Algebra

Justin Lawrence

August 31, 2024

Preface

To use the fun buzzwords that every textbook has, *a certain level of mathematical maturity is assumed*. Also assumed is an understanding of linear algebra at an introductory undergraduate level. No familiarity with categories or abstract algebra is assumed. The book may be better titled Intermediate Algebra, as its level of sophistication lies somewhere between an undergraduate and graduate approach to algebra. Sections marked with $*$ are considered optional : it is fairly safe to assume that no future sections will depend on them. There are a few exceptions, see the reading guide for details.

It is also worth mentioning what this book is not. It is not a textbook. You won't find any practice problems here, nor any long or worked through examples (except in the abstract view section). I will occasionally point the interested reader in the direction of such things, but keep in mind that the other sources I reference in doing so may use a different notation or approach than I do.

None of the ideas contained within this book were mine to begin with. Much of the book is based off of Serge Lang's Algebra [Lan05], a wonderful reference text that I suggest you pick up yourself. Other inspirations include Nathan Jacobson's Basic Algebra I [Jac09] for the earlier sections (although I do somewhat hate that book), Steven Roman's Advanced Linear Algebra [Rom07], and the lectures of Lior Silberman [Sil23] and Kalle Karu. I am of course also in debt to the many professors I've had so far in Algebra throughout my years of education. In particular, it was Vinyak Vatsal who originally taught me the basics of algebra (from that cursed Jacobson book), and Lior Silberman who taught me to see algebra in a more unified way.

Finally, I would like to thank two other people. My friend Boris for doing his own version of this, forcing me to stop procrastinating on this project. Second, I would like to thank Ben Williams, both for his wonderful insights and advice over the past year, and for making me install a spell checker. I'm sure there are still grammatical errors in this text that would cause you to lose your mind slightly, but please know that I really did try to fix them[1].

[1]Somehow, despite only being fluent in one language, I'm quite terrible at it.

Reading Guide

This text is split into five sections, and is designed to (for the most part) be read in a linear order.

The first section, *foundations*, consists only of the first chapter. It contains a collection of basic results in mathematics which are required to understand any other chapter of this text, along with an explanation of Zorn's lemma. The first of these must be fully understood before reading anything else, the second can be skipped at first and returned to whenever it is needed.

The second section, *basic algebra*, consists of chapters 2 and 3. These introduce the reader to two of the most fundamental objects in algebra; the group and ring respectively. Both of these chapters are designed to be read in order, with the section on groups first. There is one exception to the * means optional rule here, which is the final section of chapter 3 on Chinese Remainder Theorem.

The third section, *linear algebra*, covers exactly what the title says, primarily from the viewpoint of modules rather than vector spaces. It consists of two chapters, again designed to be read in an entirely linear order. This section is also more difficult to understand than the previous two.

The fourth section, *the abstract view*, details how to unify the concepts of the previous chapters and view algebra in a more abstract way. This includes a sort of warm up chapter on universal algebras, followed by the much more important topic of categories. The universal algebras chapter can be skipped, as it won't come up again, but it is rather short and may be reading just out of interest. It also may be worth reading if you've had no exposure to this level of abstraction before, as universal algebras are a bit more concrete than categories. The last chapter on categories is complete as-is, but may have more material added to it later (specifically on adjoints).

The final section, *advanced algebra*, contains chapters on a collection of topics which may be encountered in an advanced undergraduate or introductory graduate course. These include a chapter on field extensions and Galois theory, one on commutative algebra, and one on homology. The chapter on Galois theory could be read before the previous section, as it doesn't really use categories. Only the first of these chapters has been written.

Contents

Chapter 1

Foundations

Unfortunately, we cannot immediately jump into algebra without first having a strong understanding of some basic mathematical concepts. The first of these is a combination of basic facts about prime numbers, factorization, and equivalence relations. These results will be used immediately and frequently in our study of algebra, and you should have a firm understanding of them before moving on. The second is an overview of Zorn's lemma. This will not be used until a fair bit into our studies, and can be skipped for now if desired. When you come back to it, it is not essentially that you understand the proof of the lemma, just how to apply it.[1]

1.1 Primes and Equivalence Classes

This section collects results from a similar one in [Jac09] (which is honestly one of his better expository moments), and is here for your convenience. If you are not familiar with these concepts to some degree already, please read the corresponding sections in [Jac09]. We start with some very basic definitions.

Definition 1.1.1. Let $a, b \in \mathbb{Z}$. We say that a divides b, or a is a divisor of b, denoted $a \mid b$, if there exists some $x \in \mathbb{Z}$ such that $b = az$. We say that a number $p \in \mathbb{Z}$ is prime if its only divisors are $\pm 1, \pm p$. By convention, we do not consider ± 1 to be prime.

Note that if $b \mid c$ and $a \mid b$, then $a \mid c$. Indeed, we use this fact to prove one of the most fundamental theorems of mathematics.

Theorem 1.1.2 (Prime Factorization)**.** *Any number $n \in \mathbb{N}$ has a unique (up to order of primes) representation in the form*

$$n = p_1^{e_1} \cdots p_r^{e_r} \tag{1.1}$$

where $p_i \in \mathbb{N}$ are prime, and $e_i \in \mathbb{N}$.

For a proof, see [Jac09] (and apply this statement to every part of this section).

[1] It is a good exercise in dealing with abstract concepts to understand it however.

Definition 1.1.3. Let $a, b \in \mathbb{Z}\backslash\{0\}$. Then we define

1. A greatest common denominator GCD of a, b to be any number $c \in \mathbb{Z}$ such that $c \mid a, b$ and if $d \mid a, b$ then $d \mid c$.
2. A least common multiple of a, b to be any number $c \in \mathbb{Z}$ such that if $a, b \mid c$ and if $a, b \mid d$ then $c \mid d$.

Theorem 1.1.4. *Let $a, b \in \mathbb{Z}\backslash\{0\}$. Let $|a| = p_1^{e_1} \cdots p_r^{e_r}, |b| = p_1^{f_1} \cdots p_r^{f_r}$ be prime factorizations, where we allow $e_i, f_i = 0$. Set $M_i = \max(e_i, f_i), m_i = \min(e_i, f_i)$. Then*

1. *The GCDs of a, b are $\pm p_1^{m_1} \cdots p_r^{m_r}$.*
2. *The LCMs of a, b are $\pm p_1^{M_1} \cdots p_r^{M_r}$.*

Because of the above theorem, we usually denote the positive GCD/LCM by $(a, b), [a, b]$.

Proposition 1.1.5. *Let $a, b \in \mathbb{N}$. Then*

$$(a, b)[a, b] = ab$$

Theorem 1.1.6 (Division Algorithm)**.** *Suppose $a, b \in \mathbb{Z}$, with $b \neq 0$. Then there exists some $q, r \in \mathbb{Z}$ such that $0 \leq r < |b|$ and*

$$a = qb + r$$

Corollary 1.1.6.1 (Bézout's identity)**.** *Suppose $a, b \in \mathbb{Z}$ are non-zero. Then there exist $m, n \in \mathbb{Z}$ such that $ma + nb = (a, b)$ and*

$$\{ca + db \mid c, d \in \mathbb{Z}\} = \{c(a, b) \mid c \in \mathbb{Z}\}$$

Now, we move on to equivalence relations.

Definition 1.1.7. Let S be a set. A relation R on S is a subset of $S \times S$. If $(x, y) \in R$, we denote this by xRy.

Definition 1.1.8. Let S be a set. An equivalence relation $\sim$ on S is a relation satisfying the following three axioms for all $x, y, z \in S$

1. $x \sim x$ (Symmetry)
2. $x \sim y \Rightarrow y \sim x$ (Reflexivity)
3. $x \sim y, y \sim z \Rightarrow x \sim z$ (Transitivity)

Equivalence relations allow us to partition sets into what are called equivalence classes.

Definition 1.1.9. Let $x \in S$, and $\sim$ be an equivalence relation on S. The equivalence class of x, denoted $[x]_\sim$ or just $[x]$, is defined in the following manner.

$$[x] = \{y \in S \mid x \sim y\}$$

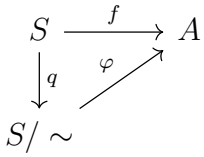
Proposition 1.1.10. *$[x] = [y]$ if and only if $x \sim y$. The set $S/\sim$ of equivalence classes in S (called the quotient set of S) is therefore a partition of S, that is a division of S into disjoint subsets whose union is S.*

The map $q : S \to S/\sim$ defined by $x \mapsto [x]$ is called the quotient function. One can actually go the other way as well.

Proposition 1.1.11. *Let $\pi \subset \mathcal{P}(S)$ be a partition. Then there exists a unique equivalence relation $\sim$ on S such that $S/\sim = \pi$. In particular, this $\sim$ is defined by two elements being equivalent if and only if they lie in the same set in the partition.*

Finally, sufficiently well-behaved functions on S will induce unique maps on the quotient set.

Theorem 1.1.12. *Suppose $f : S \to A$ is a mapping between sets, and $\sim$ is an equivalent relation on S with quotient function q. If f is such that $x \sim y \Rightarrow f(x) = f(y)$, for all $x, y \in S$, then there exists a unique function $\varphi : S/\sim \to A$ such that $\varphi \circ q = f$, that is such that the following diagram commutes.*



1.2 Zorn's Lemma

This section is a streamlined version of a similar section in [Lan05].

Zorn's lemma is to algebra what Fourier transforms are to physics. Nobody will every explicitly teach it to you, but it gets brought up constantly and at some point you seem to be expected to just learn it via osmosis. I open with it in hopes that, if you haven't learned it before, now will be your chance to learn about Zorn's lemma and its proof.

We begin with a definition.

Definition 1.2.1. Let S be a set. A partial ordering of S is a relation $\leq$ between elements of S satisfying the following axioms $\forall x, y, z \in S$

1. $x \leq x$
2. $x \leq y \wedge y \leq z \Rightarrow x \leq z$
3. $x \leq y \wedge y \leq x \Rightarrow x = y$

Note. We do not require that every pair of elements in S be comparable. If this additional condition is satisfied, we call $\leq$ a *total ordering*, and say that S is *totally ordered*. Totally ordered subsets of partially ordered sets are often called chains. If $x \leq y$ and $x \neq y$, we write that $x < y$.

We follow this up with a collection of definitions related to Definition 1.2.1.

Definition 1.2.2. Let S be an ordered set. A smallest element of S is an element $a \in S$ such that $a \leq x$ for all $x \in S$, with a greatest element defined similarly. A maximal element $m \in S$ is an element such that if $m \leq x \Rightarrow x = m$, and a minimal element is defined similarly.

Note. Maximal and greatest elements are not identical notions. Indeed, one can note that maximal elements need not be a greatest element, and that greatest elements are unique when they exist (while maximal elements may not be). A similar result hold minimal and smallest elements.

Definition 1.2.3. Let $T \subseteq S$ be a subset of a partially ordered set. An upper bound of T in S is an element $a \in S$ such that $t \leq a$ for all $t \in T$. A least upper bound of T in S is an upper bound $b \in S$ such that for any other upper bound $a \in S$, $b \leq a$. A set is inductively ordered if every non-empty totally ordered subset has an upper bound, and strictly so if it has a least upper bound.

Definition 1.2.4. Let A be a non-empty partially and strictly inductively ordered set. A map $f : A \to A$ is increasing if, for all $x \in S$, $x \leq f(x)$.

Definition 1.2.5. Let A be a non-empty partially and strictly inductively ordered set, and $f : A \to A$ an increasing map. Pick some $a \in A$, and let $B \subseteq A$. We say that B is admissible with respect to a if

1. $a \in B$
2. $f(B) \subseteq B$
3. Whenever T is a non-empty totally ordered subset of B, the least upper bound of T in A lies in B

Definition 1.2.6. Let A be a non-empty partially and strictly inductively ordered set with minimal element $a \in A$, and $f : A \to A$ an increasing map. We define $M(A, f)$ to be the intersection of all admissible subsets of A with respect to a.

Note. It is not too difficult to see that $M(A, f)$ is the smallest admissible subset of A with respect to a, and is contained in all admissible subsets of A with respect to a. Furthermore, $M(A, f)$ is strictly inductively ordered.

Definition 1.2.7. Let A be a non-empty partially and strictly inductively ordered set with minimal element $a \in A$, and $f : A \to A$ an increasing map. We say that $c \in M(A, f)$ is an *extreme point* of $M(A, f)$ if $x \in M(A, f), x < c \Rightarrow f(x) \leq c$. We further define for such points that

$$M_c(A, f) = \{x \in M(A, f) \mid x \leq c \vee f(c) \leq x\}$$

Note. $a \in M_c(A, f)$, so this set is necessarily non-empty.

We next build up a series of lemmas related to these definitions. The point of this is to prove the third of the lemmas, which will be used to prove a subsequent theorem.

Lemma 1.2.8. *Let A be a non-empty partially and strictly inductively ordered set with minimal element $a \in A$, $f : A \to A$ an increasing map, and $c \in M(A, f)$ be an extreme point. Then $M_c(A, f) = M(A, f)$.*

Proof. It suffices to prove that $M_c(A, f)$ is admissible with respect to a. We already have $a \in M_c(A, f)$. Suppose $x \in M_c(A, f)$. If $x < c$, then since c is an extreme point we get $f(x) \leq c$, so $f(x) \in M_c(A, f)$. If $x = c$, then $f(c) \leq f(x) \Rightarrow f(x) \in M_c(A, f)$. If $f(c) \leq x$, then $f(c) \leq x \leq f(x) \Rightarrow f(c) \in M_c(A, f)$. Thus, $f(M_c(A, f)) \subseteq M_c(A, f)$ as required. Let $T \subseteq M_c(A, f)$ be a non-empty totally ordered subset, and $b \in M(A, f)$ the least upper bound of T in $M(A, f)$. Pick any $x \in T$. If $f(c) \leq x$, then $f(c) \leq b$ so $b \in M_c(A, f)$. If $x \leq c$ for all $x \in T$, then $b \leq c \Rightarrow b \in M_c(A, f)$, as required. □

Lemma 1.2.9. *Let A be a non-empty partially and strictly inductively ordered set with minimal element $a \in A$, and $f : A \to A$ an increasing map. Then every element of $M(A, f)$ is an extreme point.*

Proof. Let $E \subseteq M(A, f)$ be the set of all extreme points. Again, it suffices to show that E is admissible with respect to a. a is vacuously extreme, so $a \in E$. Now, pick any $c \in E$, $x \in M(A, f)$. Suppose $x < f(c)$. By lemma 1.2.8, $M_c(A, f) = M(A, f)$, so $x \leq c$ or $f(c) \leq x$. If $x < c$, then $f(x) \leq f(c)$. If $x = c$, then $f(x) \leq f(c)$. This proves that $f(c) \in E$ as desired. Finally, let $T \subseteq E$ be a non-empty totally ordered subset, and $b \in M(A, f)$ the least upper bound of T in $M(A, f)$. Suppose $x \in M(A, f)$ and $x < b$. Then $\exists c \in T$ such that $x \leq c$ (indeed, we otherwise get by lemma 1.2.8 that $f(c) \leq x$ for all $c \in T$, and hence $c \leq x$ for all $c \in T$, so $b \leq x$). If $x < c$, then $f(x) \leq c \leq b$ so $f(x) \leq b$. If $x = c$, then by lemma 1.2.8 we must get $f(x) \leq b$ (as otherwise $b \leq x$). This shows that $b \in E$, and hence completes the proof. □

Lemma 1.2.10. *Let A be a non-empty partially and strictly inductively ordered set with minimal element $a \in A$, and $f : A \to A$ an increasing map. Then $M(A, f)$ is totally ordered.*

Proof. Pick any $x, y \in M(A, f)$. By lemma 1.2.9, y is an extreme point of $M(A, f)$, so by lemma 1.2.8 either $x \leq y$ or $y \leq f(y) \leq x \Rightarrow y \leq x$. □

Using this result, we prove a powerful theorem of which Zorn's lemma is a corollary.

Theorem 1.2.11 (Bourbaki's Theorem)**.** *Let A be a non-empty partially and strictly inductively ordered set, and $f : A \to A$ an increasing map. Then $\exists x_0 \in A$ such that $f(x_0) = x_0$, that is f has a fixed point.*

Proof. Suppose that A is totally ordered. Then since it has a least upper bound $b \in A$, $b \leq f(b) \leq b \Rightarrow b = f(b)$, as required. Otherwise, it suffices to find an admissible totally ordered subset of A. Pick some $a \in A$ and let B be the set of elements $x \in A$ such that $x < a$. Then $A \backslash B$ is admissible with respect to a, and a is a minimal element of $A \backslash B$, so we may assume without loss of generality that A has a minimal element $a \in A$. By lemma 1.2.10, $M(A, f)$ is the desired totally ordered admissible subset. □

Corollary 1.2.11.1 (Weak Zorn's Lemma)**.** *Let A be a non-empty partially and strictly inductively ordered set. Then A has a maximal element.*

Proof. Suppose not. Then for any $x \in A$, there exists some $y_x \in A$ such that $x < y_x$[2], as otherwise x would be maximal. Let $f : A \to A$ be defined by $f : x \mapsto y_x$. Then f is increasing, so by Theorem 1.2.11 f has a fixed point, which is impossible. □

Corollary 1.2.11.2 (Zorn's Lemma)**.** *Let A be a non-empty partially and inductively ordered set. Then A has a maximal element.*

Proof. Let B be the set of non-empty totally ordered subsets of A. Then B is not empty, as any singleton is totally ordered. If $X, Y \in B$, we define a partial order $\leq$ on B by $X \leq Y \iff X \subseteq Y$. In fact, this makes B strictly inductively ordered. To see this, let $T = \{X_i\}_{i \in I} \subset B$ be totally ordered, and let $Z = \cup_{i \in I} X_i$. Pick any $x, y \in Z$. Then $x \in X_i, y \in Y_j$ for some $i, j \in I$. Since T is totally ordered, we get (without loss of generality) $X_i \subseteq X_j$, so since $X_j \in B$ we must have $x \leq y$ or $y \leq x$. Thus, Z is totally ordered, and hence clearly a least upper bound of T. Therefore, B is a non-empty partially and strictly inductively ordered set, and therefore has a maximal element $X_0 \in B$. Since A is inductively ordered, X_0 has an upper bound $m \in A$. We'll show that m is a maximal element of S. Indeed, suppose that $x \in S$ and $m \leq x$. Then $X_0 \cup \{x\}$ is totally ordered, so by the maximality of X_0 we must get $X_0 = X_0 \cup \{x\} \Rightarrow x \in X_0 \Rightarrow x \leq m$, so $x = m$, as was to be shown. □

Note. The non-empty condition comes from the definition of an inductively ordered set, and isn't really needed. Indeed, suppose that A is a non-empty partially ordered set such that every totally ordered subset has an upper bound. Then in particular every non-empty totally ordered subset has an upper bound, so by Zorn's lemma A has a maximal element.

Zorn's lemma turns out to be equivalent to the axiom of choice, but as this is not a book on set theory we won't get further into that here. If you're interested in learning more about that, it may be worth starting at this rabbit hold of a Wikipedia page[3].

[2]This choice of y_x is invoking the axiom of choice

[3]`https://en.wikipedia.org/wiki/Axiom_of_choice#Equivalents`

Part I

Basic Algebra

Chapter 2

Groups

2.1 Basic Definitions

We start with the basic definitions of group theory.

Definition 2.1.1. A monoid is a tuple $(M, 1, \cdot)$, where M is a set, $1 \in M$, and $\cdot : M \times M \to M$ is an operation such that for all $a, b, c \in M$

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c \qquad\qquad 1 \cdot a = a \cdot 1 = a$$

We usually call this operation *multiplication*, and omit the $\cdot$ when writing it. We also denote monoids with just their set, M. A monoid is called *Abelian* if its multiplication is commutative.

Example 2.1.1. The natural numbers $\mathbb{N}$ under addition or multiplication are a monoid. Another good example is $\mathbb{R}^{n\times n}$ under multiplication, which is not an Abelian monoid.

Definition 2.1.2. Let M be a monoid. For any $a \in M$, we call $b \in M$ the right-inverse of a if $ab = 1$, and the left-inverse if $ba = 1$. b is called the inverse of a if it is both a left and right inverse. A monoid where every element has an inverse is called a group.

Example 2.1.2. $\mathbb{R}^{n\times n}$ under multiplication is not a group, due to the lack of inverses. The set of all invertible matrices in $\mathbb{R}^{n\times n}$ is a group.

Theorem 2.1.3. *Let $a \in M$ be an invertible (has an inverse) element of a monoid. Then its inverse is unique.*

Proof. Suppose $b, c \in M$ are inverses of a. Then $ab = ac \Rightarrow b(ab) = b(ac) \Rightarrow (ba)b = (ba)c \Rightarrow 1b = 1c \Rightarrow b = c$. □

As a result of Theorem 2.1.3, we denote the unique inverse of an element $a \in M$ by a^{-1}.

Definition 2.1.4. A sub-monoid (sub-group) of a monoid M is a subset of M which is itself a monoid (group).

Theorem 2.1.5. *Let $\{U_\alpha\}_{\alpha\in I}$ be a collection of sub-monoids of a monoid M. Then $U = \bigcap_{\alpha\in I} U_\alpha$ is a sub-monoid of M.*

Proof. Since $1 \in U_\alpha$ for every $\alpha \in I$, $1 \in U$. Associativity of multiplication is inherited from the monoid M, so it suffices to check that U is closed under multiplication. Pick any $a, b \in U$. Then $a, b \in U_\alpha$ for all $\alpha \in I$, so $ab \in U_\alpha$ for all $\alpha \in I$. Thus, $ab \in U$, as was to be shown. □

Note. An identical result holds for groups.

Definition 2.1.6. Let $S \subseteq M$ be a subset of a monoid M. The sub-monoid generated by S, denoted by $\langle S \rangle$, is the intersection of all sub-monoids of M containing S.

Note. By the result of Theorem 2.1.5, $\langle S \rangle$ is in fact a sub-monoid of M.

We next state a result which provides a much more practical way of expressing the module generated by a set.

Theorem 2.1.7. *$\langle S \rangle$ is the set of all elements of M that may be written as a product of 1 and the elements of S.*

Proof. Let X be the collection of products of elements of S and 1. Since $\langle S \rangle$ is a sub-monoid containing S and 1, any product of those elements is in $\langle S \rangle$, so $X \subseteq \langle S \rangle$. But X is closed under multiplication, and hence a sub-monoid of M containing S. Therefore, $\langle S \rangle \subseteq X$, as was to be shown. □

Note. We can make an identical definition for groups. Theorem 2.1.7 still holds if you include the inverse of every element in S in the products.

Definition 2.1.8. Let $a \in M$ be an element of a monoid. The order of a, denoted $o(a)$, is the smallest $n \in \mathbb{N}$ such that $a^n = 1$. If no such natural number exists, we write that $o(a) = \infty$.

Theorem 2.1.9. *If an element $a \in M$ has finite order, it is invertible.*

Proof. Suppose that $o(a) = n$. Then $a^n = 1 = a^{n-1}a = aa^{n-1}$, so $a^{n-1} = a^{-1}$. □

Theorem 2.1.10. *Every element of a finite group G has finite order.*

Proof. Suppose $a \in G$ has infinite order. Pick any $n \geq m \in \mathbb{N}$. If $a^n = a^m$, then $a^{n-m} = 1$. Then a has finite order if $n \neq m$, so $n = m$. It follows that a^k is distinct for each $k \in \mathbb{N}$. But this would imply that G is infinite, a contradiction. □

Definition 2.1.11. Let M, N be monoids. A homomorphism $\varphi : M \to N$ is a map such that for any $a, b \in M$, $\varphi(ab) = \varphi(a)\varphi(b)$.

Note. Since $\varphi(a) = \varphi(1a) = \varphi(1)\varphi(a)$ and $\varphi(a) = \varphi(a1) = \varphi(a)\varphi(1)$ for any homomorphism φ and monoid element a, $\varphi(1) = 1$ for any homomorphism.

An injective homomorphism is called a *monomorphism*, and a surjective one an *epimorphism*. A bijective homomorphism is called an *isomorphism*. If there exists an isomorphism between two monoids M, N, we call them *isomorphic* and write that $M \cong N$. Note that $\cong$ is an equivalence relation.

2.2 Group of Transformations

This section is partially based on (and uses some proofs from) a similar section in [Jac09]. For the next few definitions, let X be an arbitrary set.

Definition 2.2.1. The monoid $M(X)$ of all transformations on X is the set of all set functions $f : X \to X$, with multiplication given by function composition. The group $\mathrm{Sym}(X)$ of all invertible elements of $M(X)$ is the symmetric group of X.

It is not too difficult to check that these are indeed a monoid and group respectively. Any sub-monoid of $M(X)$ is called a monoid of transformations, and any sub-group of $\mathrm{Sym}(X)$ a group of transformations.

Definition 2.2.2. Suppose that X is a finite set with $|X| = n$. Then we call $\mathrm{Sym}(X) = S_n$ the permutation group on n elements. A sub-group of S_n is called a permutation group on n elements.

