zeta_1, \ldots, a_n, \zeta_n)$$

of normal radical extensions. For notational convenience define $F_0 = F$ and $F(a_1, \zeta_1, \ldots, a_i, \zeta_i)$ as F_i for $1 \leq i \leq n$.

Note F_n/F is Galois by construction which implies that every F_n/F_i is Galois.

Considering the chain $F_i \leq F_{i+1} \leq F_n$ and using the Galois correspondence yields that

$$\mathrm{Aut}_{F_{i+1}}(F_n) \trianglelefteq \mathrm{Aut}_{F_i}(F_n),$$

with

$$\mathrm{Aut}_{F_i}(F_n) / \mathrm{Aut}_{F_{i+1}}(F_n) \cong \mathrm{Aut}_{F_i}(F_{i+1}).$$

Note that $\mathrm{Aut}_{F_i}(F_{i+1})$ is abelian as any automorphism $\varphi : F_{i+1} \to F_{i+1}$ fixing F is uniquely determined by $\varphi(a_{i+1}) \in \{a_{i+1}, \ldots, \zeta_i^{n-1} a_{i+1}\}$.

This induces a chain

$$\mathrm{id} = \mathrm{Aut}_{F_n}(F_n) \trianglelefteq \ldots \trianglelefteq \mathrm{Aut}_F(F_n)$$

whose quotients are abelian. Thus $\mathrm{Aut}_F(F_n)$ is solvable by Theorem 2.50.

[16] By general form, I mean an expression with the operations $+, \times, (-)^q$ over the coefficients of the polynomial (as variables).

To be entirely clear, $\mathrm{Aut}_F(F_n)$ is not the Galois group of f; however, it does contain the Galois group. Since the subgroup of a solvable group is solvable, this is sufficient.

Now suppose K/F is a splitting field extension for f and that $\mathrm{Aut}_F(K)$ is solvable. This induces a chain

$$\mathrm{id} = G_0 \leq \ldots \leq G_n = \mathrm{Aut}_F(K)$$

with abelian quotients.

The structure theorem for finite abelian groups implies

$$G_{i+1}/G \cong \frac{\mathbb{Z}}{\left(p_1^{e_1}\right)} \oplus \ldots \oplus \frac{\mathbb{Z}}{\left(p_k^{e_k}\right)},$$

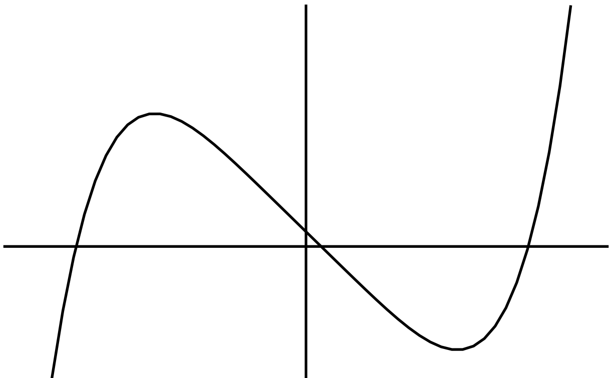
i.e. G_{i+1}/G is the direct sum of cyclic groups. This means we may further decompose G_{i+1}/G into a chain

$$G_i = H_{i,0} \trianglelefteq H_{i,1} \trianglelefteq \ldots \trianglelefteq H_{i,m} = G_{i+1}$$

where every $H_{i,j+1}/H_{i,j}$ is cyclic. And any cyclic Galois group must be the result of a radical extension.[17] Since the composition of radical extensions is obviously radical, we are done. □

To show that there is no formula for the roots of a degree 5 polynomial, we simply have to show the Galois group of $x^5 - 80x + 16$ is unsolvable. So we will do exactly that.

First note that $x^5 - 80x + 16$ is irreducible over $\mathbb{Q}$; we may check this with the Rational Root Theorem. Now note $x^5 - 80x + 16$ has three roots. Informally, we consider the graph of $f(x) = x^5 - 80x + 16$.