We generally represent permutations $\sigma \in S_n$ as the product of disjoint cycles. That is, denoting $X = \{1, 2, \ldots, n\}$ we first look at the sequence

$$1, \sigma(1), \sigma^2(1), \ldots$$

Since $|X| < \infty$, this sequence is finite. Since σ is invertible, the image of the final element of the sequence is 1. We call this sequence a disjoint cycle, and write it $(1\sigma(1)\sigma^2(1)\cdots)$. This represents a function on X that takes 1 to $\sigma(1)$, $\sigma(1)$ to $\sigma^2(1)$ and so on, and does nothing on any element not in the cycle. Repeating this process with an element not in any previous cycle, we get the following result.

Proposition 2.2.3. *Any permutation can be written as the product of disjoint cycles, and this product is unique up to the order of the cycles. Furthermore, disjoint cycles commute.*

Note. The presentation of any given cycle is not unique, for example $(123) = (312) = (231)$. In general, doing a cyclic permutation of the elements of a cycle does not change the cycle that it represents.

Now, we present two extremely important results on symmetric groups.

Theorem 2.2.4. *For any $n \in \mathbb{N}$, $|S_n| = n!$.*

Proof. We do this constructively. Pick any $\sigma \in S_n$. We have n valid choices for $\sigma(1)$. Then since σ is bijective, we have $n - 1$ for $\sigma(2)$, $n - 2$ for $\sigma(3)$, and so on until we have only one choice for $\sigma(n)$. Thus, $|S_n| = n!$, as was to be shown. □

Theorem 2.2.5 (Cayley's Theorem)**.** *Let G be a finite group. Then there exists a finite set X and group of transformations H on X such that $G \cong H$.*

Proof. Set $X = G$, and for any $g \in G$ consider the transformation $L_g : X \to X$ given by

$$L_g(x) = gx$$

We'll prove that $H = \{L_g \mid g \in G\}$ is a group, and that the map $\varphi : g \mapsto L_g$ is an isomorphism. Let $L_a, L_b, L_c \in H$. Then we get

$$\begin{aligned}(L_aL_1)(x) = L_a(L_1(x)) = L_a(1x) = L_a(x) = ax\\ (L_1L_a)(x) = L_1(L_a(x)) = L_1(ax) = 1(ax) = ax\\ (L_aL_b)(x) = L_a(L_b(x)) = L_a(bx) = a(bx) = (ab)x = L_{ab}(x)\\ ((L_aL_b)L_c)(x) = (L_aL_b)(cx) = (ab)(cx) = a(bcx) = (L_a(L_bL_c))(x)\\ (L_aL_{a^{-1}})(x) = L_a(L_{a^{-1}}(x)) = a(a^{-1}x) = x\\ (L_{a^{-1}}L_a)(x) = L_{a^{-1}}(L_a(x)) = L_{a^{-1}}(ax) = a^{-1}(ax) = x\end{aligned}$$

The first two lines prove that L_1 is our unit 1, the third closure under multiplication, the fourth associativity, and the last two that $L_{a^{-1}}$ is the inverse of L_a, so H is a group. Suppose that $a, b \in G$ satisfy $\varphi(a) = \varphi(b)$. Then in particular, $b = b1 = \varphi(b)(1) = \varphi(a)(1) = a1 = a$, so φ is injective. φ is clearly surjective, and is hence bijective. It remains to show that it is a homomorphism. Indeed, we get for any $x \in G$

$$\varphi(ab)(x) = (ab)(x) = a(bx) = \varphi(a)(\varphi(b)(x)) = (\varphi(a)\varphi(b))(x)$$

so φ is a homomorphism, completing the proof. □

Note. Cayley's theorem allows us to study any finite group (in theory) by studying subgroups of S_n, which simplifies our life considerably. The only trick is that you often need to understand the structure of the group to find its corresponding subgroup of S_n, which has an unfortunate circular quality.

Note. Cayley's theorem extends to infinite monoids/groups, which are isomorphic to monoids/groups of transformation. The proof is essentially identical.

Finally, we'll develop the notion of the *sign* of a permutation. Before starting on this journey, we'll need the following lemma.

Lemma 2.2.6. *Let $a, b, c_1, \ldots, c_m, d_1, \ldots d_k$ be distinct elements in a finite set X. Then the following two equations hold.*

1. $$(ab)(ac_1 \cdots c_m bd_1 \cdots d_k) = (ac_1 \cdots c_m)(bd_1 \cdots d_k)$$

2. $$(ab)(ac_1 \cdots c_m)(bd_1 \cdots d_k) = (ac_1 \cdots c_m bd_1 \cdots d_k)$$

Proof. The first of these is obtained by tracing through the result of applying both sides to $a, b, c_1, \ldots, c_m, d_1, \ldots d_k$. We also note that $(ab)^2 = 1$. Then applying (ab) to both sides of the second equality, we see that it is in fact equivalent to the first. □

Note. Result (1) in the proceeding lemma implies that any disjoint cycle can be split into two smaller disjoint cycles. The above results also still hold if $m = 0$ or $k = 0$.

Using this, we can find an alternative way of representing permutations.

Theorem 2.2.7. *Any permutation can be written as the product of transpositions (cycles of the form (ab), where $a \neq b$).*

Proof. Since every permutation can be written as the product of disjoint cycles, it suffices by Lemma 2.2.6 to prove this for cycles of the form (abc), where a, b, c are all distinct. Indeed, we note that $(abc) = (ab)(bc)$, as required. □

This transposition decomposition is closely related to the sign of a permutation, which we now (finally) define.

Definition 2.2.8. Let $\sigma = C_1 C_2 \cdots C_m$ be the disjoint cycle representation of a permutation σ, where each C_i is a cycle of length d_i. Then the sign of σ is defined as

$$\operatorname{sgn}(\sigma) = (-1)^{\sum_{i=1}^{r}(d_i - 1)}$$

Note. Since the disjoint cycle decomposition of a permutation is unique up to the order of cycles, sgn is in fact well-defined.

Proposition 2.2.9. *Let σ be a permutation and τ a transposition. Then* $\operatorname{sgn}(\tau\sigma) = -\operatorname{sgn}(\sigma)$.

Proof. Let $\tau = (ab)$. There are then two cases to consider.
Case 1 : Suppose that a, b are in two different cycles in σ. Without loss of generality (since we can just cyclically permute the elements of a cycle, and we can permute disjoint cycles), these cycles are $C_1 = (ac_1 \cdots c_m)$ and $C_2 = (bd_1 \cdots d_k)$. Then writing $\sigma = C_1 C_2 \cdots C_n$ as a product of disjoint cycles with lengths l_k, we get by Lemma 2.2.6 the following

$$\tau\sigma = (ab)(ac_1 \cdots c_m)(bd_1 \cdots d_k)C_3 \cdots C_n = (ac_1 \cdots c_m bd_1 \cdots d_k)C_3 \cdots C_n$$

This new cycle is disjoint from the other C_i, and has length $l_1 + l_2$. Thus, we get

$$\operatorname{sgn}(\tau\sigma) = (-1)^{(l_1 + l_2 - 1) + \sum_{i=3}^{n}(l_i - 1)} = (-1)^{1 + \sum_{i=1}^{n}(l_i - 1)} = -\operatorname{sgn}(\sigma)$$

as required.
Case 2 : Suppose that a, b are in the same cycle in σ. Without loss of generality, we write this cycle as $C_1 = (ac_1 \cdots c_m bd_1 \cdots d_k)$. Then writing $\sigma = C_1 \cdots C_n$ as a product of disjoint cycles with lengths l_k, we get by Lemma 2.2.6 the following

$$\tau\sigma = (ab)(ac_1 \cdots c_m bd_1 \cdots d_k)C_2 \cdots C_n = (ac_1 \cdots c_m)(bd_1 \cdots d_k)C_2 \cdots C_n$$

These two new cycles are disjoint from the other C_i, and have lengths $l_{11} + l_{12} = l_1$. Thus, we get

$$\operatorname{sgn}(\tau\sigma) = (-1)^{(l_{11} - 1) + (l_{12} - 1) + \sum_{i=3}^{n}(l_i - 1)} = (-1)^{\sum_{i=1}^{n}(l_i - 1) - 1} = -\operatorname{sgn}(\sigma)$$

as required. □

This proposition turns out to be quite useful in proving the following theorem.

Theorem 2.2.10. *For any permutations σ_1, σ_2,* $\operatorname{sgn}(\sigma_1\sigma_2) = \operatorname{sgn}(\sigma_1)\operatorname{sgn}(\sigma_2)$.

Proof. By Theorem 2.2.7, we may write $\sigma_1 = \tau_1 \cdots \tau_n$ as a product of transposition. Then by repeated application of Proposition 2.2.9 we get

$$\operatorname{sgn}(\sigma_1\sigma_2) = (-1)^n \sigma \operatorname{sgn}(\sigma_2) = \operatorname{sgn}(\sigma_1)\operatorname{sgn}(\sigma_2)$$

as required. □

This finally brings us to the full relation between transpositions and the sign of a permutation, which is an immediate consequence of Theorem 2.2.10.

Corollary 2.2.10.1. *Suppose a permutation σ can be written as the product of k transpositions. Then $\operatorname{sgn}(\sigma) = (-1)^k$. Furthermore, any representation of σ as a product of transpositions must have an even amount of transpositions if $\operatorname{sgn}(\sigma) = 1$, and an odd amount otherwise.*

2.3 Cosets and Quotient Groups

Definition 2.3.1. Let G be a group, and $H \subseteq G$ a subgroup. We define two equivalence relations $\sim_L, \sim_R$ by, for any $x, y \in G$

$$x \sim_L y \iff \exists h \in H \mid hx = y, x \sim_R y \iff \exists h \in H \mid xh = y$$

We call the equivalence classes in $G/\sim_L, G/\sim_R$ left and right cosets respectively. They are given more explicitly by, for any $x \in G$

$$xH = \{xh \mid h \in H\}, Hx = \{hx \mid h \in H\}$$

Note. While not proven in the definition, it is not too hard to prove that these are indeed equivalence relations.

What we're really after here is some unified notion of the quotient group, so we want to find out what left and right cosets have in common.

Lemma 2.3.2. *Let G be a group, and $H \subseteq G$ a subgroup. For any $x \in G$, $|xH| = |Hx| = |H|$.*

Proof. Consider the function $f : H \to xH$ given by $f(h) = xh$. This is surjective, and injective since $f(h_1) = f(h_2) \Rightarrow xh_1 = xh_2 \Rightarrow h_1 = h_2$ (by cancelling the x on both sides). Thus, it is bijective. The proof for Hx is identical. □

This gives an immediate corollary.

Corollary 2.3.2.1. *If G is finite, then for any subgroup $H \subseteq G$ there are the same number of left and right cosets.*

We use this to build a new definition.

Definition 2.3.3. Let G be a finite group and $H \subseteq G$ a subgroup. We denote the number of left/right cosets (xH or Hx) in G by $[G : H]$, and call it the index of H in G.

Which brings us to our first big result of the section.

Theorem 2.3.4 (Lagrange's Theorem)**.** *Let G be a finite group and $H \subseteq G$ a subgroup. Then $|G| = [G : H]|H|$.*

Proof. By lemma 2.3.2, each coset xH (we choose left cosets here for convenience) in G has size $|H|$. The result is then immediate from the definition of $[G : H]$. □

This has a couple of immediate corollaries.

Corollary 2.3.4.1. *Let $g \in G$ be any element of a finite group. Then $o(g) \mid |G|$.*

Proof. It suffices to note that $o(g) = |\langle g \rangle|$. □

Corollary 2.3.4.2. *Let $g \in G$ be any element of a finite group. Then $g^{|G|} = 1$.*

Proof. Since $o(g) \mid |G|$, $\exists k \in \mathbb{N}$ such that $o(g)k = |G|$. Thus

$$g^{|G|} = (g^{o(g)})^k = 1^k = 1$$

□

We now move on towards quotient groups, for which we really want left and right cosets to be identical.

Definition 2.3.5. A subgroup $H \subseteq G$ is called normal, denoted $H \trianglelefteq G$, if for all $x \in G$, $xH = xH$.

When $H \trianglelefteq G$, we often just refer to the coset of an element, and don't bother specifying left or right.

Theorem 2.3.6. *$H \trianglelefteq G$ if and only if for all $h \in H$ and $g \in G$, $ghg^{-1} \in H$.*

Proof. Suppose $H \trianglelefteq G$, and pick any $h \in H$. Then for any $g \in G$, $gh \in Hg$. Thus, $\exists h' \in H$ such that $gh = h'g \Rightarrow ghg^{-1} \in H$, as was to be shown. Now, suppose that $ghg^{-1} \in H$ for all $g \in G, h \in H$. Let $h' = ghg^{-1}$. Then $h'g = gh$, so $gh \in Hg$. Since this was for arbitrary g, h, this implies that $gH = Hg$, as was to be shown. □

Theorem 2.3.7. *Suppose $H \trianglelefteq G$. Then the set of cosets of G relative to H, with multiplication defined by $(xH)(yH) = (xy)H$, is a group.*

Proof. All the properties of group multiplication are inherited from G if this multiplication is well-defined, so we just need to check this. Suppose $xH = yH$ and $aH = bH$. Then since H is normal, $\exists h, h' \in H$ such that $hx = y, ah' = b$ (this is using the same technique as in the proof of Theorem 2.3.6, just taking one of the values in H to be 1). Since multiplying by elements of H doesn't change the coset, it follows that

$$(xH)(aH) = (xa)H = (xah')H = (xb)H = H(xb) = H(hxb) = H(yb) = (yb)H$$

completing the proof. □

Definition 2.3.8. The group from Theorem 2.3.7 is denoted G/H, and called the quotient group of G by H, or G mod H.

Note. When G/H is a quotient group, the canonical quotient map becomes a homomorphism.

2.4 Homomorphism Theorems

Again, this is a reworking of a similar section in [Jac09].

Definition 2.4.1. Let $\varphi : G_1 \to G_2$ be a group homomorphism. The kernel of φ is given by

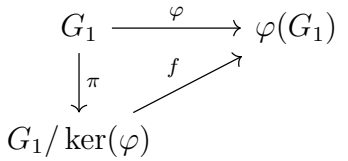
$$\ker(\varphi) = \{g \in G_1 \mid \varphi(g) = 1\}$$

Lemma 2.4.2. $\ker(\varphi) \trianglelefteq G_1$.

Proof. We know that $\varphi(1) = 1$, so this condition is satisfied. It remains to check closure under multiplication and the existence of multiplicative inverses. For the former, we note that if $x, y \in \ker(\varphi)$, then $\varphi(xy) = \varphi(x)\varphi(y) = 1$, so $xy \in \ker(\varphi)$. For the latter, we note that for any $x \in G_1$, $\varphi(xx^{-1}) = 1 = \varphi(x)\varphi(x^{-1})$, so $\varphi(x)^{-1} = \varphi(x^{-1})$. Thus, if $x \in \ker(\varphi)$, $\varphi(x^{-1}) = \varphi(x)^{-1} = 1^{-1} = 1$, so $x^{-1} \in \ker(\varphi)$. □

We denote the set of all homomorphisms between two groups G_1, G_2 by $\mathrm{Hom}(G_1, G_2)$. If the homomorphism is between a group and itself, we drop the second group in that notation. We denote all the isomorphisms by $\mathrm{Isom}(G_1, G_2)$.

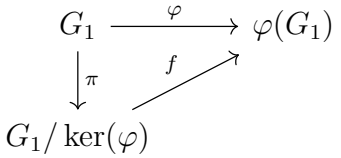
Theorem 2.4.3 (First Fundamental Theorem of Homomorphisms)**.** *Let G_1, G_2 be groups and pick any $\varphi \in \mathrm{Hom}(G_1, G_2)$. Let $\pi \in \mathrm{Hom}(G_1, G_1/\ker(\varphi))$ be the quotient map. Then there exists an isomorphism $f \in \mathrm{Isom}(G_1/\ker(\varphi), \varphi(G_1))$ which makes the following diagram commute.*



Proof. We proceed by directly constructing f. Pick any $g\ker(\varphi) \in G_1/\ker(\varphi)$. We define that $f(g\ker(\varphi)) = \varphi(g)$. This is clearly a homomorphism if it is well-defined, so we prove that it is in fact well-defined. Suppose that $x, y \in G$ are elements such that $x\ker(\varphi) = y\ker(\varphi)$. Then $\exists h \in \ker(\varphi)$ such that $xh = y$, so $\varphi(y) = \varphi(xh) = \varphi(x)\varphi(h) = \varphi(x)$, as required. f is by definition surjective onto $\varphi(G_1)$, so it remains only to prove that it is injective. Suppose that $f(x\ker(\varphi)) = f(y\ker(\varphi))$. Then $\varphi(x) = \varphi(y) \Rightarrow \varphi(x)\varphi(y)^{-1} = 1$, so $\varphi(xy^{-1}) = 1$. Then $xy^{-1} \in \ker(\varphi)$, so $x\ker(\varphi) = y\ker(\varphi)$, as was to be shown. □

Corollary 2.4.3.1. *Any $\varphi \in \mathrm{Hom}(G_1, G_2)$ is injective if and only if $\ker(\varphi) = \{1\}$.*

Proof. Suppose φ is injective. Then by definition, $G_1 \cong \varphi(G_1)$. But by Theorem 2.4.3, $\varphi(G_1) \cong G_1/\ker(\varphi)$, so $G_1 \cong G_1/\ker(\varphi)$. Thus, the projection map $\pi : G_1 \to G_1/\ker(\varphi)$ is injective, so $\pi^{-1}(\ker(\varphi)) = \ker(\varphi) = \{1\}$, as required. Now, suppose that $\ker(\varphi) = \{1\}$. Then $x\ker(\varphi) = y\ker(\varphi)$ if and only if $x = y$, so π is injective. By Theorem 2.4.3, the following diagram commutes



where f is an isomorphism, so it follows that φ is injective. □

Lemma 2.4.4. *The image or pre-image of a subgroup under a homomorphism is a subgroup.*

Proof. Let G, H be groups and $\varphi : G \to H$ a homomorphism. Let $K \subseteq G$ be a subgroup. Then $1 \in K$, so $\varphi(1) = 1 \in \varphi(K)$. Pick any $a, b \in \varphi(K)$. Then $\exists x, y \in K$ such that $\varphi(x) = a, \varphi(y) = b$, so $\varphi(xy) = ab \in \varphi(K)$, making it a subgroup of H. Suppose $K \subseteq H$ is a subgroup. Then $1 \in K$, so $1 \in \varphi^{-1}(K)$ as required. Suppose $x, y \in \varphi^{-1}(K)$. Then $\varphi(xy) = \varphi(x)\varphi(y) \in K$, so $xy \in \varphi^{-1}(K)$ making it a group. □

Theorem 2.4.5 (Second Fundamental Theorem of Homomorphisms)**.** *Let G_1, G_2 be groups and $\varphi : G_1 \to G_2$ a surjective homomorphism. Then*

1. *There exists a bijection between all subgroups of G_1 containing $\ker(\varphi)$ and all subgroups of G_2.*
2. *If $H \supseteq \ker(\varphi)$ is a subgroup of G_1, $H \trianglelefteq G_1$ if and only if $\varphi(H) \trianglelefteq G_2$.*
3. *If $\ker(\varphi) \subseteq H \trianglelefteq G_1$, then $G_1/H \cong G_2/\varphi(H)$.*

Proof. Let $\pi : G_1 \to G_1/\ker(\varphi)$ be the quotient map. Then by lemma 2.4.4, π maps the subgroups of G_1 onto all subgroups of $G_1/\ker(\varphi)$. By Theorem 2.4.3, $G_1/\ker(\varphi) \cong G_2$, so by lemma 2.4.4 the isomorphism $f : G_1/\ker(\varphi) \to G_2$ is a bijective map between subgroups of $G_1/\ker(\varphi)$ and subgroups of G_2. Since the pre-image of any subgroup of $G_1/\ker(\varphi)$ under π contains $\ker(\varphi)$, $f \circ \pi$ is a surjection between subgroups of G_1 containing $\ker(\varphi)$ and subgroups of G_2. Finally, suppose that H_1, H_2 are subgroups of G_1 containing the kernel and $\pi(H_1) = \pi(H_2)$. Then $H_1 = \{h \ker \varphi\}_{h \in H_1} = \{h \ker \varphi\}_{h \in H_2} = H_2$, so $f \circ \pi = \varphi$ is injective between the sets of subgroups, making it a bijection.
Let $H \supseteq G_1$ contain the kernel. Suppose it is normal. Pick any $h \in \varphi(H), g \in G_2$. Then $\exists g' \in G_1, h' \in H$ such that $\varphi(g') = g, \varphi(h') = h$. Thus, $ghg^{-1} = \varphi(g')\varphi(h')\varphi(g'^{-1}) = \varphi(g'h'g'^{-1})$. Since H is normal, $g'h'g'^{-1} \in H$, so $ghg^{-1} \in \varphi(H)$, making it normal. Suppose, conversely, that $\varphi(H)$ is normal. Pick any $g \in G_1, h \in H$. Note, by part (1) of this theorem, that $\varphi^{-1}(\varphi(H)) = H$. Thus, $\varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g)^{-1} \in \varphi(H)$ implies that $ghg^{-1} \in H$, as required.
Finally, suppose that $\ker(\varphi) \subseteq H \trianglelefteq G_1$. By part (2), we know that $\varphi(H) \trianglelefteq G_2$. Let $\pi' : G_2 \to G_2/\varphi(H)$ be the quotient map, and let $\psi = \pi' \circ \varphi$. Then by Theorem 2.4.5, $G_1/\ker(\psi) \cong G_2/\varphi(H)$. But $\ker(\psi) = \varphi^{-1}(\pi'^{-1}(1)) = \varphi^{-1}(\varphi(H)) = H$, so $G_1/H \cong G_2/\varphi(H)$. □

Definition 2.4.6. Let $H, K \subseteq G$ be subgroups. We define the product of the subgroups as

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

Lemma 2.4.7. *If $K \trianglelefteq G$, then HK is a subgroup of G.*

Proof. Since $1 \in H, K$, $1 \in HK$. Pick any $hk, h'k' \in HK$. Then

$$(hk)(h'k') = (hh')(h'^{-1}kh')k' = h''k''$$

for some $h'' \in H, k'' \in K$, as required. We just need to check inverses.

$$(hk)^{-1} = k^{-1}h^{-1} = h^{-1}(hk^{-1}h^{-1}) = h^{-1}k'$$

for some $k' \in K$, as required. □

Theorem 2.4.8 (Third Fundamental Theorem of Homomorphisms)**.** *Let G_1, G_2 be groups, $\varphi : G_1 \to G_2$ a surjective homomorphism, $K = \ker(\varphi)$, and $H \subseteq G_1$ any subgroup. Then*

$$\varphi(H) \cong \frac{HK}{K} \cong \frac{H}{H \cap K}$$

Proof. Let $\psi = \varphi_{\restriction HK}, \vartheta = \varphi_{\restriction H}$. Then since $K \subseteq HK$, $\psi^{-1}(1) = K$ and $\vartheta^{-1}(1) = K \cap H$. Furthermore, we can note that since $\psi(hk) = \phi(hk) = \phi(h)$, $\psi(HK) = \varphi(H)$. Thus, by Theorem 2.4.3 we have the following two commutative diagrams.

$$\begin{array}{ccc} HK & \xrightarrow{\psi} & \varphi(H) \\ \downarrow & \overset{\cong}{\nearrow} & \\ HK/K & & \end{array} \qquad \begin{array}{ccc} H & \xrightarrow{\vartheta} & \varphi(H) \\ \downarrow & \overset{\cong}{\nearrow} & \\ H/(K \cap H) & & \end{array}$$

as required. □

Note. Many books will put these theorems in a different order, give them different names, or add/remove conclusions from each. I do not claim that these are **the** authoritative correct fundamental theorems of homomorphisms.

2.5 Cyclic Groups

This section is essentially that of the same name presented in the first chapter of [Jac09], and aims to introduce properties of the simplest kind of groups; cyclic groups. I've done my best to rework it in a way that hopefully increases clarity.

Definition 2.5.1. A cyclic group is a group generated by a single element.

We refer to an element that generates a cyclic group as a *generator* of that group. In general, a cyclic group will have many possible generators.

Lemma 2.5.2. *If G is Abelian, every subgroup of G is normal.*

Proof. If $K \subseteq G$, $k \in K, g \in G$, then $gkg^{-1} = gg^{-1}k = k \in K$. □

Theorem 2.5.3. *Suppose G is cyclic. Then if G is infinite, $G \cong \mathbb{Z}$ (the additive group of integers) and otherwise $G \cong \mathbb{Z}/|G|\mathbb{Z}$ (the additive group of integers modulo $|G|$).*

Proof. Let $G = \langle g \rangle$. If G is infinite, we can define a homomorphism $\varphi : G \to \mathbb{Z}$ by $g \mapsto 1$. Since $\varphi(g^n) = n\varphi(g) = n$, φ is injective and surjective, so $G \cong \mathbb{Z}$. Suppose $|G| = n$, that is $o(g) = n$. Then we can define a homomorphism $\varphi : G \to \mathbb{Z}/n\mathbb{Z}$ by $\varphi(g) = [1]$. This is certainly surjective. $\varphi(g^r) = \varphi(g^m) \Rightarrow [r] = [m] \Rightarrow r = kn + m$ for some $k \in \mathbb{Z}$. But then $g^r = g^{kn+m} = g^m$, so φ is injective. Finally, we note that since $g^n = 1$, $g^m = g^{m \bmod n}$ for any $m \in \mathbb{Z}$, so this map is well-defined. Thus, $G \cong \mathbb{Z}/n\mathbb{Z}$. □

Combining this with the transitivity of being isomorphic, we get an immediate corollary.

Corollary 2.5.3.1. *Any two cyclic groups of the same order are isomorphic.*

Theorem 2.5.4. *Any subgroup of a cyclic group is cyclic. If the cyclic group is infinite, the set of all non-trivial subgroups is in bijection with $\mathbb{N}$. If $G = \langle g \rangle$, where $o(g) = n$, then there is one and only subgroup of order q for every $q \mid n$.*

Proof. Let $G = \langle g \rangle$. For any subgroup $H \subseteq G$, there exists some smallest $n > 0$ such that $g^n \in H$ (note we can assume $n > 0$ since if $n < 0$, we simply take its inverse). Any element of the form g^m, where $n \mid m$, can be generated by this element. Suppose $\exists g^m \in H$ such that $n \nmid m$. Write $m = kn + r$, where $k \in \mathbb{Z}, 0 < |r| < n$. Then $g^{m-kn} = g^r \in H$. But $r < n$, contradicting the minimality of n. Hence, $H = \langle g^n \rangle$, as was to be shown.
Now, suppose that $o(g) = \infty$. Then $\langle g^n \rangle \subseteq G$ is a subgroup for each $n \in \mathbb{N}$. Furthermore, all non-trivial subgroups are of this form, as if the generator is g^n for $n < 0$ we simply take its inverse, and $g^0 = 1$. It suffices then to show that each of these is unique. Suppose that $\langle g^n \rangle = \langle g^m \rangle$ for some $m, n \in \mathbb{N}$. Then there exists some k such that $g^{mk} = g^n \Rightarrow g^{mk-n} = 1$. There also exists some $r \in \mathbb{N}$ such that $g^{rn} = g^m \Rightarrow g^{rn-m} = 1$. Since g has infinite order, $mk - n, rn - m = 0 \Rightarrow n \mid m, m \mid n$, so $m = n$ as required.
Suppose $o(g) = n$, and pick any subgroup $H \subseteq G$. Let $m \in \mathbb{N}$ be the minimal number such that $g^m \in H$. From the above, we know that $H = \langle g^m \rangle$. Pick any $q \mid n$. For existence, it suffices to note that $o(g^{n/q}) = q$. For uniqueness, let $m \in \mathbb{N}$ be the minimal number such that $o(g^m) = q$. Suppose $o(g^k) = q$. Then $\exists a, b \in \mathbb{N}$ such that $qm = an, qk = bn$. Then $m = \frac{an}{q}, k = \frac{bn}{q}$. Since m is minimal, it follows that $a = 1$ (as $a = 1$ certainly works), so $g^k \in \langle g^m \rangle$, as required. □

Corollary 2.5.4.1. *Let G be a group, and $a \in G$ an element such that $o(a) < \infty$. Then $\langle a \rangle = \{g \in G \mid o(g) \mid o(a)\}$.*

Proof. Suppose $g \in \langle a \rangle$. Then $g = a^n$ for some $n \in \mathbb{N}$. Thus, $o(g) \mid o(a)$, as required. Now, suppose that $o(g) \mid o(a)$. By Theorem 2.5.4, there is exactly one group of order $o(g)$, and g must generate this group. But $o(a^{o(a)/o(g)}) = o(g)$, so it follows that this group lies in $\langle a \rangle$, and in particular $g \in \langle a \rangle$. □

Lemma 2.5.5. *Let a, b be elements of an abelian group of finite orders $o(a) = n, o(b) = m$ such that $(n, m) = 1$. Then $\langle a \rangle \cap \langle b \rangle = \langle 1 \rangle$, $\langle a, b \rangle = \langle ab \rangle$, and $o(ab) = nm$.*

Proof. Suppose $x \in \langle a \rangle \cap \langle b \rangle$. Then $o(x) \mid n, m$, so $o(x) = 1 \Rightarrow x = 1$ as required. Since the group is Abelian, $(ab)^k = a^k b^k$ for any $k \in \mathbb{N}$. In particular, if $k \in \mathbb{N}$ we get $a^k = b^{-k} \in \langle a \rangle \cap \langle b \rangle$, so $a^k = b^{-k} = 1 \Rightarrow n, m \mid k$. The smallest such k is nm, and $(ab)^{nm} = 1$, so $o(ab) = nm$. Finally, note that we can write every $x \in \langle a, b \rangle$ in the form $x = a^r b^q$, where $0 \leq r < n, 0 \leq q < m$. It follows that $|\langle a, b \rangle| \leq nm$. But $\langle ab \rangle \subseteq \langle a, b \rangle$, so $\langle a, b \rangle = \langle ab \rangle$. □

Lemma 2.5.6. *If G is a finite Abelian group, then it contains an element whose order is divisible by the order of every element of G.*

Proof. Since G is finite, it suffices to show that we can do this with any two elements. Let $a, b \in G$, and take the prime decomposition of both orders

$$o(a) = n = p_1^{e_1} \cdots p_r^{e_r} \cdots p_l^{e_l} \qquad o(b) = m = p_1^{f_1} \cdots p_r^{f_r} \cdots p_l^{f_l}$$

We order things such that $e_i \geq f_i$ for $i \leq h$, and $f_i \geq e_i$ for $i > h$. We can see then that

$$[n, m] = p_1^{e_1} \cdots p_r^{e_r} p_{r+1}^{f_{r+1}} \cdots p_l^{f_l}$$

Let $q = p_1^{f_1} \cdots p_r^{f_r}, s = p_{r+1}^{e_{r+1}} \cdots p_l^{e_l}$. Then $[n, m] = \frac{n}{s}\frac{m}{q}$, and $(n/s, m/q) = 1$. $o(a^s) = n/s, o(b^q) = m/q$, so by lemma 2.5.5 $o(a^s b^q) = [n, m]$, as required. □

Theorem 2.5.7. *Let G be a finite Abelian group. Then G is cyclic if and only if $|G|$ is the smallest positive integer n such that $a^n = 1$ for all $a \in G$.*

Proof. Suppose G is cyclic, with generator g. Then $o(g) = |G|$, so by Lagrange's theorem $a^{|G|} = 1$ for any $a \in G$. Since $o(g) = |G|$, this is the smallest such integer. Now, suppose that $|G|$ is the smallest positive integer n such that $a^n = 1$ for all $a \in G$. By lemma 2.5.6, $\exists g \in G$ whose order is divisible by the order of every element in G. Then $a^{o(g)} = 1$ for any $a \in G$, and $o(g) \leq |G|$, so $o(g) = |G|$. Thus, $G = \langle g \rangle$, completing the proof. □

2.6 Group Actions

This section is based on similar sections in [Jac09] and [Lan05]. In it, we introduce one of the most important applications of groups in mathematics, *group actions*.

Definition 2.6.1. An action of a group G on a set S is a homomorphism $f \in \mathrm{Hom}(G, \mathrm{Sym}(S))$.

If a group G acts on a set S, we call S a G-set. Before looking at examples, let's prove a result that makes the reasoning for calling this a group action more clear.

Theorem 2.6.2. *Let G be a group and S a set. S is a G-set if and only if there exists a map $\cdot : G \times S \to S$ satisfying the following axioms for any $x \in S$, $g, h \in G$*

1. $1 \cdot x = x$
2. $(gh) \cdot x = g \cdot (h \cdot x)$

Proof. Suppose S is a G-set. Then there exists a homomorphism $\varphi : \mathrm{Hom}(G, \mathrm{Sym}(S))$, in which case we simply define that $g \cdot x = \varphi(g)(x)$ (one can check that this has the desired properties). Now, suppose that the map $\cdot$ exists. Then $x \mapsto g \cdot x$ is a permutation of S, and $\varphi(g) = (x \mapsto g \cdot x)$ is the desired homomorphism. □

Note. The above version of a group action is also called a *left group action*, with *right group actions* being similar but with a map $\cdot S \times G \to S$. These are essentially identical to left group actions, so we won't be bothering with them here. They are occasionally a clearer notation though.

Example 2.6.1. Given any group G and subgroup $H \subseteq G$, G acts on G/H by $g \cdot (kH) = (gk)H$ (for any $g, k \in G$).

A group action is *effective* if the map $\varphi \in \mathrm{Hom}(G, \mathrm{Sym}(S))$ is injective. It is *faithful* if $g \cdot x$ for all $x \in S$ implies that $g = 1$, and *free* if $g \cdot x$ for some $x \in S$ implies that $g = 1$.

The rest of this section is admittedly a little haphazard, exploring the varied directions of inquiry we could take with group actions. Let's start by figuring out how to move between group actions.

Definition 2.6.3. Let X, Y be G-sets. A map $f : X \to Y$ is called a morphism of G-sets if for all $g \in G, x \in X$, $f(g \cdot x) = g \cdot f(x)$.

Two G-sets are said to be *isomorphic* or *equivalent* if there exists an invertible morphism between them, whose inverse is also a morphism of G-sets.

Let X be a G-set. The G-orbit of an element $x \in S$ is $Gx = \{g \cdot x \mid g \in G\}$. Its clear that S can be partitioned into disjoint G-orbits. A G-set S is called *transitive* (or the action of G on S called transitive) if S has exactly one G-set.

Definition 2.6.4. The stabilizer of an element $x \in X$ of a G-set X is defined as

$$\mathrm{Stab}_G(x) = \{g \in G \mid g \cdot x = x\}$$

Lemma 2.6.5. *For any $x \in X$, $Stab_G(x)$ is a subgroup of G.*

Proof. Suppose $a, b \in \mathrm{Stab}_G(x)$. Then $(ab) \cdot x = a \cdot (b \cdot x) = a \cdot x = x$, so $ab \in \mathrm{Stab}_G(x)$. $1 \in \mathrm{Stab}_G(x)$ is clear. Finally, we get

$$x = 1 \cdot x = (a^{-1}a) \cdot x = a^{-1} \cdot (a \cdot x) = a^{-1} \cdot x$$

so $a^{-1} \in \mathrm{Stab}_G(x)$. □

Using this, we can get a very strong result on transitive group actions.

Theorem 2.6.6. *Suppose X is a transitive G-set. For any $x \in S$, set $H = Stab_G(x)$. Then the action of G on X is equivalent to the left action of G on G/H, as given in example 2.6.1.*

Proof. Pick any $x \in X$. We'll define $f : X \to G/H$ by $f(g \cdot x) = gH$, for any $g \in G$ (note that this only works since X is transitive). First, we need to show that this is well-defined. Suppose $a, b \in G$ are such that $a \cdot x = b \cdot x$. Then $\exists g \in G$ such that $b = ag$, so $(ag) \cdot x = a \cdot (g \cdot x) \Rightarrow g \cdot x = 1 \cdot x = x$, and thus $g \in H$ and $aH = bH$. Next, we show that this is a morphism of G-sets. Pick any $x \in X, g \in G$. Then we get

$$f(g \cdot x) = gH = g \cdot (1H) = g \cdot f(1 \cdot x) = g \cdot f(x)$$

as required. Finally, we show that it is bijective. That f is surjective is clear from the definition. Suppose that f were not injective. Then $\exists a, b \in G$ such that $a \cdot x \neq b \cdot x$ but $aH = bH$. Thus, $\exists h \in H$ such that $b = ah$, so $b \cdot x = (ah) \cdot x = a \cdot (h \cdot x) = a \cdot x$, which is impossible. f is therefore bijective, as was to be shown. □

This leads immediately to a very useful corollary.

Corollary 2.6.6.1. *If X is a transitive G-set, then $|X| = [G : Stab_G(x)]$ for any $x \in X$.*

This is actually more powerful than it looks at first glance for group actions on finite sets. Suppose X is a finite G-set. Then for any $x \in X$, we can regard Gx as a finite G-set by restricting the action on X to Gx. It's clear in this case that G acts transitively on Gx. Furthermore, if X is finite, then it divides up into finitely many disjoint G-orbits. This, combined with corollary 2.6.6.1, gives us the following important result.

Theorem 2.6.7. *Let X be a finite G-set, and $\bigsqcup_{i=1}^{n} Gx_i = X$ be a decomposition of X into disjoint G-orbits. Then*

$$|X| = \sum_{i=1}^{n} [G : Stab_G(x_i)]$$

Proof. Since the Gx_i are disjoint, $|X| = \sum_{i=1}^{n} |Gx_i|$, and by corollary 2.6.6.1 we know that since G acts on Gx_i transitively, $|Gx_i| = [G : \mathrm{Stab}_G(x_i)]$. □

We now take a quick detour into the world of *primitive group actions*.

Definition 2.6.8. Let X be a G-set, and $\pi(X)$ a partition of X. We say that $\pi(X)$ is stabilized by the G-action if $g \cdot Y \in \pi(X)$ for all $g \in G, Y \in \pi(X)$.

Note. There are three partitions of a set X which will always have this property. The partition X, the partition $\{x\}_{x \in X}$ and the partition of X into G-orbits.

Definition 2.6.9. Let X be a G-set. We say that G acts primitively on X if the only two partitions of X stabilized by G are X and $\{x\}_{x \in X}$. Otherwise, we say it acts imprimitively.

Note. Since the partition of X into G-orbits is always stabilized by G, G acts primitively only if it acts transitively.

The next two results have proofs taken directly from [Jac09].

Lemma 2.6.10. *G acts imprimitively on a set X if and only if $\exists A \subsetneq X$ such that $|A| > 1$ and for any $g \in G$, either $g \cdot Y = Y$ or $(g \cdot Y) \cap Y = \emptyset$.*

Proof. Suppose there exists $Y \subsetneq X$ meeting the above conditions. Then for any $g_1, g_2 \in G$, $g_1 \cdot Y$ and $g_2 \cdot Y$ are either equal or disjoint. Let $Z = X \backslash (\bigcup_{x \in X} g \cdot Y)$. Then $g_1 \cdot B \cap g_2 \cdot Y = \emptyset$ for all $g_1, g_2 \in G$, so $g_1 \cdot B = B$ for all $g_1 \in G$. Thus, the set of all distinct subsets of the form $g \cdot Y$, along with B, forms a partition of X stabilized by G, making the action of G imprimitive. Now, suppose that G acts imprimitively on X. Then there exists a partition $\pi(X)$ which G stabilizes, which must contain some $Y \in \pi(X)$ such that $|Y| > 1, Y \subsetneq X$. Since this partition is stabilized, it follows that for any $g \in G$, either $g \cdot Y = Y$ or $(g \cdot Y) \cap Y = \emptyset$. □

Theorem 2.6.11. *Suppose G acts transitively on a set X, where $|X| > 1$. Then G acts primitively if and only if, for any $x \in X$, $Stab_G(x)$ is a maximal subgroup of G (i.e. there exists no group H such that $Stab_G(x) \subsetneq H \subsetneq G$).*

Proof. Suppose $\exists x \in X$ such that $\text{Stab}_G(x)$ is not maximal, and let H be a subgroup such that $\text{Stab}_G(x) \subsetneq H \subsetneq G$. Since G acts transitively, we get by Theorem 2.6.6 that the G-action on X is equivalent to the G-action on $G/\text{Stab}_G(x)$, and thus G acts imprimitively on X if and only if it acts imprimitively on $G/\text{Stab}_G(x)$. Let Y be the set of cosets of the form $h\text{Stab}_G(x)$, where $h \in H$. Since $\text{Stab}_G(x) \subsetneq H \subsetneq G$, $|Y| > 1$ and $Y \neq G/\text{Stab}_G(x)$. It's also clear that $h \cdot Y = Y$ for any $h \in H$. If $g \notin H$, then since $gh \notin H$ for any $h \in H$ we have that $(g \cdot Y) \cap Y = \emptyset$. Thus, by lemma 2.6.10, G acts imprimitively on $G/\text{Stab}_G(x)$ and hence on X.

Now, suppose that G acts transitively and imprimitively on X. Then by lemma 2.6.10, there exists a proper subset $Y \subsetneq X$ such that $|Y| > 1$, and for any $g \in G$, either $g \cdot Y = Y$ or $(g \cdot Y) \cap Y = \emptyset$. Let $H = \{h \in G \mid h \cdot Y = Y\}$. H is clearly a subgroup of G which contains $\text{Stab}_G(x)$ for any $x \in Y$, since for any $g \in G$ we get $g \cdot x = x$ implies that $(g \cdot Y) \cap Y \neq \emptyset$, so $g \cdot Y = Y$. Since $Y \neq X$ and g acts transitively, $\exists g \in G$ such that $g \cdot Y \neq Y$. Thus, $g \notin H$, so $H \neq G$. Finally, let $x, y \in Y$ be distinct elements. Then $\exists g \in G$ such that $g \cdot x = y$. Thus, $g \in H, g \notin \text{Stab}_G(x)$, so $H \neq \text{Stab}_G(x)$, making $\text{Stab}_G(x)$ not a maximal subgroup of G. □

Next, we look at possibly the most important kind of group action; *conjugation*.

Definition 2.6.12. Let G by any group. The action of G on itself by conjugation is given by, for any $g, x \in G$, $g \cdot x = gxg^{-1}$. The orbits of this action are called the *conjugacy classes* of G.

Example 2.6.2. The conjugacy classes of S_n are all the disjoint cycle decompositions of the same time, in the sense that two cycles are conjugate if and only if their decomposition has the same number of disjoint cycles of each size.

We also bring up now the centralizer $C(S)$ of a subset S of a group G, which is the set of all elements which commute with every element in G. This can be found to be a subgroup of G. $C(G) = C$ is called the centre of a group. For action by conjugation, it's clear that $\text{Stab}_G(x) = C(x)$. Thus, we get by Theorem 2.6.7 that for finite groups

$$|G| = \sum_{i=1}^{n} [G : C(x_i)]$$

where $x_i \in G$ are representatives of the conjugacy classes of G. This is called the *class equation of a finite group*. We can also note that $C(x) = |G|$ and the conjugacy class of x is x for elements in C, so this is also often written as

$$|G| = |C| + \sum_{i=1}^{n} [G : C(y_i)]$$

where y_i are representatives of the conjugacy classes which are not C. We use this now to do a cute little proof which is often useful when working with finite groups.

Theorem 2.6.13. *Any finite group G of prime power order has a non-trivial centre.*

Proof. Let $|G| = p^n$, and let $y_i \in G$ be the representatives for the conjugacy classes other than C. We know that $p \mid |G|$, and $p \mid |C(y_i)|$ for each y_i (as $|C(y_i)| \neq 1$ and $C(y_i)$ is a subgroup of G). Thus, by the class equation we must get $p \mid |C| \Rightarrow |C| \neq 1$. □

2.7 Free Groups*

This section will be much more informal than the rest, as a formal treatment of free groups requires delving into a level of category theory that is best left for a more advanced course in algebra. For a formal treatment, see [Lan05].

Definition 2.7.1. Let S be an arbitrary set. The free group on S, denoted F_S, is the group of finite strings of the elements x and x^{-1} for $x \in S$, along with the unit string 1. Multiplication is concatenation of strings, with the rule that $xx^{-1} = 1$.

Example 2.7.1. The free group on two elements, F_2 is the set of all strings of the letters a, b, a^{-1}, b^{-1}. In it, we'd have that $abb^{-1}a^{-1} = 1$, but this group is not commutative and $aba^{-1}b^{-1} \neq 1$ cannot be simplified.

Free groups are often used to specify arbitrary finitely generated groups in a simple manner. To show how this is done, we need the following definition.

Definition 2.7.2. Let G be a group, and $S \subseteq G$ a subset. The *normal subgroup generated by* S, denoted $\langle S \rangle_N$, is the intersection of all normal subsets of G containing S, or equivalently the smallest normal subgroup of G containing S.

There is unfortunately no simple way to write elements of a normal subgroup generated by a set, as there was for the regular subgroup generated by a set. However, there is a simple way to write elements of $G/\langle S \rangle_N$.

Definition 2.7.3. Let S be sets, and R a subset of F_S. The group with generators S and relations R is the free group S, with the addition rule enforced that any string in R is equal to 1. The group presentation of this group is $\langle S \mid R \rangle$.

Theorem 2.7.4. $\langle S \mid R \rangle \cong \frac{F_S}{\langle R \rangle_N}$.

Proof. We start with the natural surjective homomorphism $\varphi : F_S \to \langle S \mid R \rangle$. Then clearly $R \subseteq \ker(\varphi)$, so $\langle R \rangle_N \subseteq \ker(\varphi)$. Let $J = \ker(\varphi) \backslash \langle R \rangle$. Then $\varphi(r) = 1$ for any $r \in J$, so adding J to R wouldn't change $\langle S \mid R \rangle$. Thus, we must get $\ker(\varphi) = \langle R \cup J \rangle_N = \langle R \rangle_N$, so by the first fundamental theorem of homomorphisms $\langle S \mid R \rangle \cong \frac{F_S}{\langle R \rangle_N}$ as claimed. □

Example 2.7.2. The presentation of the group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is $\langle a, b \mid a^2, b^2, aba^{-1}b^{-1} \rangle$.

The final things I'll mention here is a construction called the *amalgamated product*.

Definition 2.7.5. Let G, H, K be groups. The free product $G * J$ of two groups is $F_{G,H}$, with simplification by in-group multiplication in G or H also enforced. Let $\varphi \in \text{Hom}(K, G), \psi \in \text{Hom}(K, H)$. Then the amalgamated product of G, H over K with respect to φ, ψ, which is denoted $G *_K H$, is given by

$$G *_K H = \frac{G * H}{\langle \{\varphi(k)\psi(k)^{-1}\}_{k \in K} \rangle_N}$$

All of these constructions are quite important in algebraic topology, but won't be used much by us for the remainder of this text.

2.8 Sylow's Theorems

This section focuses on Sylow's theorems, an important tool in classifying and understanding the structure of finite groups. The structure and proofs in the section are primarily based on those in [Lan05], with some inspiration taken from [Jac09]. We open, however, with a quick detour to talk about exponents.

Definition 2.8.1. Let G be a group. The *exponent* of G, denoted $\exp(G)$, is the minimal number $n \in \mathbb{N}$ such that $g^n = 1$ for all $g \in G$. If no such number exists, we write that $\exp(G) = \infty$.

Lemma 2.8.2. *Let G be a group with $\exp(G) = n < \infty$. Then given any $g \in G$, $o(g) \mid n$.*

Proof. By definition, $o(g) \leq n$. Suppose that $o(g) \nmid n$. Then $\exists k \in \mathbb{N}, 0 < \ell < o(g)$ such that $n = ko(g) + \ell$. Thus, since $g^n = 1$, we get that $g^\ell = 1$. But $\ell < o(g)$ is non-zero, so this is impossible. □

Lemma 2.8.3. *Let G be a finite Abelian group with $\exp(G) = n < \infty$. Then $|G| \mid n^k$ for some $k \in \mathbb{N}$.*

Proof. We proceed by induction on $|G|$, assuming that $\exp(G) = n$. If $|G| = \exp(G) = n$, then the result is trivial. Suppose that the result holds for all groups G with $n \leq |G| < m$. Let G be a group of size m, with $\exp(G) = n$. Pick any non-trivial $h \in G$. By lemma 2.8.2, $o(h) \mid n$, so $H = \langle h \rangle$ has an order which divides n. Since G is Abelian, $H \trianglelefteq G$, so G/H is again an Abelian group and $\exp(G/H) \mid n$. Thus, by induction we get that $|G/H| \mid n^k$ for some $k \in \mathbb{N}$, so by Lagrange's theorem $m = [G : H]|H| \mid n^{k+1}$, completing the proof. □

Using these results, we can start to work on Sylow's theorems.

Definition 2.8.4. Let G be a finite group, and p a prime which divides $|G|$. A subgroup $H \subseteq G$ is called a p-subgroup if $|H| = p^n$ for some $n \in \mathbb{N}$, and a p-Sylow subgroup if this n is the maximal natural number such that $p^n \mid |G|$.