Formally, we could look at the inflection points and the derivative of f. We will not bother to perform this formal analysis.

[17] This is not immediately obvious. First note that two Galois extensions K_1/F and K_2/F with the same Galois groups are isomorphic, for the Galois correspondence implies that K_1 and K_2 share the same subfield structure. Then note a cyclic Galois group of order n can be generated by a polynomial of the form $x^n - 1$, so the extension must be radical.

Suppose a is a root of $x^5 - 80x + 16$ and that K/F is a splitting field extension over this polynomial. Note $[F(a) : F] = 5$ so by Theorem 5.6 $[K : F]$ is divisible by 5. Since K/F is Galois (separability is free), the Galois Correspondence implies that $|\mathrm{Aut}_F(K)|$ is divisible by 5 too. Cauchy's says $\mathrm{Aut}_F(K)$ contains an element of order 5, which must be a 5-cycle.

Furthermore, the conjugation $z \mapsto \overline{z}$ fixes the real roots and swaps the complex roots, so it is a 2-cycle.

Exercise 5.45: Show that for prime p, any two-cycle $(1, 1 + k)$ and p-cycle $(1, 2, ..., p)$ in S_p generate the entirety of S_p as follows:

- Denote $(1, ..., p)$ as σ. Show that $\sigma(i, j)\sigma^{-1} = (i + 1, j + 1)$ in general.
- This allows us to generate $(a, a + k)$ for every $1 \leq a \leq p$ (of course, all taken modulo p). Now $1 \to 1 + k \to 1 + 2k \to ... \to 1 + pk$ allows us to swap any two elements we want.
- Now it is fairly obvious that repeatedly swapping elements can get us any permutation.

  If you really want to see for yourself, you may inductively show this: to construct a desired permutation φ, swap to get $\varphi(p)$ be correct, and then permute $\varphi \upharpoonright \{1, ..., p - 1\}$ correctly with the inductive hypothesis.

This implies that $\mathrm{Aut}_F(K) = S_5$. And since S_5 is not solvable (the specific reasons of which we will not go into), f is not solvable.

Of course, the same line of reasoning works for any f with three real roots and two complex roots. For example, $x^5 - 80x + 5$ also is unsolvable over radicals. It is irreducible (Eisenstein's Criterion is sufficient to show this), and it has the correct number of real/complex roots.

5.6 Characterizing the Finite Fields

It is a fact that every finite field has order p^n. Conversely, for every p^n there is a unique finite field of that order. We set out to prove this and a few other facts about finite fields.

> **Theorem 5.46**: Suppose F is a finite field. Let $F^* = (F \setminus \{0\}, \times)$ be the multiplicative group of F. Then F^* is cyclic.

Proof of Theorem 5.46: The polynomial $x^p - 1$ has at most p roots in F — one of which is 1 — so there are at most $p - 1$ elements in F^* whose order is p.

The structure theorem for finite abelian groups says that

$$F^* \cong \frac{\mathbb{Z}}{(p_1^{e_1})} \oplus ... \oplus \frac{\mathbb{Z}}{(p_n^{e_n})}.$$

If there exist $i \neq j$ such that $p_i = p_j$, then we may find $p - 1$ elements of order p in each of $\mathbb{Z}/\left(p_i^{e_i}\right)$ and $\mathbb{Z}/\left(p_j^{e_j}\right)$. This gives us $2p - 2 > p - 1$ elements of order p.

So $p_1, ..., p_n$ are all distinct. And thus the group is cyclic, for it is generated by the element $(1, ..., 1)$. □

Note that every finite field F has prime characteristic p, implying the prime subfield is isomorphic to $\mathbb{F}_p$. And if $P \leq F$ is the prime subfield, then obviously $[F : P]$ is finite. Suppose $(a_1, ..., a_n)$ is a basis of F as a vector space in P. Then each of the linear combinations

$$p_1 a_1 + ... + p_n a_n$$

where $p_i \in P$ is unique for each distinct tuple $(p_1, ..., p_n)$. This implies that F has p^n elements. So every field has cardinality of the form p^n for some prime p and positive integer n.