Lemma 2.8.5. *Let G be a finite Abelian group, and p a prime such that $p \mid |G|$. Then there exists an element of G of order p.*

Proof. Again, we proceed by induction on $|G|$. The result is clear for $|G| \leq p$. Suppose it holds for all $|G| < n$, where $p < n$, and that $|G| = n$. By lemma 2.8.3, $n \mid \exp(G)^k$ for some $k \in \mathbb{N}$, so $p \mid \exp(G)$. Since $\operatorname{lcm}\{o(g) \mid g \in G\} \mid \exp(G)$, it follows that there exists some $g \in G$ such that $p \mid o(g)$. If G is cyclic, the result is then clear, and otherwise the result then holds by induction on $H = \langle g \rangle$. □

Theorem 2.8.6 (Sylow I)**.** *Let G be a finite group, and p a prime number dividing the order of G. Then there exists a subgroup $H \subseteq G$ of order p^k for each $k \in \mathbb{N}$ such that $p^k \mid |G|$. In particular, G has a p-Sylow subgroup.*

Proof. Again, we proceed by induction on $|G|$. If $|G| \leq p$ then the result is trivial. Suppose the result holds for all $|G| < n$, where $n > p$, and that $|G| = n$. If there exists some subgroup $H \subsetneq G$ such that $p \nmid [G : H]$, then the result is immediate by induction, since $|H| < |G|$. Thus, we may assume that $p \mid [G : H]$ for all subgroups $H \subsetneq G$. From the class equation, we know that

$$n = |C(G)| + \sum_{i=1}^{n} [G : C(x_i)]$$

where the x_i are representatives of the non-trivial conjugacy classes of G. Since $p \mid [G : C(x_i)]$ for each x_i, it follows that $p \mid |C(G)|$, that is G has a non-trivial centre. Then by lemma 2.8.5, there exists an element $g \in C(G)$ of order p. If $p^k \mid |G|$ only for $k = 1$, we're done. Otherwise, we note that $H = \langle g \rangle \trianglelefteq G$, so G/H is a group of order $\frac{|G|}{p}$. Let N be the maximal number such that $p^N \mid |G|$ By induction, we can find a subgroup $K' \in G/H$ of order p^k for $1 \leq k < N$. Let $\varphi : G \to G/H$ be the canonical quotient map, and $K' = \varphi^{-1}(K)$. Then by the first fundamental theorem of homomorphisms $\frac{K'}{H} \cong K$, so by Lagrange's theorem $|K'| = |K||H| = p^{k+1}$, completing the proof. □

To prove the second Sylow theorem, we need one more lemma.

Lemma 2.8.7. *Let G be a p-group acting on a finite set X. Then the number of $x \in X$ such that $Stab_G(x) = G$ is equivalent to $|X|$ modulo p.*

Proof. Let S be the set of $x \in X$ such that $\operatorname{Stab}_G(x) = G$. Then by Theorem 2.6.7

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

where the x_i are representative of all the other G-orbits. Since G is a p-group, $p \mid [G : \operatorname{Stab}(x_i)]$ for each x_i, which immediately gives the desired result. □

Theorem 2.8.8 (Sylow's Theorem II)**.** *Let G be a finite group, and p a prime such that $p \mid |G|$. Then the following all hold.*

1. *Every p-subgroup of G is contained in a p-Sylow subgroup of G.*
2. *Every p-Sylow subgroup of G is conjugate, that is if P_1, P_2 are p-Sylow subgroups then $\exists g \in G$ such that $gP_1g^{-1} = P_2$.*

3. The number of p-Sylow subgroups of G is equivalent to one modulo p.

Proof. Consider the action of G on the set Γ of p-Sylow subgroups by conjugation. We give a special name to $\mathrm{Stab}_G(P)$, where $P \in \Gamma$; we call it the normalizer of P and denote it $N(P)$. Let H be a p-subgroup. First, let's suppose that $H \subseteq N(P)$. Then $HP \subseteq N(P)$ and $P \trianglelefteq HP$, so by the third fundamental theorem of homomorphisms

$$\frac{HP}{P} \cong \frac{H}{H \cap P}$$

Thus, if $HP \neq P$ then HP is a p-subgroup with order larger than P, which is impossible. Hence, $HP = P \Rightarrow H \subseteq P$, as was required for (1). We move now to a more general case. Let H act on the set of all conjugate subgroups of P, called S, by conjugation. We can see that $|S| = \frac{|N(P)|}{|P|}$, so $p \nmid |S|$. Thus, it follows by lemma 2.8.7 that there exists at least one element of $gPg^{-1} \in S$ which is unchanged by any element of H, meaning $H \subseteq gPg^{-1}$ and hence by the previous case $H \subseteq gPg^{-1}$. This completes the proof of (1). If we take H to be a p-*Sylow* subgroup, then $|H| = |P|$, so we instead get that $H = gPg^{-1}$ for some $g \in G$, giving (2). For (3), we note that since a general p-subgroup $H \subset gPg^{-1}$ is contained in some p-Sylow subgroup, and so no other p-Sylow subgroup is unchanged by every element of H under conjugation. Therefore, there is exactly one element of S has $\mathrm{Stab}_G(x) = G$, so by lemma 2.8.7 we get $|S| \equiv 1 \bmod p$, proving (3). □

We'll end this section off by taking a look at a simple yet very powerful application of Sylow's theorems : the structure theorem of finite Abelian groups. First, we need to introduce the concept of direct products.

Definition 2.8.9. Let G_1, G_2 be two groups. The *direct product* of the groups, denoted $G_1 \times G_2$ is the group whose set is $G_1 \times G_2$ (the Cartesian product) and operation element-wise multiplication in the two groups.

We'll leave it to the reader to show that this is indeed a group, and that the direct product of groups is associative and commutative up to isomorphism. Indeed, taking the direct product of an arbitrary number of groups is simply taking the Cartesian product of those groups with element-wise multiplication. Before moving on, we do need one more result on direct products.

Theorem 2.8.10. *Let G be a group and $\{H_i\}_{i=1}^n$ a finite collection of subgroups of G. Then $G \cong H_1 \times \cdots \times H_n$ if the following two conditions are met.*

1. *Every element of G can be expressed uniquely in the form $h_1 \cdots h_n$, where $h_i \in H_i$*
2. *Every element of H_i commutes with every element of H_j, for all $1 \leq i, j \leq n, i \neq j$*

Proof. Suppose $\{H_i\}_{i=1}^n$ is a collection of subgroups meeting the above conditions. The first condition gives us a natural bijective map $\varphi : H_1 \times \cdots \times H_n \to G$ defined by $\varphi : (h_1, \ldots, h_n) \mapsto h_1 \cdots h_n$. All that remains is to check that this is a homomorphism. Let $(h_1, \ldots, h_n), (h'_1, \ldots, h'_n) \in H_1 \times \cdots \times H_n$. Then

$$\varphi((h_1, \ldots, h_n)(h'_1, \ldots, h'_n)) = \varphi((h_1h'_1, \ldots, h_nh'_n)) = (h_1h'_1) \cdots (h_nh'_n)$$

By condition (2), all the elements on the right-hand side commute, so

$$\varphi((h_1,\dots,h_n)(h_1',\dots,h_n')) = (h_1h_2\cdots h_n)(h_1'\cdots h_n') = \varphi((h_1,\dots,h_n))\varphi((h_1',\dots,h_n'))$$

as required. □

Theorem 2.8.11 (Structure Theorem of Finite Abelian Groups)**.** *Let G be a finite Abelian group. Then there exist some $p_i, e_i \in \mathbb{N}$ such that*

$$G \cong \mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_n^{e^n}\mathbb{Z}$$

where each p_i is a prime such that $p_i^{e_i} \mid |G|$. Furthermore, this decomposition is unique up to the order of terms.

Proof. We start by proving existence. First, suppose that $|G| = p^n$, for some prime p. We proceed by induction on n. Suppose that the theorem holds for all $|G| = p^m$, where $m < n$. Let $h \in G$ be an element of maximal order p^N, and $H = \langle h\rangle$. We have by induction that $G/H \cong \langle g_1H\rangle \times \cdots \times \langle g_rH\rangle$, where $g_i \in G\backslash H$ and the order of each group in the product is a power of p. We'll first show that we can choose $g_i \in \langle g_iH\rangle$ such that $o(g_i) = |\langle g_iH\rangle|$. We know that there exists some minimal $\ell \in \mathbb{N}$ such that $g_i^{p^\ell} \in H$, and a minimal $q \in \mathbb{N}$ such that $g_i^{p^\ell} = h^{zp^q}$, where $z \in \mathbb{N}$ and $1 \le z < p$. Suppose $o(g_i) = p^r$. Then clearly $r \ge \ell$, and we get

$$1 = g_i^{p^r} = (g^{p^\ell})^{p^{r-\ell}} = h^{zp^{q+r-\ell}}$$

Thus, $q + r - \ell = N$ so $o(g_i) = p^{\ell+N-q}$, and by the maximality of N and Sylow II we know that $q \ge 1, \ell$. We want for $o(g_i) = p^\ell$. This is simply achieved by multiplying g_i by $h^{-zp^{q-\ell}}$, as since g_iH generates $\langle g_iH\rangle$ we know that $o(g_ih) \ge p^\ell$ for any $h \in H$, and one can verify that $o(g_ih^{-zp^{q-\ell}}) \le p^\ell$. Thus, we may assume that $o(g_i) = |\langle g_iH\rangle|$. In this case, we define a map $\varphi : H \times G/H \to G$ by the rule

$$\varphi : (h^q, (g_1^{e_1}H, \dots, g_r^{e_r}H)) \mapsto h^q g_1^{e_1}\cdots g_r^{e_r}$$

We'll show that this is an isomorphism, which will complete the proof in this case. This map is bijective if it is surjective by a simple counting argument, since $|G| = |H|[G : H]$. To check surjectivity, pick any $x \in G$. Then there exists some $y \in H$ and $e_1, \dots, e_r \in \mathbb{N}$ such that $x = yg_1^{e_1}\cdots g_n^{e_n}$. Since H is cyclic, y is a power of h, and thus we get the desired surjectivity. We just then need to show that it's a homomorphism. Indeed, since $o(g_i) = |\langle g_iH\rangle|$ we get

$$\begin{aligned}\varphi((h^q, (g_1^{e_1}H, \dots, g_r^{e_r}H))(h^\ell, (g_1^{x_1}H, \dots, g_r^{x_r}H))) &= \varphi((h^{q+\ell}, (g_1^{e_1+x_1}H, \dots, g_r^{e_r+x_r}H)))\\ = h^{q+\ell}g_1^{e_1+x_1}\cdots g_r^{e_r+x_r} &= \varphi((h^q, (g_1^{e_1}H, \dots, g_r^{e_r}H)))\varphi((h^\ell, (g_1^{x_1}H, \dots, g_r^{x_r}H)))\end{aligned}$$

as required.

Next, we prove existence in the general case. Since G is Abelian, we get by Sylow II that there exists a unique p_i-Sylow subgroup P_i of G for each prime $p_i \mid |G|$. By the above case, it suffices to prove that $G \cong P_1 \times \cdots \times P_n$. By Theorem 2.8.10, since G is Abelian it suffices for this to prove that each element of G can be uniquely expressed as a product of one element

from each P_i. Indeed, since $\prod_{i=1}^{n} |P_i| = |G|$, uniqueness comes for free with existence of the representation as a product. Again, we can proceed by induction on $|G|$. Suppose this holds for groups of size less than G. Since G is Abelian, $P_1 \trianglelefteq G$, so G/P_1 is an Abelian group with p-Sylow subgroups being the images of $P_2, \ldots, P_n$ under the quotient map. Pick any $g \in G$. Then there exists $h \in G \backslash P_1, k \in P_1$ such that $g = hk$. By the inductive hypothesis, there exists $h_2 \in P_2, \ldots, h_n \in P_n$ such that $hP_1 = (h_2 \cdots h_n)P_1$. Thus, $\exists h_1 \in P_1$ such that $g = h_1 \cdots h_n$, as required. This completes the proof of existence.

Finally, we prove uniqueness. Suppose $G \cong \mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_n^{e_n}\mathbb{Z}$, with isomorphism $\varphi : \mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_n^{e_n}\mathbb{Z} \to G$. By Sylow I, $\varphi(\mathbb{Z}/p_i^{e_i}\mathbb{Z})$ is contained in the p_i-Sylow subgroup, so it suffices to show uniqueness of the decomposition of each p_i-Sylow subgroup. Hence, we may assume that $|G| = p^k$ for some prime p, and that $p_1 = p_2 = \cdots = p_n = p$. In this case, let N be the number of distinct cyclic subgroups of G of order $\exp(G) = p^m$. It follows immediately that $e_i \leq m$, and furthermore that each $e_i = m$ in our decomposition corresponds to a distinct such cyclic subgroup. In fact, one can also see that if j is the number of $e_i = m$, then the number of distinct cyclic subgroups of order p^m in our decomposition is

$$\sum_{f=1}^{j} (p^m - p^{m-1})(p^{k-fm}) = N$$

giving a unique solution for j (as $\sum_{f=1}^{j} (p^m - p^{m-1})(p^{k-fm})$ strictly increases with j). We can then divide out all the terms of the form $\mathbb{Z}/p^m\mathbb{Z}$ (or their images) from both sides, and repeat this argument to get the desired result. □

Note. This clever proof is certainly not my own, but I forgot to write down the source when I initially wrote it up and was never able to find it again. I've provided here another source giving a similar proof [Lan11]. If you are able to find the original source of this proof, please let me know!

2.9 Solvable Groups

This section of a combination of similar sections in [Lan05] and [Jac09]. We start with a series of definitions.

Definition 2.9.1. Let G be a group, and $\{G_i\}_{i=1}^{n}$ a collection of subgroups. If $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n$, we call this sequence of groups a *tower* of subgroups. A tower of subgroups is normal if $G_{i+1} \trianglelefteq G_i$ for all $0 \leq i < n$, Abelian if it is normal and G_i/G_{i+1} is Abelian for all $0 \leq i < n$, and cyclic if it is normal and G_i/G_{i+1} is cyclic for all $0 \leq i < n$.

Note. We assume that groups are never repeated in a tower. This results in essentially no loss of information, since we can always just toss out the repeated group, and makes proofs a little easier.

At this point, we're already prepared to give the definition of a solvable group.

Definition 2.9.2. A group G is solvable if it has an Abelian tower of subgroups, which terminates with the subgroup $\{1\}$.

Note. The last requirement in this definition is primarily for convenience, as we can always just add $\{1\}$ to the end of any tower that doesn't already have it.

We would also like to be able to have some notion of how "fine" or "course" a normal tower of subgroups is, leading us to the following definition.

Definition 2.9.3. Let G be a group, and $\{G_i\}_{i=1}^n$ a normal tower of subgroups. A *refinement* of this tower is any normal tower of subgroups of G which contains $\{G_i\}_{i=1}^n$, and is called *proper* if it contains subgroups which are not in the original tower. A normal tower of subgroups is called a *composition series* if $G_n = \{1\}$ and the tower has no proper refinements.

Note. There's another way to characterize composition series. We can note that a normal H subgroup of G is maximal if and only if G/H is *simple*, that is only has normal subgroups of G/H and the trivial group. Thus, a composition series is a normal tower terminating at the trivial group such that each G_i/G_{i+1} is simple.

It turns out that we can simplify our consideration of solvable groups considerably with these refinements. To begin with, we'll need the following lemmas.

Lemma 2.9.4. *Let G, H be groups, $\varphi \in \mathrm{Hom}(G, H)$ a surjective homomorphism, and $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n \supseteq \ker(\varphi)$ a tower of subgroups. This tower is normal/Abelian/cyclic if and only if the tower $H \supseteq \varphi(G_1) \supseteq \cdots \supseteq \varphi(G_n)$ is normal/Abelian/cyclic.*

Proof. By the second fundamental theorem of homomorphisms, $G_{i+1} \trianglelefteq G_i$ if and only if $\varphi(G_{i+1}) \trianglelefteq \varphi(G_i)$, for all $0 \leq i < n$, so the normal part of this lemma is clear. Furthermore, we know from this theorem that

$$\frac{G_i}{G_{i+1}} \cong \frac{\varphi(G_i)}{\varphi(G_{i+1})}$$

which gives us the Abelian/cyclic part of the lemma. □

Lemma 2.9.5. *Let G be a solvable group, and $H \trianglelefteq G$. Then G/H is solvable. Furthermore, if G is any group and there exists $H \trianglelefteq G$ such that G/H, H are solvable, then G is solvable.*

Proof. Suppose that G is solvable, and $H \trianglelefteq G$. Since G is solvable, it has an Abelian tower $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$. Let $\pi : G \to G/H$, and set $H_i = \pi(G_i)$. We'll first show that $H_{i+1} \trianglelefteq H_i$. Pick any $g_i \in G_i, g_{i+1} \in G_{i+1}$. Then $(g_iH)^{-1}(g_{i+1}H)(g_iH) = (g_i^{-1}g_{i+1}g_i)H \in G_{i+1}H$, as required. Next, we'll show that there exists a surjective homomorphism from G_i/G_{i+1} to H_i/H_{i+1}, implying the latter is Abelian and proving that G/H is solvable. We'll define this map by $\varphi : gG_{i+1} \mapsto \pi_i(gH)$, where $\pi_i : H_i \to H_i/H_{i+1}$ is the standard projection map. We show that this is a homomorphism first.

$$\varphi((g_1G_{i+1})(g_2G_{i+1})) = \varphi((g_1g_2)G_i) = \pi_i((g_1g_2)H) = \pi_i(g_1H)\pi_i(g_2H) = \varphi(g_1G_{i+1})\varphi(g_2G_{i+1})$$

That it is surjective is immediate. Now, suppose that G is some group and there exists $H \cong G$ such that G/H is solvable. Let $G/H = H_0 \supseteq H_1 \supseteq \cdots \supseteq H_n = \{1\}$ be an Abelian tower, and $\pi : G \to G/H$ the projection map. Then by the second fundamental theorem

of homomorphisms, we can get a normal tower of subgroups $G = G_0 = \pi^{-1}(H_0) \supseteq G_1 = \pi^{-1}(H_1) \supseteq \cdots \supseteq G_n = \pi^{-1}(H_n) = H$. Furthermore,

$$\frac{G_i}{G_{i+1}} \cong \frac{H_i}{H_{i+1}}$$

is Abelian, so since H is solvable we're done. □

Theorem 2.9.6. *A finite group G is solvable if and only if it has a cyclic composition series.*

Proof. The having a cyclic composition series implies that a group is solvable is immediate. Now, suppose that G is a solvable finite group. First, we show that G has an Abelian composition series. Since G is solvable, it has an Abelian tower $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$. Any refinement of this tower remains Abelian, so since G is finite we can take a maximal refinement to get an Abelian composition series. Thus, we may assume that our Abelian tower is a composition series. Suppose G_i/G_{i+1} were not cyclic. Then since its Abelian, picking any non-trivial $x \in G_i/G_{i+1}$ we'd get a normal subgroup $G_{i+1} \subsetneq \pi^{-1}(\langle x \rangle) \subsetneq G_i$, where $\pi : G_i \to G_{i+1}$ is the standard projection map. But this contradicts our assumption that the normal tower was a composition series. □

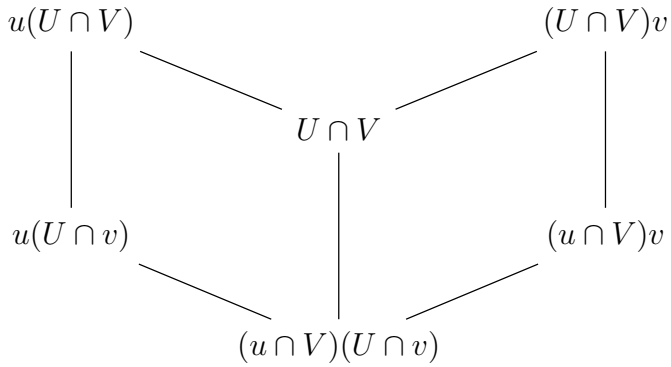
Note. Since cyclic simple groups are necessarily of prime power order, this also implies that the composition series has $G_i/G_{i+1} \cong \mathbb{Z}/p_i\mathbb{Z}$, where p_i are prime.

This can be strengthened even further, there's a notion in which all composition series are equivalent. Thus, a finite would group would be solvable if and only if all of its composition series are cyclic. Let's work towards proving this equivalence now, following [Lan05].

Lemma 2.9.7 (Buttefly Lemma)**.** *Let G be a group, $U, V \subseteq G$ subgroups, and $u \trianglelefteq U, v \trianglelefteq V$ normal subgroups. Then*

1. $u(U \cap v) \trianglelefteq u(U \cap V)$
2. $(u \cap V)v \trianglelefteq (U \cap V)v$
3. $\frac{u(U\cap V)}{u(U\cap v)} \cong \frac{(U\cap V)v}{(u\cap V)v}$

Proof. This is our first instance of proof by pretty picture.



First, we'll show that the intersection of two lines going down is the intersection of two groups, and the intersection of two lines going up is the group generated by the two groups. Clearly $U \cap V \subseteq u(U \cap V) \cap (U \cap V)v$. Suppose $g \in u(U \cap V) \cap (U \cap V)v$. Then $g = xy = wz$, where $x \in u, y, w \in U \cap V, z \in v$. Thus, $x = w(zy^{-1}) = (wzw^{-1})(wy^{-1})$. Since $z \in v$ and $w \in V$, $k = wzw^{-1} \in v$. Since $y^{-1}, w \in U \cap V$, $wy^{-1} = r \in U \cap V$, so $x \in u \cap v(U \cap V) \Rightarrow (v \cap U)(U \cap V) \Rightarrow x \in U \cap V$, and hence $g \in U \cap V$. Therefore, $U \cap V = u(U \cap V) \cap (U \cap V)v$, as claimed. We can again see that $u(U \cap v) \cap (u \cap V)v \supseteq (u \cap V)(U \cap v)$. Suppose $g \in u(U \cap v) \cap (u \cap V)v$. Then $g = xy = wz$, where $x \in u, y \in U \cap v, w \in u \cap V, z \in v$. Thus, $x = wzy^{-1} \in V$, so $x \in u \cap V$ and hence $u(U \cap v) \cap (u \cap V)v = (u \cap V)(U \cap v)$ as claimed. Again, we can see that $(u \cap V)(U \cap v) \subseteq (U \cap V) \cap u(U \cap v)$. Suppose $g \in (U \cap V) \cap u(U \cap v)$. Then $g = xy$, where $x \in u$ and $y \in U \cap v$. Since $g \in U \cap V$, it follows that $x = gy^{-1} \in V$, so $x \in u \cap V$ and hence $(u \cap V)(U \cap v) = (U \cap V) \cap u(U \cap v)$ as claimed. The remaining case proceeds by an identical argument. The claim about lines going up being generating groups is clear. Second, we can see from the diagram that (1) and (2) are equivalent, so we prove only (1). Suppose $g \in u(U \cap V), h \in u(U \cap v)$. Then $g = xy, h = wz$, where $x, w \in u, y \in U \cap V, z \in U \cap v$. Thus,

$$ghg^{-1} = (xy)(wz)(y^{-1}x^{-1}) = x(ywy^{-1})(yzy^{-1})x^{-1}$$

$r = ywy^{-1} \in u, k = yzy^{-1} \in U \cap v$, so we get

$$ghg^{-1} = xrkx^{-1} = (xr)(kx^{-1}k^{-1})k$$

$xr \in u, kx^{-1}k^{-1} \in u$, so $ghg^{-1} \in u(U \cap v)$ and hence $u(U \cap v) \trianglelefteq u(U \cap V)$, as claimed. Finally, we prove (3). By the third fundamental theorem of homomorphisms

$$\begin{aligned}\frac{u(U \cap V)}{u(U \cap v)} &= \frac{(U \cap V)u(U \cap V)}{u(U \cap v)} \cong \frac{U \cap V}{u(U \cap v) \cap (U \cap V)} = \frac{U \cap V}{(u \cap V)(U \cap v)} \\ &= \frac{U \cap V}{(u \cap V)v \cap (U \cap V)} \cong \frac{(U \cap V)(u \cap V)v}{(u \cap V)v} = \frac{(U \cap V)v}{(u \cap V)v}\end{aligned}$$

as claimed. □

Now, we need to define what we mean when we say two towers are equivalent.

Definition 2.9.8. Let G be a group, and $G = G_1 \supseteq G_2 \supseteq \cdots \supseteq G_n = \{1\}, G = H_1 \supseteq H_2 \supseteq \cdots \supseteq H_m = \{1\}$ be two normal towers. We call these towers *equivalent* if $r = s$ and there exists $\sigma \in S_{n-1}$ such that $\frac{G_i}{G_{i+1}} \cong \frac{H_{\sigma(i)}}{H_{\sigma(i)+1}}$ for all $1 \leq i < n$.

Which will allow us to, finally, develop results on these equivalences.

Theorem 2.9.9 (Schreier's Theorem)**.** *Suppose $G = G_1 \supseteq G_2 \supseteq \cdots \supseteq G_n = \{1\}, G = H_1 \supseteq H_2 \supseteq \cdots \supseteq H_m = \{1\}$ are two normal towers. Then they have equivalent refinements.*

Proof. For each $1 \leq i < n, 1 \leq j \leq m$, define $G_{ij} = G_{i+1}(G_i \cap H_j)$. We can note that $G_{i1} = G_i, G_{im} = G_{i+1}$, so by the butterfly lemma

$$G \supseteq G_{12} \supseteq \cdots \supseteq G_{1m} \supseteq G_{21} \supseteq \cdots \supseteq G_{(n-1)1} \supseteq \cdots \supseteq \{1\}$$

is a refinement of the first normal tower. Similarly, setting $H_{ji} = (H_j \cap G_i)H_{j+1}$ for all $1 \leq i \leq n, 1 \leq j < m$, we get

$$G \supseteq H_{12} \supseteq \cdots \supseteq H_{1n} \supseteq H_{21} \supseteq \cdots \supseteq H_{(m-1)1} \supseteq \cdots \supseteq \{1\}$$

is a refinement of the second normal tower. By the butterfly lemma

$$\frac{G_{ij}}{G_{i(j+1)}} \cong \frac{H_{ji}}{H_{j(i+1)}}$$

for all $1 \leq i < n, 1 \leq j < m$ completing the proof. □

Theorem 2.9.10 (Jordan-Hölder)**.** *If G is solvable, then any two composition series of G are equivalent*

Proof. By Schreier's theorem, the two composition series must have equivalent refinements. But composition series have no proper refinements, so it follows that the composition series are equivalent. □

At this point, we've proven everything I find particularly enlightening (at least at this point in my life) about solvable groups, without of course getting into their major role in Galois theory. We'll end this section by taking a look at the connection between solvability and commutators, following [Jac09].

Definition 2.9.11. Let G be a group and $g, h \in G$. We denote the commutator by

$$[g, h] = g^{-1}h^{-1}gh$$

The *derived* group $G' \subseteq G$ is the subgroup generated by all the commutators of two elements in G. We define the n-th derived group iteratively by $G^{(n)} = (G^{(n-1)})'$.

Lemma 2.9.12. *For any $k \in \mathbb{N}$, $G^{(k)} \trianglelefteq G$.*

Proof. We can note that since $[g, h]^{-1} = h^{-1}g^{-1}hg = [h, g]$, G' consists of elements of the form $[g_1, h_1] \cdots [g_n, h_n]$, where $g_i, h_i \in G$. Let H be another group, and $\varphi \in \mathrm{Hom}(G, H)$. Then $\varphi([g, h]) = [\varphi(g), \varphi(h)] \in H'$, so $\varphi(G') \subseteq H'$. Now, pick any $K \trianglelefteq G$ and $a \in G$. The map $\varphi : g \mapsto aga^{-1}$ induces an automorphism of K, so $\varphi(K') \subseteq K'$. Since a was arbitrary, this implies that $K' \trianglelefteq G$. In particular, $G \trianglelefteq G \Rightarrow G' \trianglelefteq G \Rightarrow \cdots \Rightarrow G^{(k)} \trianglelefteq G$. □

Lemma 2.9.13. *G/G' is Abelian and if $K \trianglelefteq G$ is such that G/K is Abelian then $G' \subseteq K$.*

Proof. Let $\pi \in \mathrm{Hom}(G, G/G')$ be the projection map, and pick any $g, h \in G$. Then $\pi(g)\pi(h) \equiv g(g^{-1}hgh^{-1})h \equiv hg \equiv \pi(h)\pi(g)$, so G/G' is Abelian. Now, suppose that $K \trianglelefteq G$ is such that G/K is Abelian, and pick any $g, h \in G$. Then

$$(g^{-1}h^{-1}gh)K = (gK)^{-1}(hK)^{-1}(gK)(hK) = (gK)^{-1}(gK)(hK)^{-1}(hK) = K$$

so $g^{-1}h^{-1}gh \in K \Rightarrow G' \subseteq K$, as was to be shown. □

Theorem 2.9.14. *A group G is solvable if and only if there exists $n \in \mathbb{N}$ such that $G^{(n)} = \{1\}$.*

Proof. One direction is simple, as $G \supseteq G' \supseteq \cdots \supseteq G^{(n)}$ is an Abelian tower by lemmas 2.9.12 and 2.9.13. Now, suppose that G is solvable, and that $G = H_0 \supseteq H_1 \supseteq \cdots \supseteq H_k = \{1\}$ is an Abelian tower. By lemma 2.9.13, $G' \subseteq H_1$, so $G'' \subseteq H_1' \subseteq H_2$ and so on, forcing $G^{(k)} = \{1\}$ as desired. □

Corollary 2.9.14.1. *Any subgroup H of a solvable group G is solvable.*

Proof. Since G is solvable, there exists some $n \in \mathbb{N}$ such that $G^{(n)} = \{1\}$. For any $k \in \mathbb{N}$, $H^{(k)} \subset G^{(k)}$. Thus, $H^{(n)} = 1$, making H solvable. □

2.10 Group Representations*

As you may have surmised from the past chapter, understanding the structure of groups can be quiet difficult. One of the ways of getting around this is group representations, which we'll introduce (but not develop all that much) here. The study of representations is a field unto itself; those interested should look at [Lan05] or one of the numerous textbooks dedicated to the subject.

Given a vector space V, one can see that the set of all invertible linear maps from V to V, with composition as multiplication, forms a group. Since we understand linear transformations far more than we do groups, our aim is to shift the study of groups to the study of linear algebra.

Definition 2.10.1. Let G be a group and V a vector space. A representation of G in V is the image of a homomorphism from G to the invertible linear transformations of V to itself.

Example 2.10.1. Given any group G and vector space V, we have the trivial representation given by $\varphi : g \mapsto \mathrm{Id}_V$. Needless to say this one is not particularly useful.

Example 2.10.2. Consider the group $\Gamma = \{e^{i\frac{2\pi j}{n}}\}_{1 \leq j \leq n}$, where the operation is multiplication. We can represent this in $\mathrm{GL}_2(\mathbb{C})$ using the map

$$\varphi : e^{i\frac{2\pi j}{n}} \mapsto \begin{pmatrix} 1 & 0 \\ 0 & e^{i\frac{2\pi j}{n}} \end{pmatrix}$$

There are also many ways of classifying representations, we introduce some of them below.

Definition 2.10.2. Let G be a group represented in a vector space V, with the associated homomorphism being φ. A representation is said to *irreducible* if, given any non-trivial subspace $U \subsetneq V$, there exists some $\underline{v} \in U$ and $g \in G$ such that $\varphi(g)(\underline{v}) \notin U$. Otherwise, it is called *reducible*. A representation is called *faithful* if φ is injective.

Chapter 3

Rings

3.1 Basic Definitions

When studying group theory, two of our most common examples were the monoids $(\mathbb{Z}, +)$ and $(\mathbb{Z}^*, \cdot)$ (non-zero integers under multiplication). Of course, there is something rather unnatural about treating these two monoids as separate objects; we know intuitively that they are parts of a single object, the integers. Our resolution to this is the ring.

Definition 3.1.1. A ring R is a set with elements $0, 1 \in R$ (which we call the zero and identity) and binary operations $+, \cdot$ such that

1. $(R, +, 0)$ is an Abelian group.
2. $(R, \cdot, 1)$ is a monoid.
3. For any $x, y, z \in R$, distributivity is respected. That is,

$$(x + y) \cdot z = x \cdot z + y \cdot z$$
$$z \cdot (x + y) = z \cdot x + z \cdot y$$

Note. Like with groups, we often drop the $\cdot$ when writing products. We will often denote $R\backslash\{0\}$ by R^*. The definition of a ring will vary from text to text, some older sources do not assume that all rings have an identity. Others call rings without an identity rngs (this is a ploy used by [Jac09]). Either way, it will not be of much interest to us here. All of our rings, by assumption, will have an identity.

Example 3.1.1. Perhaps the best example of a ring is $M_n(\mathbb{R})$, the set of $n \times n$ real matrices with matrix addition and multiplication. Our identity here is the identity matrix, and our zero the zero matrix. Note that multiplication here is **not** commutative: assuming it to be so is a common mistake when working with rings.

There are many types of properties a ring can have, we list them here.

Definition 3.1.2. A ring R such that $0 \neq 1$ is

1. Commutative if $(R, \cdot, 1)$ is Abelian.

2. A (integral) domain if $(R^*, \cdot, 1)$ is a sub-monoid.

3. A division ring if $(R^*, \cdot, 1)$ is a sub-group.

4. A field if it is a commutative division ring.

Note. The assumption that $0 \neq 1$ here is very common, so much so that it will often be assumed without being stated.

We'll come back to all of these in a moment to explore their connections more deeply. For now, we note some basic results on arithmetic in rings.

Proposition 3.1.3. *Let R be a ring, and $x, y \in R$ ring elements. Then*

1. $1x = x1 = x$

2. $0x = x0 = 0$

3. $(-1)x = -x$

4. *If $xy = yx$, then for any $n, m \in \mathbb{N}$ we get $x^n y^m = y^m x^n$, and*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

The proof of these basic identities is left to the reader. Like with groups, we of course have subrings and generated subrings.

Definition 3.1.4. A subring $S \subseteq R$ of a ring R is a subset which is also a ring. For an arbitrary subset $A \subset R$, then subring generated by A, $\langle A \rangle$, is the smallest subring of R containing A.

Like with group and monoids, the intersection of subrings is a subring, and hence $\langle A \rangle$ is the intersection of all subrings of R containing A. The elements of $\langle A \rangle$ are $0, 1$, and all finite sums of finite products of elements of A. With these basic concepts out of the way, we return to the problem of characterizing the properties of rings.

Definition 3.1.5. If $a \in R$ is such that there exists some non-zero $b \in R$ for which $ab = 0$ ($ba = 0$), then a is called a left (right) zero-divisor of b.

Theorem 3.1.6. *The following are equivalent (where R is a ring and $R \neq 0$).*

1. *R is a domain.*

2. *R has no non-zero zero-divisors.*

3. *For any $x, y, z \in R$, $xy = xz \Rightarrow y = z$ or $x = 0$ and $xy = zy \Rightarrow x = z$ or $y = 0$. This condition is called the cancellation law.*

Proof. First, suppose that R is a domain. Then since R^* is a sub-monoid, the product of any pair of non-zero elements is non-zero, and hence R has only 0 as a zero divisor. Now, suppose that R has no non-zero zero-divisors. If $xy = xz$, then $x(y - z) = 0$, so either $x = 0$ or $y - z = 0 \Rightarrow y = z$. If $xy = zy$, then $(x - z)y = 0$, so either $y = 0$ or $x - z = 0 \Rightarrow x = z$. Finally, suppose that the cancellation law holds, and pick any pair of non-zero elements $x, y \in R$. If $xy = 0$, then $xy = x0 \Rightarrow x = 0$ or $y = 0$, a contradiction. Thus, $xy \in R^*$, making R^* a sub-monoid. □

There are two final objects to define before we move on.

Definition 3.1.7. The subgroup of $(R^*, \cdot, 1)$ consisting of elements with a multiplicative inverse is called the units of R, and denoted

$$R^\times = \{x \in R \mid \exists y \in R, xy = yx = 1\}$$

Note. We again denote the multiplicative inverse of $x \in R$ by x^{-1}.

Definition 3.1.8. Let R, R' be rings. A ring homomorphism is a map $\varphi : R \to R'$ which is a group homomorphism $(R, 0, +) \to (R', 0', +')$ and a monoid homomorphism $(R, 1, \cdot) \to (R', 1', \cdot')$. The set of all ring homomorphisms between two rings is denoted $\mathrm{Hom}(R, R')$

Note. By definition, we must get $\varphi(0) = 0'$ and $\varphi(1) = 1'$.

3.2 Matrix Rings

We start with one of the simplest yet most important types of rings, the matrix ring. This section is based on a similar one in [Jac09].

Definition 3.2.1. Let R be an arbitrary ring, and $n \in \mathbb{N}$. The matrix ring of R, denoted $M_n(R)$, is the set of all $n \times n$ matrices with entries in R. Endowed with standard matrix addition and multiplication, $M_n(R)$ is a ring.

Note. We can embed R in $M_n(R)$ via the monomorphism $x \mapsto \mathrm{diag}(x, x \ldots, x)$. Thus, if $A \in M_n(R)$, by xA we mean $\mathrm{diag}(x, x, \ldots, x)A$. We can pull a similar trick and map $\mathbb{Z}$ into any ring R, with the map $f \in \mathrm{Hom}(\mathbb{Z}, R)$ being defined by $f(1) = 1$. In this notation, for any given $n \in \mathbb{Z}$ we write nx for $f(n)x$.

Note. Even if R is commutative, $M_n(R)$ will be non-commutative if $n \geq 2$.

We denote by e_{ij} the matrix with entries all zero except in the i, jth position. Note then that

$$(a_{ij}) = \sum_{i,j} a_{ij} e_{ij}$$

I'll cut right to the chase, let's figure out determinants shall we?

Theorem 3.2.2. *For any commutative ring R, there exists a unique function $f : M_n(R) \to R$ such that*

1. $f(\mathrm{Id}_n) = 1$.
2. *If A' is the matrix A with its rows permuted by some $\sigma \in S_n$, then $f(A') = \mathrm{sgn}(\sigma) f(A)$.*
3. *f is linear in each row of $M_n(R)$, keeping all other rows fixed.*

Proof. Suppose such a function f existed. Pick an arbitrary matrix $A = (a_{ij})$. Then

$$\begin{aligned} f(A) = f\Big(\sum_{i,j} a_{ij} e_{ij}\Big) &= \sum_{k=1}^{n} a_{1k} f\Big(e_{1k} + \sum_{i=2}^{n}\sum_{j=1}^{n} a_{ij} e_{ij}\Big) \\ &= \sum_{k_1=1}^{n} \cdots \sum_{k_n=1}^{n} a_{1k_1} \cdots a_{nk_n} f(e_{1k_1} + \cdots + e_{nk_n}) \end{aligned}$$

Note that if some $k_i = k_j$ in any given sum, then there's a permutation of the rows of $e_{1k_1} + \cdots + e_{nk_n}$ of sign one which doesn't change the matrix. That is, we get

$$f(e_{1k_1} + \cdots + e_{nk_n}) = -f(e_{1k_1} + \cdots + e_{nk_n})$$

Thus, all the terms with $k_i = k_j$ for some $1 \leq i < j \leq n$ cancel out to zero, and we're left with

$$f(A) = \sum_{\sigma \in S_n} a_{1\sigma(1)} \cdots a_{n\sigma(n)} f(e_{1\sigma(1)} + \cdots + e_{n\sigma(n)})$$

But of course each $e_{1\sigma(1)} + \cdots + e_{n\sigma(n)}$ is just the identity with its rows permuted by σ, so

$$f(A) = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)} f(\mathrm{Id}_n) = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}$$

Therefore, f is unique if it is well-defined. We just need to check then that the function

$$f(A) = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}$$

satisfies all three properties. This verification is not too difficult, and left to the reader. □

We call this unique function the determinant, and denote it det. This is of course the same determinant as you would have encountered in linear algebra, although perhaps it doesn't seem that way for now. Let's look at different ways of formulating this same function.

Definition 3.2.3. Let $A \in M_n(R)$. The i,jth minor of A, denoted $M_{A,i,j}$, is the determinant of the matrix obtained be removing the ith row and jth column of A, multiplied by $(-1)^{i+j}$.

This definition leads to the determinant which you may be more familiar with. Before that though, we need a quick lemma.

Proposition 3.2.4. *If R is a commutative ring, then properties (2) and (3) of* det *also hold for columns.*

Note. This hints at the fact that our choice to define the determinant properties in terms of columns in Theorem 3.2.2 was arbitrary. In fact, it would be entirely equivalent to phrase Theorem 3.2.2 in terms of columns, and then re-phrase and prove the lemma 3.2.4 in terms of rows.

Corollary 3.2.4.1. *If R is commutative, then for any $A \in M_n(R)$ and $1 \leq i \leq n$*

$$\det(A) = \sum_{j=1}^{n} a_{ij} M_{A,i,j}$$

and

$$\det(A) = \sum_{j=1}^{n} a_{ji} M_{A,j,i}$$

Proof. We note that

$$\det(A) = \sum_{j=1}^{n} a_{ij} \det\left(e_{ij} + A - \sum_{k \neq j} e_{ik}\right)$$

Let's examine $e_{ij} + A - \sum_{k \neq j} e_{ik}$ more closely. Each of these is a matrix of the form

$$\begin{pmatrix} a_{11} & \cdots & a_{1(j-1)} & a_{1j} & a_{1(j+1)} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{(i-1)1} & \cdots & a_{(i-1)(j-1)} & a_{(i-1)j} & a_{(i-1)(j+1)} & \cdots & a_{(i-1)n} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ a_{(i+1)1} & \cdots & a_{(i+1)(j-1)} & a_{(i+1)j} & a_{(i+1)(j+1)} & \cdots & a_{(i+1)n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{n(j-1)} & a_{nj} & a_{n(j+1)} & \cdots & a_{nn} \end{pmatrix}$$

We can apply a row and column swap to this to end up with the matrix

$$\begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{(i-1)j} & \cdots & a_{(i-1)(j-1)} & a_{(i-1)1} & a_{(i-1)(j+1)} & \cdots & a_{(i-1)n} \\ a_{1j} & \cdots & a_{1(j-1)} & a_{11} & a_{1(j+1)} & \cdots & a_{1n} \\ a_{(i+1)j} & \cdots & a_{(i+1)(j-1)} & a_{(i+1)1} & a_{(i+1)(j+1)} & \cdots & a_{(i+1)n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{nj} & \cdots & a_{n(j-1)} & a_{n1} & a_{n(j+1)} & \cdots & a_{nn} \end{pmatrix}$$

It's clear from the permutation definition of the determinant (and the row/column properties it satisfies) then that $\det\left(e_{ij} + A - \sum_{k \neq j} e_{ik}\right)$ is $(-1)^{i+j}$ multiplied by the determinant of the (1,1) minor of the above matrix. But of course that's just $M_{A,i,j}$, giving us the desired result. The result for column expansion is proved in an essentially identical manner. □

The determinant has one more relevant property, the proof of which we omit[1].

Proposition 3.2.5. $\det : M_n(R) \to R$ *is a monoid homomorphism with respect to matrix multiplication.*

Note. This is just a very fancy way of saying $\det(AB) = \det(A)\det(B)$. The point of viewing it in the above manner is that we can then embed R into $M_n(R)$, giving us a natural map $\det : M_n(R) \to M_n(R)$ defined by (abusing notation a bit) $\det(A) = \text{diag}(\det(A), \ldots, \det(A))$.

We need one more result before we can cover the main application of the determinant.

Lemma 3.2.6. *Suppose R is commutative, $A \in M_n(R)$, and $1 \leq i, j \leq n$ are such that $i \neq j$. Then*

$$\sum_{k=1}^{n} a_{ik} M_{A,j,k} = 0$$

$$\sum_{k=1}^{n} a_{ki} M_{A,k,j} = 0$$

Proof. We prove only the first identity, the proof for the second is essentially the same. Note that by corollary 3.2.4.1, the first expression is just the determinant of the matrix A with its jth row replaced by its ith row. Since $i \neq j$, such a matrix has a repeated row, which by property (2) of determinants implies it has determinant zero. □

Let us now cover the aforementioned application of determinants, invertibility and adjugate matrices.

Definition 3.2.7. Let $A \in M_n(R)$, where R is commutative. The cofactor matrix of A, denoted $\text{adj}(A)$, is the $n \times n$ matrix who's (i, j)th entry is $M_{A,j,i}$.

Note. [Jac09] calls this the adjoint matrix, which is terrible form and should not be done.

Theorem 3.2.8. *Suppose R is commutative and $A \in M_n(R)$. Then A is invertible if and only if* $\det(A)$ *is a unit, and the inverse of A (if it exists) is* $\det(A)^{-1} adj(A)$.

Note. In this statement, and its proof, we use $\det(A)$ to refer both to the determinant of a matrix and the embedding of that determinant back into $M_n(R)$.

Proof. First, suppose that A is invertible. Then by proposition 3.2.5 we get $1 = \det(AA^{-1}) = \det(A)\det(A^{-1})$, so $\det(A)$ must be a unit. Now, suppose that $\det(A)$ is a unit. For any $1 \leq i, j \leq n$, we get

$$(A\text{adj}(A))_{ij} = \sum_{k=1}^{n} a_{ik}(\text{adj}(A))_{kj} = \sum_{k=1}^{n} a_{ik} M_{A,j,k}$$

Thus, by lemma 3.2.6 and corollary 3.2.4.1 we conclude that $A\text{adj}(A) = \det(A)$. A similar calculation implies that $\text{adj}(A)A = \det(A)$, so since $\det(A)$ is a unit we get $A^{-1} = \det(A)^{-1}\text{adj}(A)$, as required. □

In particular, this is just the familiar result from an introductory linear algebra course.

Corollary 3.2.8.1. *If F is a field, then $A \in M_n(F)$ is invertible if and only if* $\det(A) \neq 0$.

[1] The proof is basically symbol pushing

3.3 Ideals and Quotient Rings

As will become a common theme in this text, we wish now to replicate the homomorphism theorems for groups in the setting of rings. In order to do this, we need to figure out how to construct quotient rings. To that end, let us consider a ring R with subring A. Ideally, we would like for the following equations to hold in R/A, for any $x, y \in R$

$$(x + A) + (y + A) = (x + y) + A, (x + A)(y + A) = xy + A$$

The first of these we get for free, in particular it's just a manifestation of the additive group in the ring being Abelian, and all subgroups of Abelian groups being normal. The second is not at all guaranteed, leading us to the following definition.

Definition 3.3.1. A left (right) ideal $I \subseteq R$ is a sub-ring of R such that $RI \subset I$ ($IR \subset I$). A subset which is both a left and right ideal is simply called an ideal.

Note. R is necessarily an ideal of R.

Ideals are in fact the structure we need to generate quotient rings. Indeed, one can expand[2] to get

$$(x + A)(y + A) = xy + xA + Ay + A^2$$

so we need a guarantee that $xA, Ay \subset A$, which is exactly to say that A is an ideal.

Definition 3.3.2. Let R be a ring and $I \subset R$ an ideal. The quotient ring R/I is the quotient group with multiplication defined by, for any $x, y \in R$

$$(x + I)(y + I) = xy + I$$

Again, that this is a well-defined ring follows from the properties of an ideal. That being said, let us explore more properties of ideals.

Proposition 3.3.3. *Let R be a ring, and $\{I_j\}_{j\in J}$ a collection of ideals in R. Then $\bigcap_{j\in J} I_j \subset R$ is an ideal.*

Proof. That it's an additive subgroup follows from Theorem 2.1.5. We just need to check then closure under multiplication. Pick any $a \in \bigcap_{j\in J} I_j$ and $x \in R$. Then for each I_j, $a \in I_j$, and hence $xa, ax \in I_j$. It follows that $xa, ax \in \bigcap_{j\in J} I_j$, as required. □

This, along with the above note about R being an ideal of itself, allows us to define the sub-ring generated by a set.

Definition 3.3.4. Let R be a ring, and $S \subset R$. The ideal generated by S, denoted (S), is the smallest ideal in R containing S (i.e. the intersection of all ideals in R containing S).

Like with rings and monoids, we can explicitly write out the elements of this ideal.

[2]Strictly speaking this is not a "proper" expansion, but you get the idea.

Proposition 3.3.5. *Let R be a ring, and $S \subset R$. Then*

$$(S) = \Big\{ \sum_{a \in S} x_a a y_a \mid x_a, y_a \in R \Big\}$$

where all the above sums have finitely many non-zero terms.

Note. The above proposition and definition have fairly immediate equivalent formulations for right and left ideals.

The proof of this is left to the reader.

Note. I'm going to be doing a lot more "left to the reader" or "it is clear" explanations from now on. The hope is that the rigour of the previous section has given you the intuition to follow and be comfortable with such explanations.

3.4 Homomorphism Theorems

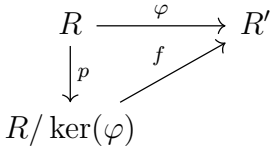
The theorems so nice we cover them twice. These are the theorems presented in [Jac09], although the order and proofs have been changed. We start with a quick lemma.

Lemma 3.4.1. *Let $\varphi \in \mathrm{Hom}(R, R')$, where R, R' are rings. Then $\ker(\varphi) \subset R$ is an ideal.*

Proof. That it's an additive subgroup is immediate by a similar result on group homomorphisms, so it suffices to show that $R\ker(\varphi), \ker(\varphi)R \subset \ker(\varphi)$. To that end, pick any $r \in R, x \in \ker(\varphi)$. Then $\varphi(rx) = \varphi(r)\varphi(x) = 0\varphi(x) = 0$, and $\varphi(xr) = \varphi(x)\varphi(r) = \varphi(x)0 = 0$, so $rx, xr \in \ker(\varphi)$, as required. □

No point in wasting time, let's jump right into these.