If fields F_1 and F_2 both have p^n elements, then we show F_1 and F_2 are isomorphic. We can assume their prime subfields are $\mathbb{F}_p$, otherwise we simply find fields isomorphic to F_1 and F_2 whose prime subfields are $\mathbb{F}_p$.

Now

$$f(x) = x^{p^n - 1} - 1$$

is a polynomial whose roots are every $a \in F_1$ besides 0. So evidently F_1 splits over $f \in \mathbb{F}_p[x]$. Obviously $F_1 / \mathbb{F}_p$ is a splitting field extension over f. Similarly $F_2 / \mathbb{F}_p$ is a splitting field extension over f, and the uniqueness of the splitting field extension implies $F_1 \cong F_2$.

We use very similar reasoning to conclude a field of order p^n exists for every prime p and positive integer n. Because $f(x) = x^{p^n - 1} - 1 \in \mathbb{F}_p[x]$ has derivative -1, it is separable. Now take a splitting field extension $K / \mathbb{F}_p$, and take the map $\varphi : K \to K$ where $\varphi : a \mapsto a^p$. (This is known as the **Frobenius Endomorphism**.)

Define m to be $\left[K : \mathbb{F}_p\right]$ and note that $K \cong F_{p^m}$. Thus $\varphi\left(a^{p^{m-1}}\right) = a^{p^m} = a$, showing that φ is surjective, and any surjection from a finite set to itself must also be an injection. Obviously φ preserves multiplication, and the Binomial Theorem shows that

$$\varphi(a + b) = (a + b)^p = a^p + b^p = \varphi(a) + \varphi(b)$$

as the characteristic of K is still p. So φ preserves addition as well.

Finally, note that $\varphi^m : a \mapsto a^{p^m}$ is an automorphism which fixes precisely the roots of f and 0. Putting this all together, $\varphi^m \in \operatorname{Aut}_{\mathbb{F}_p}(K)$ and $|\operatorname{Fix}(\varphi^m)| = p^n$.[18]

Thus a field of order p^n exists.

[18] Strictly speaking, we really should talk about the fixed field of the subgroup of $\operatorname{Aut}_{\mathbb{F}_p}(K)$ generated by φ^m to generate a subfield. But because this subgroup is cyclic, it suffices to just consider the generator φ^m.

Chapter 6

Concluding Remarks

Congratulations on making it through. If you fully understood the text, you should now have a strong conceptual understanding of the content in an introductory graduate algebra course. In particular, the hope is that you understand how topics are connected and have internalized why the important theorems are true.

I realize that it may be a little rich to call this text a primer when it clocks in at over a hundred pages, but I stick by it. This primer is thorough in many respects, but there are many gaps which it leaves.

Most importantly, there are startlingly few worked examples and exercises in this text. We define only the concepts we really need and move on immediately afterwards. Just looking at group theory, we shown very few examples of using Sylow's to characterize groups of a certain finite order. For instance, we can use Sylow's and some combinatorics to show that no group of order 56 is finite. Nor have we done very many Orbit-Stabilizer exercises. For instance, we have not mentioned that for groups $H \leq G$, $[G : H] = 2$ implies $H \trianglelefteq G$. And in general, if $|G|$ is finite and p is the smallest prime dividing G, $[G : H] = p$ implies $H \trianglelefteq G$ as well.

This is not to mention that important concepts and topics such as characteristic subgroups and representation theory have been completely omitted. Nor do we really pay much heed to non-commutative ring theory or module theory.[1] There is so much more even to the basics of abstract algebra, and I believe it is a field worth further studying.

Thank you for reading this primer. If you have any feedback, please email me at dchen@dennisc.net.

[1] In particular, it is possible to define left and right R-modules and simultaneously consider them.