Theorem 3.4.2 (First Fundamental Theorem of Homomorphisms)**.** *Let $\varphi : R \to R'$ be a ring homomorphism. Then the natural projection map $p : R \mapsto R/\ker(\varphi)$ is a ring homomorphism, and the map $f : R/\ker(\varphi) \to \mathrm{Im}(\varphi)$ given by $f : x + \ker(\varphi) \mapsto \varphi(x)$ is a well-defined ring isomorphism. Finally, the following diagram commutes.*



Proof. First, pick any $x, y \in R$. Then $p(xy) = xy + I = (x + I)(y + I)$ (here we're doing arithmetic with cosets), and $p(x + y) = (x + I) + (y + I) = (x + y) + I$, making p a ring homomorphism. f is certainly a homomorphism if well-defined (by an essentially identical check), so we check instead that it is indeed well-defined. Pick any $x, y \in R$ such that $x + \ker(\varphi) = y + \ker(\varphi)$. Then $\exists z \in \ker(\varphi)$ such that $x = y + z$. Thus, $\varphi(x) = \varphi(y) + \varphi(z) = \varphi(y)$, making f well-defined. Finally, we show that f is an isomorphism onto $\mathrm{Im}(\varphi)$ (the above diagram commuting is immediate from this). But this is immediate from our verification of f being well-defined. □

Theorem 3.4.3 (Second Fundamental Theorem of Homomorphisms)**.** *Suppose $\varphi : R \to R'$ is a surjective ring homomorphism. Then*

1. *An additive subgroup $S \subset R$ containing $\ker(\varphi)$ is a subring (ideal) of R if and only if $\varphi(S)$ is a subring (ideal) of R'.*
2. *The map $S \mapsto \varphi(S)$ on subrings (ideals) of R containing $\ker(\varphi)$ is a bijection onto subrings (ideals) of R'.*
3. *If $I \subset R$ is an ideal containing $\ker(\varphi)$, then $R/I \cong R'/\varphi(I)$.*

Proof. We start with the first statement. If S is a subring (ideal), then since φ is surjective it is immediate that $\varphi(S)$ is a subring (ideal). Now, suppose that $\varphi(S)$ is a subring. Pick any $x, y \in S$. Then $\varphi(xy) = \varphi(x)\varphi(y) \in \varphi(S)$, so it follows that there exists some $s \in S, z \in \ker(\varphi)$ such that $xy = s + z$. But of course $\ker(\varphi) \subset S$, so this implies that $xy \in S$, and hence S is a subring as it is an additive subgroup. Suppose further that $\varphi(S)$ is an ideal. Pick any $x \in S, r \in R$. Then $\varphi(xr) = \varphi(x)\varphi(r) \in \varphi(S)$, and similar with $\varphi(rx)$. Thus, both differ from an element of S only by some element of $\ker(\varphi)$, which again means that $xr, rx \in S$ and hence S is an ideal.

Now for the second statement. Suppose that S, S' are two subrings (ideals) of R containing $\ker(\varphi)$. Then $\varphi(S) = \varphi(S')$ implies that any element of S not in S' differs only by addition of an element in $\ker(\varphi)$, and vice-versa. But of course both subrings (ideals) contain $\ker(\varphi)$, so this implies that $S = S'$, and hence $S \mapsto \varphi(S)$ is injective. Now, suppose that $S' \subset R'$ is a subring (ideal). It suffices to show that $\varphi^{-1}(S')$ is a subring (ideal). But of course φ is an additive group homomorphism and S' an additive subgroup, so $\varphi^{-1}(S')$ is an additive subgroup of R and our result follows from part (1) of this theorem.

Finally, we prove the third statement. Suppose that $I \subset R$ is an ideal containing $\ker(R)$. By part (1) of this theorem, $\varphi(I)$ is an ideal in R'. We define a map $f : R/I \to R'/\varphi(I)$ by, for any $x \in R$, $f : x + I \mapsto \varphi(x) + \varphi(I)$. We first check that this is well-defined. Suppose $x, y \in R$ are such that $x + I = y + I$. Then $\exists z \in I$ such that $x = y + z$. Thus, $\varphi(x) = \varphi(y) + \varphi(z)$, so since $\varphi(z) \in \varphi(I)$ we get $f(x + I) = f(y + I)$, as required. Next, we check that this is a homomorphism. Suppose that $x, y \in R$. Then $f((x + y) + I) = \varphi(x + y) + \varphi(I) = (\varphi(x) + \varphi(y)) + \varphi(I) = (\varphi(x) + \varphi(I)) + (\varphi(y) + \varphi(I)) = f(x) + f(y)$, and $f(xy + I) = \varphi(xy) + \varphi(I) = \varphi(x)\varphi(y) + \varphi(I) = (\varphi(x) + \varphi(I))(\varphi(y) + \varphi(I)) = f(x)f(y)$, as required. Finally, we show that this is an isomorphism. Pick any $x, y \in R$ and suppose that $f(x + I) = f(y + I)$. Then $\varphi(x), \varphi(y)$ differ only by an element of $\varphi(I)$, and hence $\varphi(x - y) \in \varphi(I)$. Thus, by part (2) of this theorem, $x - y \in I \Rightarrow x + I = y + I$, making f injective. Surjectivity follows from the surjectivity of φ. □

Corollary 3.4.3.1. *Suppose that $I \subset J$ are both ideals in a ring R. Then*

$$R/J \cong \frac{R/I}{J/I}$$

Proof. This is just the third part of the proceeding theorem applied to the surjective homomorphism $p : R \to R/I$ (the projection map). □

Theorem 3.4.4 (Third Fundamental Theorem of Homomorphisms)**.** *Suppose S is a subring and I an ideal in a ring R. Then $S + I = \{x + y \mid x \in S, y \in I\}$ is a subring of R containing I, $S \cap I$ is an ideal of S, and*

$$\frac{S}{S \cap I} \cong \frac{S + I}{I}$$

Proof. The proofs of $S + I$ being a subring and $S \cap I$ an ideal of S are left to the reader (the proof is a direct verification). For the last part, we define our map by $f : s + (S \cap I) \mapsto s + I$. That this is a homomorphism if it is well-defined is immediate, so we just check that it's well-defined and bijective. For well-defined, suppose that $s + (S \cap I) = s' + (S \cap I)$. Then $\exists z \in S \cap I$, and in particular $z \in I$, such that $s = s' + z \Rightarrow s + I = s' + I$, as required. For injectivity, suppose that $s + I = s' + I$. Then $\exists z \in I$ such that $s = s' + z$. Since $s - s' = z$, it follows that $z \in S$, so $z \in S \cap I$ and hence $s + (S \cap I) = s' + (S \cap I)$, as required. Finally, we do surjectivity. Pick any $x + y \in S + I$. Then $(x + y) + I = x + I = f(x + (S \cap I))$, making f surjective. □

It's worth at the end here taking a moment to compare these theorems to those in section 2.4, and noting any similarities or differences between them. In fact, there's a sense in which the two sets of theorems are in fact identical, which will be explored further in chapter 6.

3.5 Field of Fractions

Note. **For the rest of the chapter, all rings are assumed to be commutative unless otherwise stated.**

This is following a similar section in [Jac09], although it has been re-written significantly. The question we explore here is quite simple. Given an arbitrary domain, can we embed it into a field? The answer turns out to be no in general, but it turns out that for commutative rings we can always do this. The natural construction to prove this is actually much more intuitive than one may think. Indeed, at this point you've probably seen a construction of the rational numbers from the integers. This will, in fact, work for any ring.

Definition 3.5.1. The field of fractions of a ring R, denoted $FF(R)$, is the set

$$\{(a, b) \in R \times R^* \mid b \neq 0\}/\sim$$

where $\sim$ is the equivalence on $R \times R^*$ given by $(a, b) \sim (c, d) \iff ad = bc$, equipped with binary operations

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

Proposition 3.5.2. *$FF(R)$ is a well-defined field, with zero element $[(0, 1)]$ and identity $[(1, 1)]$ which R embeds into via the map $x \mapsto (x, 1)$.*

Note. We often refer to the map in the above proposition as the natural embedding of a domain into its field of fractions (although this is technically bad form and shouldn't be

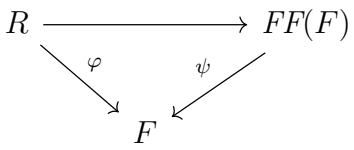
done). Since $\mathrm{FF}(R)$ is a field, we will also (suggestively) denote $[(a,b)]$ by a/b or $\frac{a}{b}$. In this notation, the above operations become

$$a/b + c/d = \frac{ad+bc}{bd} \qquad \frac{a}{b}\cdot\frac{c}{d} = \frac{ac}{bd}$$

Proof. Left as an exercise to the reader (it's good practice for working with fields of fractions in calculations, but not particularly enlightening from a conceptual standpoint). The important thing to note is that $(a/b)^{-1} = b/a$. □

Of course, domains can be embedded into multiple fields. What makes the field of fractions special is that it is, in a sense, the smallest such field. This is characterized by the following *universal property*[3].

Theorem 3.5.3. *Suppose R is a domain embedded in some field F via $\varphi : R \hookrightarrow F$. Then there exists a unique $\psi \in Hom(FF(R), F)$ such that the following diagram commutes*



where the unlabelled arrow is he the natural embedding.

Proof. To start, suppose such a $\psi \in \mathrm{Hom}(\mathrm{FF}(R), F)$ existed. Then we'd require that $\varphi(a) = \psi(a/1)$. Furthermore, it would follow that

$$\psi(a/b) = \psi\Big(\frac{a}{1}\frac{1}{b}\Big) = \psi(a/1)\psi(b/1)^{-1} = \varphi(a)\varphi(b)^{-1}$$

But this completely characterizes ψ, making it the desired unique homomorphism if it is well-defined. So, we just need to check that it is in fact well-defined. To that end, we first note that for $b \neq 0$, $\varphi(b) \neq 0$ (as φ is injective), and hence $\varphi(b)^{-1}$ is well-defined. Thus, $\varphi(a)\varphi(b)^{-1}$ is well-defined. To end off then, we just need to check that $a/b = c/d \Rightarrow \psi(a/b) = \psi(c/d)$. But since $ad = bc$, we get

$$\varphi(ad) = \varphi(bc) \Rightarrow \varphi(a)\varphi(b)^{-1} = \varphi(c)\varphi(d)^{-1}$$

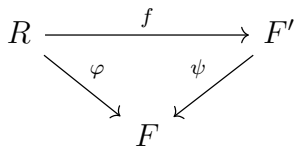
as required. □

In fact, Theorem 3.5.3 implies something a little stronger, that $\mathrm{FF}(R)$ is characterized up to *unique isomorphism*[4]. By this, we mean the following.

Corollary 3.5.3.1. *Suppose F' is a field, and $f : R \hookrightarrow F'$ an embedding of a domain. If, for every field F and homomorphism $\varphi : R \to F$ there exists a unique $\psi \in Hom(F', F)$ such that the following diagram commutes*

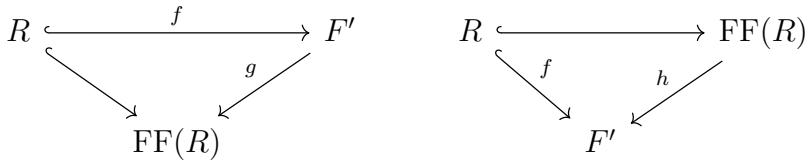
[3]These are very important, but we won't worry too much about them at the moment. They will be formally introduced in chapter 7.

[4]This is the actual universal property.

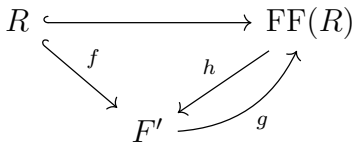


Then there exists one, and only one, isomorphism $g : FF(R) \to F'$ such that $\varphi = g \circ \iota$ always holds.

Proof. If F' satisfies the above property, then there exist by Theorem 3.5.3 unique homomorphisms g, h such that the following diagrams commute



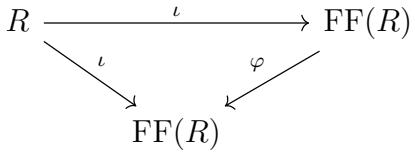
where the unlabelled arrows are the natural embedding. Combining these diagrams, we get one larger commutative diagram



Naming the natural inclusion ι, we see that this implies

$$\iota = g \circ f, h \circ \iota = f \Rightarrow \iota = g \circ h \circ \iota$$

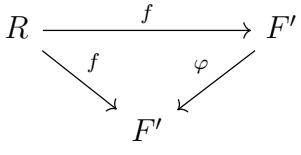
Since ι is injective, we can conclude that $(g \circ h)|_{\mathrm{Im}(\iota)} = \mathrm{Id}_{\mathrm{Im}(\iota)}$. But of course by Theorem 3.5.3 there exists a unique $\varphi \in \mathrm{Hom}(\mathrm{FF}(R), \mathrm{FF}(R))$ such that the following diagram commutes



Since $\varphi = \mathrm{Id}_{\mathrm{FF}(R)}$ works, it follows that $\varphi = \mathrm{Id}_{\mathrm{FF}(R)}$. But of course and φ will do as long as $\varphi|_{\mathrm{Im}(\iota)} = \mathrm{Id}_{\mathrm{Im}(\iota)}$, so it follows that $(g \circ h)|_{\mathrm{Im}(\iota)} = \mathrm{Id}_{\mathrm{Im}(\iota)} \Rightarrow g \circ h = \mathrm{Id}_{\mathrm{FF}(R)}$. Similarly, we also see that

$$f = h \circ g \circ f$$

Since f is injective, we can conclude that $(h \circ g)|_{\mathrm{Im}(f)} = \mathrm{Id}_{\mathrm{Im}(f)}$. But of course by assumption there exists a unique $\varphi \in \mathrm{Hom}(F', F')$ such that the following diagram commutes



Since $\varphi = \mathrm{Id}_{F'}$ works, it follows that $\varphi = \mathrm{Id}_{\mathrm{FF}(R)}$. But of course and φ will do as long as $\varphi|_{\mathrm{Im}(f)} = \mathrm{Id}_{\mathrm{Im}(f)}$, so it follows that $(h \circ g)|_{\mathrm{Im}(f)} = \mathrm{Id}_{\mathrm{Im}(f)} \Rightarrow h \circ g = \mathrm{Id}_{F'}$. Thus, h is the desired isomorphism. That it is unique (in the sense of the corollary statement) follows immediately from the uniqueness of our original choice of h. □

Don't feel too worried if that proof seemed overwhelming or hard to follow, it's our first instance of a technique known as *diagram chasing* which has a habit of being hard formally write out. Take as much time as you need to understand the above proof before moving on.

As one final note, we talked a lot about how to turn rings into fields, but what about recognizing which rings are already fields? To do this, we have the following useful result.

Proposition 3.5.4. *A ring R is a field if and only if its only two ideals are $(0) = \{0\}$ and $(1) = R$.*

Proof. First, suppose that R is a field, and $I \subset R$ an ideal. If $I \neq (0)$, then we can find some non-zero $x \in I$. x is a unit, so for any other $y \in R$ we get $(yx^{-1})x = y \in I$, and hence $I = R$. Now, suppose that there exists some non-zero $x \in R$ which is not a unit. Since R is commutative, $(x) = \{rx \mid r \in R\}$. Thus, since x is not a unit, $rx \neq 1$ always and hence $1 \notin (x)$, so $(x) \neq (1)$. Since $x \neq 0$, $(x) \neq (0)$. □

3.6 Factorial Monoids

Most of the remaining (required) sections of this chapter are all on polynomial rings. Our main concern for polynomial rings, in general, is similar to that of polynomial functions. Namely, we wish to find roots of the polynomials. However, there is an issue with this. Namely, when we view polynomials as rings instead of functions, evaluation becomes a bit of a trickier topic[5]. As such, we wish to find roots without considering evaluation, which leads naturally to the idea of factoring polynomials. To that end, we will take a step back now to understand factoring in a more general context, following a similar section in [Jac09].

For this section, we will work exclusively with commutative monoids M satisfying the cancellation law, that is in the monoid $xy = xz \Rightarrow y = z$.

Definition 3.6.1. For $a, b \in M$, we say that $a \mid b$ (a divides b or is a factor of b) if there exists some $z \in M$ such that $b = za$. a is a proper factor of b if $a \mid b$ but $b \nmid a$. An element $b \in M$ is irreducible if its only proper factors are units, and is prime if $b \mid cd$ implies that $b \mid c$ or $b \mid d$, for any $c, d \in M$.

Note. If $a \mid b$ and $b \mid a$, then one can use the cancellation law to show that $a = zb$, where $z \in M$ is a unit. Also, prime elements are necessarily irreducible. Indeed, suppose that $p \in M$ is prime and $a \mid p$, with $ab = p$. Then $p \mid a$ or $p \mid b$. In the first case we've shown that a is not a proper factor, and we're done. In the second, cancellation law implies that a is a unit, and hence we're done. Note, however, that irreducible elements need not be prime in general.

[5]It can be done without too much difficulty, but you have to be careful.

Definition 3.6.2. A factorization of an element $a \in M$ is an expression of the form $a = p_1 \cdots p_n$, where $p_k \in M$ are irreducible. Such a factorization is called unique if any other factorization can be obtained from the original one by re-ordering elements and multiplying by units.

This finally allows us to define our desired objects.

Definition 3.6.3. M is factorial if every element of M has a unique factorization.

The classic example of one of these is $\mathbb{Z}^*$ under multiplication. We'll now spend the rest of this section defining some equivalent conditions for monoids to be factorial. The first two conditions we'll explore are the following.

Definition 3.6.4. A monoid M is said to satisfy

1. The ascending chain condition (ACC) if there exists no infinite sequence of elements $a_i \in M$ such that a_{i+1} is a proper factor of a_i.

2. The primeness condition if every irreducible element of M is prime.

Theorem 3.6.5. *A monoid M is factorial if and only if it satisfies the ACC and primeness conditions.*

Proof. First, suppose that M is factorial. Suppose that $x, y, z \in M$ are such that x is irreducible and $x \mid yz$. Let $w \in M$ be the element such that $xw = yz$. By the uniqueness of irreducible decompositions, and since x is an irreducible element, we conclude that x (up to multiplication by some unit) must be in the irreducible decomposition of y or z, and hence $x \mid y$ or $x \mid z$. x is therefore prime, and M satisfies the primeness condition. Now, suppose $\{a_i\}_{i\in\mathbb{N}}$ is a sequence of elements in M violating the ACC. Take an irreducible decomposition of a_1, say $a_1 = p_1 \cdots p_n$, and one of a_2, say $a_2 = q_1 \cdots q_m$, where we assume that p_j, q_j are not units (if they were then a_2 would not be a proper factor of a_1). Let $z \in M$ be such that $a_2 z = a_1$. Then by the uniqueness of irreducible decompositions and primeness condition, $q_1 \mid p_1$ or $q_1 \mid p_2 \cdots p_n$. If $q_1 \mid p_1$, then it is just a unit multiple of p_1 as p_1 is irreducible. Otherwise, we can repeat this argument on $p_2 \cdots p_n$ and thereon, finding some $1 \leq j \leq n$ such that $q_1 = up_j$, where $u \in M$ is a unit. We may then apply cancellation law to remove the factor q_1 from both sides, and repeat the argument with q_2, eventually concluding that for each q_j we can find some p_{k_j} and unit $u_j \in M$ such that $q_j = u_j p_{k_j}$ and each k_j is distinct. Since the number of irreducible factors in the decomposition of an element in a factorial monoid is unique, we may conclude (cancelling the q_j on both sides) that since a_2 is a proper factor of a_1, it must have strictly fewer irreducible factors in its decomposition compared to a_1. We may then repeat this argument with a_2, a_3 and so on, eventually finding some $r \in \mathbb{N}$ such that a_r is irreducible. But then a_r cannot have any proper factors, violating our assumption about the nature of $\{a_i\}_{i\in\mathbb{N}}$. Therefore, M satisfies the ACC.

Now, suppose that M satisfies the ACC and primeness conditions. We first show that any $x \in M$ has an irreducible decomposition. If x is irreducible than this is immediate. Otherwise, pick some non-unit irreducible factor $a_1 \in M$ and element $b_1 \in M$ such that $x = a_1 b_1$. Then

b_1 is a proper factor of x. Indeed, suppose that $x \mid b_1$. Let $y \in M$ be such that $xy = b_1$. Then $a_1 b_1 y = b_1 \Rightarrow a_1 y = 1$ and hence a_1 is a unit contrary to our assumption. If b_1 is irreducible, then $a_1 b_1$ is an irreducible decomposition, and we're done. Otherwise, we repeat the above process on b_1, and continue until all factors are irreducible. This process must terminate, as otherwise the b_i sequence constructed along the way would violate the ACC. Thus, every element of M has an irreducible decomposition. We finish by proving the uniqueness of these decompositions. Suppose $x = p_1 \cdots p_n = q_1 \cdots q_m$ are two irreducible decompositions. Since $p_1 \mid q_1 \cdots q_m$, we conclude by the primeness condition that $p_1 \mid q_1$ or $p_1 \mid q_2 \cdots q_m$. We continue this process, repeating the same argument as in the previous part of the proof, to pair up each p_i with a distinct q_j it is a unit multiple of. Cancelling all the p_i must then leave us with only units, giving the uniqueness of the decomposition and making M a factorial domain. □

There's one more idea we'd like to generalize before we move on, namely the concept of greatest common divisors and least common multiples.

Definition 3.6.6. For any two elements $x, y \in M$, we call $z \in M$

1. A least common multiple (LCM) of x, y (or an element of $\text{lcm}(x, y)$) if $x, y \mid z$ and, for any $w \in M$, $x, y \mid w \Rightarrow z \mid w$.

2. A greatest common divisor (GCD) of x, y (or an element of $\gcd(x, y)$) if $z \mid x, y$ and, for any $w \in M$, $w \mid x, y \Rightarrow w \mid z$.

Note. One can check that, in $\mathbb{Z}$, these are equivalent to our traditional notions of LCM and GCD. Furthermore, LCM and GCD of a pair of elements are unique up to units.

Theorem 3.6.7. *Any pair of elements x, y in a factorial monoid M have a GCD and LCM.*

Proof. (Sketch) First, suppose that x is a unit. Then any element of M dividing x is also a unit, $x \mid y$, and one can quickly check that $x \in \gcd(x, y), y \in \text{lcm}(x, y)$. Now, suppose that x, y are not units. Let $x = p_1 \cdots p_n, y = q_1 \cdots q_m$ be their irreducible decompositions, where we assume that the irreducible factors are not units. Then by a similar argument as in Theorem 3.6.5, any element of M dividing x, y must have as irreducible factors only factors appearing in x, y (up to multiplication by units), and none of these factors can appear more times than they did in x, y. It follows that taking the product of all common irreducible factors of x, y (counting multiplicity and considering factors up to multiplication by units) gives a GCD of x, y. A similar argument shows that taking the product of the minimum number of irreducible factors needed to have all those in x, y gives an LCM. □

Note. This result and the GCD/LCM constructed are just extensions of the same result section 1.1.

It turns out that the GCD and factoriality of a monoid are closely related. This is related to the following two results, both of which are proven in [Jac09] and will not be proven here.

Proposition 3.6.8. *1. If any pair of elements in a monoid M have a GCD (this is referred to as the GCD condition), then so does any finite collection of elements in M.*

2. The GCD condition implies the primeness condition.

Corollary 3.6.8.1. *M is factorial if and only if it satisfies the ACC and GCD conditions.*

Proof. If M is factorial, then it satisfies ACC by Theorem 3.6.5 and GCD by Theorem 3.6.7. If M satisfies ACC and GCD, then by proposition 3.6.8 it satisfies the primeness condition, and hence by Theorem 3.6.5 is factorial. □

3.7 PIDs and Euclidean Domains

This is again following a similar section in [Jac09]. We start with a basic observation.

Proposition 3.7.1. *Let R be a domain. Then for any $x, y \in R$, $x \mid y$ if and only if $(y) \subset (x)$.*

Proof. If $x \mid y$, then there exists some $z \in R$ such that $xz = y$, and hence $y \in (x) \Rightarrow (y) \subset (x)$. If $(y) \subset (x)$, then $y \in (x)$, so there exists some $z \in R$ such that $xz = y$. Therefore, $x \mid y$. □

Note. By $x \mid y$, we mean here that $x \mid y$ in the commutative monomial $(R, \cdot)$, which since R is a domain satisfies the cancellation law.

Essentially, the above proposition says that we may study ideals generated by one element (these are called *principle ideals*) instead of studying divisibility directly. Ideally, then, we'd like all the ideals in R to be principle, and we call R a *principle ideal domain* (PID) if it satisfies this. Specifically, we'd like to find rings such that $(R, \cdot)$ is factorial. We call such rings *unique factorization domains* (UFDs).

Theorem 3.7.2. *If R is a PID, then R is a UFD.*

Proof. First, we show that R satisfies the ACC. Suppose

$$(a_1) \subset (a_2) \subset \cdots \subset (a_n) \subset \cdots$$

is a chain of principle ideals in R. Then $A = \bigcup_{i \in \mathbb{N}} (a_i)$ is an ideal, and hence since R is a PID there exists some $x \in R$ such that $A = (x)$. Thus, $x \in (a_n)$ for some $n \in \mathbb{N}$, but also by definition $(a_n) \subset (x)$. We conclude that $(a_n) = (x)$, so a_n, x differ only by a unit. The ACC follows from this and proposition 3.7.1. Next, we show that R satisfies the GCD condition. In particular, pick any $x, y \in R$. Then since R is a PID, there exists some $z \in R$ such that $(x, y) = (z)$. Since $(x), (y) \subset (z)$, $z \mid x, y$. Now, suppose that $w \in R$ is some other element such that $w \mid x, y$. Then $(x), (y) \subset (w)$, so since ideals are closed under addition we conclude that $(x, y) \subset (w) \Rightarrow (z) \subset (w) \Rightarrow w \mid z$. Thus, z is the desired GCD. The result now follows by corollary 3.5.3.1. □

Actually proving that rings are PIDs can be somewhat tricky, but (as we'll see later in this section) having a long division algorithm in R like that of $\mathbb{Z}$ is sufficient to make a ring a PID. Thus, we generalize long division.

Definition 3.7.3. A domain R is Euclidean if there exists a map $\delta : R \to \mathbb{Z}^*$ such that for any $a, b \in R^*$, there exists some $q, r \in R$ such that $a = bq + r$, where $\delta(r) < \delta(b)$.

Note. δ is our way of measuring the "size" of our remainder r. In the case of $\mathbb{Z}$, we'd have $\delta(x) = |x|$.

Theorem 3.7.4. *Euclidean domains are PIDs.*

Proof. Suppose R is Euclidean. Let I be any ideal in R, and suppose that $I \neq (0)$. Let $b \in I$ be a non-zero element in I such that $\delta(b)$ is minimal. Suppose $\exists a \in I$ such that $b \nmid a$. Since R is Euclidean, there exist $q, r \in R$ such that $a = qb + r$, and $\delta(r) < \delta(b)$. But $b \nmid a \Rightarrow r \neq 0$, and $a - qb \in I \Rightarrow r \in I$, so this contradicts the minimality of $\delta(b)$. Thus, every element of I is a multiple of b, so $I = (b)$. □

Corollary 3.7.4.1. *Euclidean domains are UFDs.*

3.8 Polynomial Rings

We follow the results of [Jac09] here again, although the organization of the material has been significantly changed.

Polynomial rings are, without a doubt, the most important example of rings for algebra. In a way, the entirety of Part IV is dedicated to the study of polynomial rings and their structure. So without further ado, let's get to it.

Definition 3.8.1. Let R be a ring. The polynomial ring over R in one variable, denoted $R[x]$, is the set $R_c^{\mathbb{Z}_{\geq 0}}$ (infinite sequences in R with finitely many non-zero elements), with element-wise addition and multiplication defined by

$$((a_i)_{i\in\mathbb{Z}_{\geq 0}}(b_i)_{i\in\mathbb{Z}_{\geq 0}})_j = \sum_{i+k=j} a_i b_k$$

Note. Our zero here is $(0, 0, 0, \dots)$, and our identity is $(1, 0, 0, \dots)$.

It's not immediate from this definition the connection between this ring and polynomials as we know them. To make this connection more clear, we usually adopt the following notation. First, denote the sequence with a 1 in the nth position by x^{n-1} for $n \geq 1$. One can note that

$$(c, 0, 0, \dots) \cdot (a_0, a_1, \dots) = (ca_0, ca_1, \dots)$$

Thus, any sequence can be written uniquely[6] in the following form

$$(a_0, a_1, \dots) = (a_0, 0, \dots) + (a_1, 0, \dots)x + (a_2, 0, \dots)x^2 + \cdots$$

Denoting $(c, 0, \dots)$ by just c, this becomes

$$(a_0, a_1, \dots) = a_0 + a_1 x + a_2 x^2 + \cdots$$

making the connection much more clear. In fact, we will always choose to use this notation, as it makes everything much more intuitive. One can check that multiplying out two expressions of the above form in the way that you normally would for polynomials will give the correct result, only furthering this connection.

[6] This technically needs to be proven, but I don't think the proof is particularly enlightening or complicated.

Note. This polynomial ring is actually distinct from the ring of polynomial functions over R, which we'll cover a bit later in this section.

We can then continue to generalize this to multivariable polynomial rings.

Definition 3.8.2. We define the multivariable polynomial ring over R in the following inductive manner. That is, we denote the polynomial ring in n variables over R by $R[x_1, \dots, x_n]$, and define it by $R[x_1, \dots, x_n] = R[x_1, \dots, x_{n-1}][x_n]$.

It's at this point where the notation we've been using, writing polynomial rings in the same way as polynomial functions, becomes incredibly convenient. For example, we'd get that in $R[x_1, x_2]$

$$\begin{aligned}((a, b, c, 0, \dots), (a, b, 0, \dots), (a, 0, \dots), (0, \dots), \dots) &= (a + bx_1 + cx_1^2) + (a + bx_1)x_2 + ax_2^2 \\ &= a + bx_1 + cx_1^2 + ax_2 + bx_1x_2 + ax_2^2\end{aligned}$$

We call terms of the form $x_1^{k_1} \cdots x_n^{k_n}$ *monomials* in $R[x_1, \dots, x_n]$. Like in $R[x]$, any element of $R[x_1, \dots, x_n]$ can be written uniquely as the sum of finitely many monomial elements summed together, with each monomial multiplied by some non-zero coefficient in R[Jac09][7].

We move away from polynomials now, for a moment, and talk instead about the related concept of *adjoined rings*.

Definition 3.8.3. Let R be a subring of a ring S, and let $U \subset S$. Then R adjoin U, denoted $R[U]$, is the subring of S generated by $R \cup U$.

Proposition 3.8.4. *Suppose R is a subring of S, and $U, V \subset S$. Then $R[U][V] = R[U \cup V]$.*

Proof. Since $R[U \cup V]$ is a subring of S containing R and U, $R[U] \subset R[U \cup V]$. Thus, $R[U], V$ are contained in $R[U \cup V]$, so $R[U][V] \subset R[U \cup V]$. Furthermore, $R[U][V]$ contains R, U, and V, so $R[U \cup V] \subset R[U][V]$. □

Proposition 3.8.5. *Suppose R is a subring of S, and $u \in S$. Then $R[u] = R[\{u\}]$ is the subring of S composed of expressions of the form*

$$\sum_k a_k u^k, a_k \in R, \textit{finite sums}$$

Proof. That all expressions of this form are in $R[u]$ is immediate. Thus, since (as can be quickly checked) since this is a subring containing R and u, we get the desired result. □

Note. The above result suggests a strong connection between $R[x]$ and $R[u]$, the latter is a way of evaluating the polynomial expressions in the former. This is also why we take the blatant abuse of notation of denoting them in the same way.

Note. The above proposition in fact works for any $R[U]$, extending the expressions in the obvious way. The proof is identical.

[7] The proof of this is an inductive argument, and rather tedious.

The above observation will now be expanded upon in the following extremely important theorem. For this theorem, we take the convention of the "constants" in our standard notation for $R[x_1, \ldots, x_n]$ being an embedding of R in $R[x_1, \ldots, x_n]$.

Theorem 3.8.6. *Let R be a ring, and S any ring containing R. Then for any $n \in \mathbb{N}$ and $u_1, \ldots, u_n \in S$, there exists a unique homomorphism $\varphi : R[x_1, \ldots, x_n] \to S$ fixing R and taking $x_i \mapsto u_i$.*

Proof. Since $R[x_1, \ldots, x_n] = R[x_1, \ldots, x_{n-1}][x_n]$, it suffices by induction to prove this for the case $n = 1$. Note that any $a \in R[x]$ can be written uniquely in the form

$$a = \sum_{k \geq 0} a_k x^k$$

where $a_k \in R$ and only finitely many $a_k \neq 0$. We'll define $\varphi : R[x] \to S$ by

$$\varphi(a) = \sum_{k \geq 0} a_k u^k$$

It is fairly simple to check that this is in fact a ring homomorphism. For uniqueness, note that if $\varphi|_R = \mathrm{Id}_R$ and $\varphi(x) = u$, then since φ is a ring homomorphism

$$\varphi(a) = \varphi(a) = \sum_{k \geq 0} \varphi(a_k)\varphi(x)^k = \sum_{k \geq 0} a_k u^k$$

□

Note. What we mean by "R is a subring of S" can often be a bit loose. We make no distinction between R being a subring of S, or R embedding into S via a ring homomorphism. Indeed, identifying R with its embedding, we can see that there really is no difference between the two situations.

Corollary 3.8.6.1. *Fix $n \in \mathbb{N}$. Suppose K is any ring containing R and distinguished elements $y_1, \ldots, y_n \in K$ such that*

1. *$R[y_1, \ldots, y_n] = K$*
2. *For any other ring S containing R and $u_1, \ldots, u_n \in S$ there exists a unique homomorphism $\varphi : K \to S$ which fixes R and satisfies $\varphi(y_i) = u_i$.*

Then $K \cong R[x_1, \ldots, x_n]$.

Proof. By assumption, there exists a unique $\varphi \in \mathrm{Hom}(K, R[x_1, \ldots, x_n])$ fixing R such that $\varphi(y_i) = x_i$. Furthermore, there exists by Theorem 3.8.6 a unique $\psi \in \mathrm{Hom}(R[x_1, \ldots, x_n], K)$ fixing R such that $\psi(x_i) = y_i$. Then $(\varphi \circ \psi)(x_i) = x_i$ and $\varphi \circ \psi$ fixes R, so by the uniqueness property in Theorem 3.8.6 $\varphi \circ \psi = \mathrm{Id}_{R[x_1, \ldots, x_n]}$. An identical argument shows that $\psi \circ \varphi = \mathrm{Id}_K$, so φ is an isomorphism. □

Note. What the above argument is really saying is that, up to isomorphism, there's only one way to define a multivariable polynomial ring over R.

Corollary 3.8.6.2. *For any permutation $\sigma \in S_n$, there exists a unique automorphism $\varphi_\sigma \in \mathrm{Isom}(R[x_1, \ldots, x_n])$ such that $\varphi(x_i) = x_{\sigma(i)}$.*

Proof. Left to the reader. This has essentially the same proof as corollary 3.8.6.1. □

We call the homomorphism given by Theorem 3.8.6 the *evaluation homomorphism*. Our main goal from now on will be to understand the kernels of these homomorphisms. Indeed, we have a good understanding of the structure of $R[x]$, and $R[u] \cong R[x]/\ker\varphi$, so understanding the kernel of evaluation homomorphisms teaches us a lot about the element of rings containing R. It will also, unsurprisingly, have deep connections to identification of roots of polynomials. On that note, let's actually define these polynomial functions.

Definition 3.8.7. The ring of polynomial functions in n variables over R, denoted $\mathcal{P}_n(R)$, is the subset of the ring of functions from $R^n \to R$ of the form

$$f(u_1, \ldots, u_n) = \sum_{(k_1,\ldots,k_n)} a_{k_1,\ldots,k_n} u_1^{k_1} \cdots u_n^{k_n}$$

where $a_{k_1,\ldots,k_n} \in R$ and the above sum is finite.

In order to understand the connections between these, our polynomial rings, and roots of polynomials, we'll need to study the factoring of polynomials.

3.9 Factoring Polynomials

Again, this follows some similar sections in [Jac09]. In order to begin factoring, we first need to understand the notion of degree.

Definition 3.9.1. Let $f \in R[x]$ be a polynomial. Then we may write, for some $n \in \mathbb{N}$ and $a_n \neq 0$

$$f = \sum_{k=0}^{n} a_k x^k$$

We define the degree of f, denoted $\deg(f)$, to be n, and call a_n the leading coefficient of f. By convention, we define that $\deg(0) = -\infty$, where $-\infty$ has the expected arithmetic properties.

Note. This is well-defined by the uniqueness of this method of expressing f. For multivariable polynomials, the notion of degree gets a little tricker. One can either look at their degree in a particular variable, or their total degree. We'll look at that a bit more in the next section.

A couple of the properties of degree are fairly immediate, and will not be proven here.

Proposition 3.9.2. *Suppose $f, g \in R[x]$. Then*

1. $\deg(f+g) \leq \max(\deg(f), \deg(g))$

2. $\deg(fg) = \deg(f)\deg(g)$

Using the notion of degree, we can immediately start proving some useful results.

Proposition 3.9.3. *If R is a domain, then $R[x_1, \ldots, x_n]$ is a domain.*

Proof. Since $R[x_1, \ldots, x_{n-1}][x_n] = R[x_1, \ldots, x_n]$, it suffices to show that $R[x]$ is a domain. Indeed, suppose that $f, g \in R[x]^*$. Then $\deg(f), \deg(g) \geq 0$, so $\deg(fg) \geq 0 \Rightarrow fg \neq 0$. □

The next result is perhaps one of the most fundamental in this section, namely that polynomial long division can be extended to arbitrary polynomial rings.

Theorem 3.9.4. *Suppose $f, g \in R[x]$, with $g \neq 0$. Let $m = \deg(g)$ and $b_m \neq 0$ be the leading coefficient of g. Then there exists some $k \in \mathbb{N}$, $q, r \in R[x]$ with $\deg(r) < \deg(g)$ such that*

$$b_m^k f = qg + r$$

Proof. If $\deg(f) < \deg(g)$, then we simply take $q = 0, r = b_m f, k = 1$ to get the desired result. Otherwise, define

$$f_1 = b_m f - a_n x^{n-m} g$$

where $n = \deg(f)$, and a_n is the leading coefficient of f. It's clear that $\deg(f_1) \leq \deg(f) - 1$. If $\deg(f_1) < \deg(g)$, then we're done. Otherwise, we can repeat this process with f_1, getting an $f_2, \ldots, f_\ell \in R[x]$, continuing until $\deg(f_\ell) < \deg(g)$. This is guaranteed to terminate, since $\deg f_[k+1) \leq \deg(f_k) - 1$. In the end, we get

$$f_\ell = b_m f_\ell - a^{(\ell-1)} x^{\deg(f_{\ell-1}-m)} g$$

where $a^{(k)}$ is the leading coefficient of f_k. But in turn we know that

$$f_{\ell-1} = b_m f_{\ell-2} - a^{(\ell-2)} x^{\deg(f_{\ell-2}-m)} g$$

and so on. Expanding out, this gives

$$f_\ell = b_m^2 f_{\ell-2} - (a^{(\ell-1)} x^{\deg(f_{\ell-1}-m)} + a^{(\ell-2)} x^{\deg(f_{\ell-2}-m)}) g$$

Continuing this process, and calling the collected terms which multiply g by $q \in R[x]$, we see that

$$f_\ell = b_m^\ell f - qg \Rightarrow b_m^\ell f = qg + f_\ell$$

Since $\deg(f_\ell) < \deg(g)$, this completes the proof. □

Note. This proof also gives you an algorithm for calculating this "long division".

If b_m is a unit, then since $b_m^k \neq 0$ we get the following, more familiar result.

Corollary 3.9.4.1. *If $f, g \in R[x]$, and $g \neq 0$, then there exists unique $q, r \in R[x]$ with* $\deg(r) < \deg(g)$ *such that*

$$f = qg + r$$

Furthermore, in FF($R[x]$), we get

$$\frac{f}{g} = q + \frac{r}{g}$$

Proof. Existence is given by Theorem 3.9.4 and dividing out by the unit. FOr uniqueness, suppose q_1, r_1 and q_2, r_2 were two such pairs. Then

$$q_1 g + r_1 = q_2 g + r_2 \Rightarrow (q_1 - q_2)g = r_2 - r_1$$

Taking the degree of both sides, we get

$$\deg(q_1 - q_2)\deg(g) \leq \max(\deg(r_2), \deg(r_1)) < \deg(g)$$

We are therefore left with two possibilities. First, suppose that $\deg(g) = 0$. Then $\deg(r_1)$, $\deg(r_2) < 0 \Rightarrow r_1 = r_2 = 0$, so $(q_1 - q_2)g = 0 \Rightarrow q_1 = q_2$. Otherwise, we must conclude that $q_1 = q_2$, which in turn implies that $r_1 = r_2$. □

Note. In this case, we call q and r the quotient and remainder of f/g.

Using this, we can start factoring our polynomials properly. In order to do so, we'll need a bit of notation. Suppose $R \subset S$ is a subring, $f \in R[x]$, and $a \in S$. Then we'll use $f(x)$ to denote the polynomial in $R[x]$, and $f(a)$ to denote the *evaluation* (i.e. image under the evaluation homomorphism) of f at a.

Corollary 3.9.4.2 (Remainder Theorem)**.** *Suppose $f(x) \in R[x]$ and $a \in R$. Then there exists a unique $q(x) \in R[x]$ such that*

$$f(x) = (x - a)q(x) + f(a)$$

Proof. By corollary 3.9.4.1, there exist some unique $q(x), r(x) \in R[x]$ such that $\deg(r(x)) < \deg(x - a)$ and

$$f(x) = (x - a)q(x) + r(x)$$

In particular, since $\deg(r(x)) < \deg(x - a) = 1$, $r(x) \in R$ (or more properly its embedding into $R[x]$). Thus, $r(x)$ is fixed by any evaluation homomorphism. In particular, we can then evaluate both sides of the above equation at a to get

$$f(a) = (a - a)q(a) + r(x) \Rightarrow r(x) = f(a)$$

□

We also get the following result immediately from the above corollary.

Corollary 3.9.4.3 (Factor Theorem)**.** *Suppose $f(x) \in R[x]$ and $a \in R$. Then $(x - a) \mid f(x)$ if and only if $f(a) = 0$.*

There are two more results we can get out of these theorems, namely on the number of roots polynomials over fields have and on the structure of polynomial rings over a field.

Definition 3.9.5. Let F be a field, $f(x) \in F[x]$ be such that $\deg(f) > 0$. We call $a \in F$ a root of f if $f(a) = 0$.

Corollary 3.9.5.1. *Let $f(x) \in F[x]$ be a polynomial of degree $n \geq 1$. Then $f(x)$ has at most n distinct roots in F.*

Proof. Let $a_1, \ldots, a_m \in F$ be distinct roots of F. We show, by induction on r, that $\prod_{k=1}^{r}(x - a_r) \mid f(x)$, from which the result immediately follows. The case of $r = 1$ is given by the factor theorem. Now, suppose this holds for some $r \geq 1$ such that $r < m$. Then there exists some $g(x) \in F[x]$ such that

$$g(x)\prod_{k=1}^{r}(x - a_k) = f(x)$$

evaluating both sides at a_{r+1}, we get that since $f(a_{r+1}) = 0$ and $(a_{r+1} - a_k) \neq 0$ for $1 \leq k \leq r$, $g(a_{r+1}) = 0$. Thus, by the factor theorem, there exists some $h(x) \in F[x]$ such that $g(x) = (x - a_{r+1})h(x)$, and so

$$h(x)\prod_{k=1}^{r+1}(x - a_k) = f(x) \Rightarrow \prod_{k=1}^{r+1}(x - a_k) \mid f(x)$$

□

Corollary 3.9.5.2. *If F is a field, then $F[x]$ is a PID.*

Proof. Let $I \subset F[x]$ be an ideal. Let $f(x) \in I$ be a non-zero polynomial of minimal degree. Take any other polynomial $g(x) \in F[x]$. By corollary 3.9.4.1, there exist $q(x), r(x) \in F[x]$ such that $\deg(r) < \deg(f)$ and

$$g(x) = q(x)f(x) + r(x) \Rightarrow r(x) = g(x) - q(x)f(x)$$

Since $r(x) \in I$, we conclude by the minimality of the degree of $f(x)$ that $r(x) = 0$. Thus, $f(x) \mid g(x)$. Since every element of I is divisible by $f(x)$, and $f(x) \in I$, it follows that $I = (f)$. □

Note. This result is false for multivariable polynomial rings, a decent example of this can be found in [Jac09].

There's a strong connection between evaluation homomorphisms and the irreducibility of polynomials we're building up to here, but first we'll need the following definitions.

Definition 3.9.6. Let $R \subset S$ be a subring, and pick $a \in S$. We call a algebraic over R if there exists a monic (i.e. polynomial with leading coefficient 1) $f(x) \in R[x]$ such that $f(a) = 0$. Otherwise, we call it transcendental over R. For algebraic elements, we call a monic polynomial $f(x) \in R[x]$ such that $f(a) = 0$ a minimal polynomial of a over R.

Proposition 3.9.7. *Let $F \subset K$ be a subfield, and pick any $a \in K$ algebraic over F. Then a has a unique minimal polynomial.*

Proof. Note that the set of all polynomials of which a is a root is an ideal I. By corollary 3.9.5.2, $F[x]$ is a PID. Thus, there exists some non-zero $f(x) \in F[x]$ such that $I = (f(x))$. In particular, since we're operating over a field, we can choose for $f(x)$ to be monic. Since any polynomial in I is a multiple of $f(x)$, there is no other monic polynomial in I of degree less than or equal to $f(x)$. □

In this case, we take to calling $f(x)$ the *minimal polynomial* of a over F.

Theorem 3.9.8. *Suppose $F \subset K$ is a subfield, and $u \in K$ is algebraic over F with minimal polynomial $f(x) \in F[x]$. Then $F[u]$ is a field if $f(x)$ is irreducible, and is not a domain otherwise.*

Proof. First, suppose that $f(x)$ is irreducible. Let $\varphi \in \mathrm{Hom}(F[x], K)$ be the evaluation homomorphism. We can first note that $F[u] \cong \varphi(F[x])$, and hence

$$F[u] \cong F[x]/\ker(\varphi)$$

Every ideal in $F[u]$ is therefore the image of an ideal in $F[x]$ containing $\ker(\varphi)$ under the quotient map. By the proof of proposition 3.9.7, $\ker(\varphi) = (f(x))$. Therefore, ideals in $F[u]$ correspond to ideals in $F[x]$ containing $f(x)$. Suppose J were such an ideal. Then since $F[x]$ is a PID and F a field, there exists a monic $g(x) \in F[x]$ such that $J = (g(x))$. Thus, since $f(x) \in (g(x)), g(x) \mid f(x)$. But this implies that $g(x)$ is a unit, and hence $g(x) \in F \Rightarrow g(x) = 1$. Thus, the only two ideals in $F[u]$ are the zero ideal and $F[u]$, making $F[u]$ a field. Now, suppose that $f(x)$ is reducible, say with factoring $f(x) = g(x)h(x)$, where $\deg(g), \deg(h) \geq 1$. By the minimality of the degree of f, $g(u), h(u) \neq 0$. However, $f(u) = g(u)h(u) = 0$. Thus, $F[u]$ is not a domain. □

You may think we're done with factoring, but you'd be wrong. We can, in fact, build up to one last much stronger result. Namely, that if R is a UFD, then so is $R[x]$. To do this, we'll need to introduce the concept of the *content* of a polynomial.

Definition 3.9.9. Let R be a UFD, and $f(x) \in R[x]^*$. Writing

$$f(x) = a_n x^n + \cdots + a_1 x + a_0$$

We define the content of f, denoted $c(f)$, by

$$c(f) = \mathrm{GCD}(a_1, \ldots, a_n)$$

If $c(f)$ is a unit, we call f primitive.

Note. The content is only-well defined up to multiplication by units.

In the following, we will need an identity on GCDs which we will not prove. A proof can be found in [Jac09].

Proposition 3.9.10. *Suppose M is a factorial monoid, and $a, b, c \in M$. Then up to multiplication by units*

1. *$GCD(GCD(a,b),c) = GCD(a,b,c)$*
2. *$GCD(ac,bc) = cGCD(a,b)$*

Note. The first part of this proposition is a little vague, but essentially means that this equality works (up to multiplication by units in M) for any choice of GCD at any point in evaluating the equations.

Using this, we get the following.

Proposition 3.9.11. *Suppose R is a UFD, and $f(x) \in R[x]^*$. Then there exists a primitive polynomial $g(x) \in R[x]^*$ and constant $a \in R$ such that $f(x) = ag(x)$. Furthermore, if $f(x) = bh(x)$ is another such decomposition, then there exists a unit $u \in R$ such that $b = ua$.*

Proof. Take some $a \in c(f)$ (i.e. choose a particular content). Let $f(x) = c_n x^n + \cdots + c_0$. Then by definition, $a \mid c_k$ for each $0 \leq k \leq n$, and setting $d_k = c_k/a$ and $g(x) = d_n x^n + \cdots + d_0$ we get by proposition 3.9.10 that $c(g)$ is a unit, and hence g is primitive. $f(x) = ag(x)$ is therefore the desired decomposition. Now, suppose that $f(x) = bh(x)$ is another such decomposition. Write $h(x) = k_n x^n + \cdot + k_0$. Since h is primitive, we get that since $c(f) = c(ag(x))$, $b \in c(f)$. Thus, there exists $u \in R$ such that $b = ua$. □

We next generalize the content of polynomials to polynomials over fields in the following manner.

Lemma 3.9.12. *Suppose R is a UFD, $F = FF(R)$, and $f(x) \in F[x]^*$. Then there exists some $a \in F$ and primitive polynomial $g(x) \in R[x]^*$ such that $f(x) = ag(x)$. Furthermore, if $f(x) = bh(x)$ is another such decomposition, then there exists a unit $u \in R$ such that $a = bu$.*

Proof. Let $f(x) = c_n x^n + \cdots + c_0$. Since $F = \mathrm{FF}(R)$, there exists some $\alpha \in R^*$ such that $\alpha c_k \in R$ for every $0 \leq k \leq n$. Then $\alpha f(x) \in R[x]^*$, and hence by proposition 3.9.11 there exists some $a \in R$ (in particular $a \in c(\alpha f(x))$) and primitive $g(x) \in R[x]^*$ such that $\alpha f(x) = ag(x)$. Thus, $f(x) = \frac{a}{\alpha} g(x)$ is the desired decomposition. Now, suppose that $ag(x), bh(x)$ are two such decompositions. Then $(\alpha a)g(x) = (\alpha b)h(x)$ are in $R[x]^*$, so by proposition 3.9.11 there exists some unit $u \in R$ such that $\alpha a = u\alpha b \Rightarrow a = ub$. □

In the case of the above lemma, we call the a the *field content* of $f(x)$.

Lemma 3.9.13 (Gauss's Lemma)**.** *The product of primitive polynomials is primitive.*

Proof. Suppose $g(x), h(x)$ are primitive, but $f(x) = g(x)h(x)$ is not. Then there exists an irreducible, and hence prime, $p \in R^*$ such that $p \nmid g(x), h(x)$, but $p \mid f(x)$. Note that since p is prime, $R' = R/(p)$ is a domain[8]. Projecting all out polynomials into $R'[x]$, we get $g(x), h(x) \neq 0$ but $f(x) = 0$. Thus, $R'[x]$ is not a domain, which contradicts proposition 3.9.3. □

[8] We technically have not proven this yet, but it is not too hard to check

We can now, finally, start proving the relevant results.

Theorem 3.9.14. *If $f(x) \in R[x]$ has degree at least one, where R is a UFD and $F = FF(D)$, then $f(x)$ is irreducible in $R[x]$ if and only if it is irreducible in $F[x]$.*

Proof. If $f(x)$ is irreducible in $F[x]$, then irreducibility in $R[x]$ is immediate. Now, suppose that $f(x)$ is irreducible in $R[x]$, and is of degree at least one. By lemma 3.9.12, there exists some $a \in F$ and primitive $g(x) \in R[x]$ such that $f(x) = ag(x)$. Suppose f were reducible in $F[x]$, say $f(x) = h(x)k(x)$. We note that since all non-zero constants in F are invertible, $\deg(h), \deg(k) \geq 1$. By lemma 3.9.12, there exist some $b, c \in F$ and primitive $f_1(x), f_2(x) \in R[x]$ such that $h(x) = bf_1(x), k(x) = cf_2(x)$. By Gauss's lemma, $f_1(x)f_2(x)$ is primitive, so since $f(x) = (bc)(f_1(x)f_2(x))$ we conclude that there exists some unit $u \in R$ such that $uf_1(x)f_2(x) = f(x)$. But then $f(x)$ would be reducible in $R[x]$, contradicting our assumption. □

Theorem 3.9.15. *If R is a UFD, then so is $R[x]$.*

Proof. Suppose $f(x) \in R[x]^*$ is irreducible. Then it is irreducible in $F[x]^*$. Since $F[x]$ is a PID, it is a UFD. Thus, $f(x)$ is prime in $F[x]$, and hence prime in $R[x]$. This shows that the primeness condition is satisfied. Now, suppose that $f(x), g(x) \in R[x]^*$ are such that f is a proper factor of g. Then either $\deg(f) < \deg(g)$, or there exists some non-unit $u \in R$ such that $g(x) = uf(x)$. Since R is a UFD, it satisfies the ACC, so there can exist no infinite chain of proper factors of u violating the ACC and hence no infinite chain of proper factors of $g(x)$ violating the ACC such that any degree less than or equal to $g(x)$ has infinitely many polynomials of that degree in the chain. Thus, $R[x]^*$ satisfies the ACC, making $R[x]$ a UFD. □

Corollary 3.9.15.1. *If R is a UFD, then so is $R[x_1, \ldots, x_n]$.*

Proof. Follows by induction on n. □

Note. Since multivariable polynomial rings over fields are not PIDs, as we saw earlier in this section, this shows that UFDs need not be PIDs.

3.10 Some Consequences of Factoring

This section, again, follows [Jac09]. We look at two interesting results which follow from our discoveries about polynomial factoring in the previous section. The first is a characterization of when polynomial rings and polynomial rings of functions are the same.

Theorem 3.10.1. *Let F be a field. Then $\mathcal{P}_n(F) \cong F[x_1, \ldots, x_n]$ if and only if F is infinite.*

Proof. If F is finite, then $|\mathcal{P}_n(F)| \leq (n^{|F|})^{|F|}$ and $|F[x_1, \ldots, x_n]| = \infty$, so $F[x_1, \ldots, x_n] \ncong \mathcal{P}_n(F)|$. Now, suppose that F is infinite. There is an obvious homomorphism $\varphi : F[x_1, \ldots, x_n] \to \mathcal{P}_n(F)$ given by $\varphi(f)(a_1, \ldots, a_n) = f(a_1, \ldots, a_n)$. We wish to show that this is an isomorphism. That it is surjective is clear, so we just need to check injectivity. For this, it suffices to show that any non-zero $f(x_1, \ldots, x_n) \in F[x_1, \ldots, x_n]$ has some $a_1, \ldots, a_n \in F$

such that $f(a_1, \ldots, a_n) \neq 0$. We proceed by induction on n. First, suppose that $n = 1$. Then if $f(a) = 0$, a is a root of f. f can have at most $\deg(f) < \infty$ roots, so since $|F| = \infty$ there exists some $a \in F$ such that $f(a) \neq 0$, as required. Now, suppose that the result holds for some $n \geq 1$, and $f \in F[x_1, \ldots, x_{n+1}]$. Then we may write that

$$f(x_1, \ldots, x_{n+1}) = \sum_{k=0}^{r} f_k(x_1, \ldots, x_n) x_{n+1}^k$$

where $f_k \in F[x_1, \ldots, x_n]$, and we may assume without loss of generality that $f_r \neq 0$. By the inductive hypothesis, there exist $a_1, \ldots, a_n \in F$ such that $f_r(a_1, \ldots, a_n) \neq 0$, and hence $f(a_1, \ldots, a_n, x_{n+1}) \in F[x_{n+1}]$ is non-zero. By the case $n = 1$, there is therefore some $a_{n+1} \in F$ such that $f(a_1, \ldots, a_{n+1}) \neq 0$, as required. □

The second concerns the structure of finite subgroups of fields, and becomes quite important in Galois theory.

Theorem 3.10.2. *Any finite subgroup of the multiplicative group F^* is cyclic.*

Proof. Let $G \subset F^*$ be a finite subgroup of the multiplicative group F^*, and let $n = \exp(G)$. Then every element of G must be a root of the polynomial $x^n - 1 \in F[x]$. But $x^n - 1$ can have at most n roots, so there are exactly n elements in G, making it cyclic. □

We can then combine these results to get the following.

Theorem 3.10.3. *Let F be a finite field such that $|F| = q$. Then $\mathcal{P}_n(F) \cong F[x_1, \ldots, x_n]/I$, where $I = (x_1^q - x_1, \ldots, x_n^q - x_n)$.*

Proof. Let $\varphi : F[x_1, \ldots, x_n] \to \mathcal{P}_n(F)$ be the homomorphism from Theorem 3.10.1. It suffices to show that $I = \ker(\varphi)$. Since F is finite, F^* is a cyclic group under multiplication. Thus, $a^q = a$ for any $a \in F$, so certainly $x_k^q - x_k$ are in I. To show that I is generated by the desired polynomials, there are two steps.

First, we show that any $f \in F[x_1, \ldots, x_n]$ of degree strictly less than q in every x_k is not in I. For this, we proceed by induction on n in an identical manner to Theorem 3.10.1. The case $n = 1$ is clear, since a polynomial of degree $< q$ cannot have q roots. The result therefore follows by induction.

Second, we show that any $f \in F[x_1, \ldots, x_n]$ can be written in the form

$$f(x_1, \ldots, x_n) = \sum_{k=1}^{n} f_k(x_1, \ldots, x_n)(x_k^q - x_k) + f_0(x_1, \ldots, x_n)$$

where $f_k \in F[x_1, \ldots, x_n]$ and f_0 is of degree $< q$ in every variable. This implies the desired result, as then $f_0 \in I$ if and only if $f_0 = 0$. It suffices to consider the case of f being a monomial. Consider any monomial of the form $x_1^{j_1} \cdots x_n^{j_n}$. Then for each $1 \leq k \leq n$, there exist $q_k, r_k \in F[x_k]$ such that $x_k^{j_k} = q_k(x_k)(x_k^q - x_k) + r_k(x_k)$, where $\deg(r_k) < q$. Thus,

$$x_1^{j_1} \cdots x_n^{j_n} = (q_1(x_1)(x_1^q - x_1) + r_1(x_1)) \cdots (q_n(x_n)(x_n^q - x_n) + r_n(x_n))$$

Expanding out the expression on the right-hand side, we see that the only term without of factor of $x_k^q - x_k$ for some $1 \leq k \leq n$ is $r_1(x_1) \cdots r_n(x_n)$, which since $\deg(r_k) < q$ is a polynomial of degree $< q$ in every variable, as required. □

3.11 Irreducibility Criteria

In section 3.9 we talked a lot about irreducible polynomials, but I never gave you any tools for recognizing them! This section aims to rectify that, following an identical section in [Lan05][9]. There's not much comment to be made here, it's just three theorems useful for this purpose.

Theorem 3.11.1 (Eisenstein's Criteria)**.** *Let R be a UFD, and $F = FF(R)$. Let $f(x) = a_n x^n + \cdots + a_0 \in R[x]$ be a polynomial of degree at least one. Let $p \in R$ be prime. Then if*

1. *$p \nmid a_n$*
2. *$p \mid a_k$ for all $0 \leq k < n$*
3. *$p^2 \nmid a_0$*

$f(x)$ is irreducible.

Proof. By proposition 3.9.11 and Theorem 3.9.14, we may assume that f is primitive. Suppose $f(x)$ were reducible, say $f(x) = g(x)h(x)$. Let

$$f(x) = a_n x^n + \cdots + a_0 \qquad g(x) = b_m x^m + \cdots + b_0 \qquad h(x) = c_k x^k + \cdots + c_0$$

Then since $p^2 \nmid a_0 = b_0 c_0$, we may assume, without loss of generality, that $p \nmid c_0$ and $p \mid b_0$. Since $p \nmid a_n = b_m c_k$, we conclude that $p \nmid b_m$. We will now show, by induction, that $p \mid b_k$ for all $0 \leq k \leq m$, a contradiction. The case $k = 0$ is done. Suppose it holds for some $k < m$, and every number before that. Then

$$a_{k+1} = b_{k+1}c_0 + b_k c_1 + \cdots + b_0 c_{k+1}$$

where we allow for $c_j = 0$ if necessary. Note that since $f(x)$ is primitive, we may assume that $g(x), c(x)$ are both of degree at least one. Hence, $m < n$, so $p \mid a_{k+1}$. Since $p \nmid c_0$, it follows then by induction that $p \mid b_{k+1}$, as claimed. □

Theorem 3.11.2 (Reduction Criteria)**.** *Let R, R' be integral domains, and $\varphi : R \to R'$ a homomorphism. Let F, F' be the fraction fields of R, R'. Let $f \in R[x]$ be such that $\varphi(f) \neq 0$ and $\deg(\varphi(f)) = \deg(f) \geq 1$. Then if $\varphi(f)$ is irreducible in $F'[x]$, f has no factorization into a product of two degree one or higher polynomials in $R[x]$.*

Proof. Suppose f is reducible in $R[x]$, say $f(x) = g(x)h(x)$, where $\deg(g), \deg(h) \geq 1$. Then $\varphi(f) = \varphi(g)\varphi(h)$, so since $\deg(\varphi(g)) \leq \deg(g)$, and similar with h, we conclude that since $\deg(\varphi(f)) = \deg(f)$, φ preserves the degrees of g, h. Hence, $\varphi(f)$ is reducible in $F'[x]$. □

Theorem 3.11.3 (Integral Root Test)**.** *Suppose R is a UFD and $F = FF(R)$. Let $f(x) = a_n x^n + \cdots + a_0$ be a polynomial in $R[x]$. Let $\alpha \in F$ be a root of f, and write $\alpha = c/d$, where $GCD(c, d) = 1$ (i.e. the GCD are only units). Then $c \mid a_0$ and $d \mid a_n$. Furthermore, if a_n is a unit in R, then $\alpha \in R$.*

[9] Most of the results can also be found in the exercises of [Jac09]

Proof. We get

$$0 = f(\alpha) = a_n(c/d)^n + \cdots + a_0$$

Multiplying both sides through by d^n gives

$$0 = a_n c^n + a_{n-1}c^{n-1}d + \cdots + a_0 d^n$$

Thus, it follows that $c, d \mid a_n c^n + a_0 d^n$. Since $\text{GCD}(c, d) = 1$, the desired result comes from this. □

3.12 Symmetric Polynomials

As we saw in previous sections, multivariable polynomials, while not PIDs, are still UFDs. It's worth asking then whether there's any structure to their factorizations. The answer turns out to be yes, but only for a certain class of multivariable polynomials called *symmetric* polynomials. This section follows similar ones in [Lan05] and [Jac09].

We begin by taking a small detour to talk about *algebraic independence*.

Definition 3.12.1. Let $R \subset S$ be a subring. A set of elements $a_1, \ldots, a_n \in S$ are called algebraically independent over R if there exists no non-zero $f \in R[x_1, \ldots, x_n]$ such that $f(a_1, \ldots, a_n) = 0$.

Note. We can extend this definition to infinite sets of elements by stating that it is algebraically independent if all finite subsets are algebraically independent.

The following result is fairly immediate from this definition, and is left to the reader.

Proposition 3.12.2. *Let $R \subset S$ be a subring, and choose any $a_1, \ldots, a_n \in S$. Then the evaluation homomorphism $\varphi : R[x_1, \ldots, x_n] \to R[a_1, \ldots, a_n]$ is an isomorphism if and only if $a_1, \ldots, a_n$ are algebraically independent.*

This will become relevant later to show a very interesting result. But for now, let's get back to the main topic at hand and work towards defining *symmetric polynomials*.

Definition 3.12.3. Let $f \in R[x_1, \ldots, x_n], \sigma \in S_n$. We define the action of σ on f, denoted $f(\sigma)$, by

$$\sigma(f(x_1, \ldots, x_n)) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

f is fixed by σ if $\sigma(f) = f$, and is symmetric if it is fixed by every permutation in S_n.

The set of all symmetric polynomials forms a subring of $R[x_1, \ldots, x_n]$, which we call $\text{Sym}_n(R)$. The main result we build towards is rather surprising, namely that

$$\text{Sym}_n(R) \cong R[x_1, \ldots, x_n]$$

For this, we'll of course need to know what x_k are mapping to. This role will be fulfilled by what we call the *elementary symmetric polynomials*.

Definition 3.12.4. Consider the polynomial $F \in R[x_1, \ldots, x_n][X]$ given by

$$F(X) = (X - x_1)(X - x_2) \cdots (X - x_n)$$

We expand this out, getting an expression of the form

$$F(X) = X^n - s_1 X^{n-1} + \cdots + (-1)^n s_n$$

Then $s_k \in R[x_1, \ldots, x_n]$ are symmetric, and we call s_k the kth elementary symmetric polynomial.

Note. The factors of ± 1 in this definition are arbitrary, and just done to make the expressions in the following proposition a bit nicer. It's also clear from this definition that each x_k is algebraic over $R[s_1, \ldots, s_n]$.

One may ask what these s_k actually look like. This, it turns out, is easy enough to answer.

Proposition 3.12.5. *Let $s_k \in Sym_n(R)$, and let Γ_k be the set of all choices of k numbers from $\{1, \ldots, n\}$. Then*

$$s_k = \sum_{\{a_1, \ldots, a_k\} \in \Gamma_k} x_{a_1} \cdots x_{a_k}$$

Proof. By looking at $F(X)$, we can see that each term in s_k comes from multiplying k different $(-1)x_j$ together. Thus,

$$(-1)^j s_k = \sum_{\{a_1, \ldots, a_k\} \in \Gamma_k} (-1)^j x_{a_1} \cdots x_{a_k}$$

□

We need two more concepts before stating our main result, *homogeneity* and *weight*.

Definition 3.12.6. Let $x_1^{k_1} \cdots x_n^{k_n}$ be a monomial in $R[x_1, \ldots, x_n]$. We define the total degree of this term to be $k_1 + \cdots + k_n$, and the weight to be $k_1 + 2k_2 + \cdots + nk_n$. A polynomial $f \in R[x_1, \ldots, x_n]$ is called homogeneous if all of its terms have the same total degree, and its total degree $t(f)$ and weight $w(f)$ are the maximum total/weight degree of all of its terms.

It's clear from proposition 3.12.5 that the elementary symmetric polynomials are homogeneous, and have total degree k. We can also now, finally, state our main theorems.

Theorem 3.12.7. *Let $f \in R[x_1, \ldots, x_n]$ be symmetric polynomial such that $t(f) = d$. Then there exists $g \in R[x_1, \ldots, x_n]$ such that $w(g) \leq d$ and*

$$f(x_1, \ldots, x_n) = g(s_1, \ldots, s_n)$$

Furthermore, if f is homogeneous, then every monomial in g has weight d.

Proof. We start with induction on n. The result is clear for $n = 1$, since then $s_1 = x$ and $\mathrm{Sym}_n(R) = R[x]$. Now, suppose that the result holds for symmetric polynomials in $n-1$ variables, where $n \geq 2$. We define $s_k^{(0)}$ to be the kth elementary symmetric polynomial in $R[x_1, \ldots, x_n]$ with x_n evaluated to zero. Note that $s_n^{(0)} = 0$, so

$$X(X - x_1) \cdots (X - x_{n-1}) = X(X^{n-1} - s_1^{(0)} X^{n-2} + \cdots + (-1)^{n-1} s_{n-1}^{(0)})$$

Thus, $s_k^{(0)}$ is the kth elementary symmetric polynomial in $R[x_1, \ldots, x_{n-1}]$ for $1 \leq k \leq n-1$. Now, we proceed by induction on d. If $d = 0$, then the result is clear. Suppose the result holds for all symmetric polynomials in $n-1$ variables with total degree $< d$, where $d \geq 1$. Let $f \in \mathrm{Sym}_n(f)$ be such that $t(f) = d$. By the induction on n, there exists some polynomial $g \in R[x_1, \ldots, x_{n-1}]$ of weight $\leq d$ such that

$$f(x_1, \ldots, x_{n-1}, 0) = g(s_1^{(0)}, \ldots, s_{n-1}^{(0)})$$

Note that $t(g) \leq d$ in $R[x_1, \ldots, x_n]$. Thus, we conclude that

$$f_1(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) - g(s_1, \ldots, s_{n-1})$$

is a symmetric polynomial such that $t(f_1) \leq d$. Since $f_1(x_1, \ldots, x_{n-1}, 0) = 0$, we conclude that $x_n \mid f_1$. But f_1 is symmetric, and therefore $x_k \mid f_1$ for all $1 \leq k \leq n$. Therefore, there exists $f_2 \in \mathrm{Sym}_n(R)$ such that

$$f_1(x_1, \ldots, x_n) = x_1 \cdots x_n f_2(x_1, \cdots, x_n)$$

Since $t(f_1) \leq d$, $t(f_2) \leq d - n < d$. Thus, there exists by the inductive hypothesis some $h \in R[x_1, \ldots, x_n]$ such that $w(h) \leq w(f_2)$ and

$$f_2(x_1, \ldots, x_n) = h(s_1, \ldots, s_n)$$

Plugging this back into the above equations we get

$$f(x_1, \ldots, x_n) = x_1 \cdots x_n h(s_1, \ldots, s_n) - g(s_1, \ldots, s_{n-1}) = s_n h(s_1, \ldots, s_n) - g(s_1, \ldots, s_{n-1})$$

Calling $s_n h(s_1, \ldots, s_n) - g(s_1, \ldots, s_{n-1}) = r(s_1, \ldots, s_n)$, we see that $w(r) \leq d$, as required. For the second part of this theorem, we do the same induction. It clearly holds in the base cases, so suppose $f \in \mathrm{Sym}_n(f)$ such that $t(f) = d$ is homogeneous. By the induction on n, there exists some polynomial $g \in R[x_1, \ldots, x_{n-1}]$ with every monomial of weight d such that

$$f(x_1, \ldots, x_{n-1}, 0) = g(s_1^{(0)}, \ldots, s_{n-1}^{(0)})$$

Note that $t(g) = d$ in $R[x_1, \ldots, x_n]$. Thus, we conclude that

$$f_1(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) - g(s_1, \ldots, s_{n-1})$$

is either zero (in which case we're done) or a symmetric polynomial such that $t(f_1) = d$. Since $f_1(x_1, \ldots, x_{n-1}, 0) = 0$, we conclude that $x_n \mid f_1$. But f_1 is symmetric, and therefore $x_k \mid f_1$ for all $1 \leq k \leq n$. Therefore, there exists a homogeneous $f_2 \in \mathrm{Sym}_n(R)$ such that

$$f_1(x_1, \ldots, x_n) = x_1 \cdots x_n f_2(x_1, \cdots, x_n)$$

Since $t(f_1) = d$, $t(f_2) = d - n$. Thus, there exists by the inductive hypothesis some $h \in R[x_1, \ldots, x_n]$ such that every monomial in h has weight $d - n$ and

$$f_2(x_1, \ldots, x_n) = h(s_1, \ldots, s_n)$$

Plugging this back into the above equations we get

$$f(x_1, \ldots, x_n) = x_1 \cdots x_n h(s_1, \ldots, s_n) - g(s_1, \ldots, s_{n-1}) = s_n h(s_1, \ldots, s_n) - g(s_1, \ldots, s_{n-1})$$

Calling $s_n h(s_1, \ldots, s_n) - g(s_1, \ldots, s_{n-1}) = r(s_1, \ldots, s_n)$, we see that the weight of every monomial in r is d, as required. □

Theorem 3.12.8. *The elementary symmetric polynomials are algebraically independent over R.*

Proof. The result is clear for the case $n = 1$, so we proceed by induction. Let $n \geq 2$, suppose the result holds for $< n$, and that there existed some $f \in R[x_1, \ldots, x_n]$ such that $f(s_1, \ldots, s_n) = 0$. In particular, choose f to have a minimal (non-zero) total degree, call it $t(f) = d$. Write

$$f(x_1, \ldots, x_n) = f_0(x_1, \ldots, x_{n-1}) + \cdots + f_m(x_1, \ldots, x_{n-1}) x_n^m$$

where $f_k \in R[x_1, \ldots, x_n]$ and $f_m \neq 0$. We can note that, using the notation of the proof of Theorem 3.12.7

$$0 = f(s_1^{(0)}, \ldots, s_{n-1}^{(0)}, 0) = f_0(s_1^{(0)}, \ldots, s_{n-1}^{(0)})$$

By the inductive hypothesis, $s_1^{(0)}, \ldots, s_{n-1}^{(0)}$ are algebraically independent over $R[x_1, \ldots, x_{n-1}]$, so $f_0 = 0$. But then if $f \neq 0$

$$f(x_1, \ldots, x_n) = x_n(f_1(x_1, \ldots, x_{n-1}) + \cdots + f_m(x_1, \ldots, x_{n-1}) x_n^{m-1})$$

so $f_1(x_1, \ldots, x_{n-1}) + \cdots + f_m(x_1, \ldots, x_{n-1}) x_n^{m-1}$ is a polynomial of lower total degree which evaluates to zero on $s_1, \ldots, s_n$, a contradiction. Hence, $f = 0$ as required. □

This of course gives the following immediate result.

Corollary 3.12.8.1. $R[s_1, \ldots, s_n] \cong R[x_1, \ldots, x_n]$.

3.13 Complex Numbers and Quaternions*

The topics in this section will be pretty much completely irrelevant for the rest of the text, the author just gave a lecture on them once that they liked and doesn't want that work to go to waste. The material contained here is pulled from a combination of [Jac09], [Gub21], and the author's own head. It assumes basic knowledge of the complex numbers.

The real numbers, from an analytic perspective, are wonderful. They are however, from an algebraic perspective, terrible. Why? Well, in algebra we're usually in the business of solving for the roots of polynomials. And it turns out that a lot of very simple polynomials in $\mathbb{R}[x]$

have no roots in $\mathbb{R}[x]$, and in fact are irreducible. To fix this, we're going try defining a new ring which adds solutions to polynomials to $\mathbb{R}$. Let's start by adding a solution to the simplest polynomial in $\mathbb{R}[x]$ with no roots, x^2+1. To do this, we define our new ring to be

$$C = \frac{\mathbb{R}[x]}{(x^2+1)}$$

What we've done here is sort of a sleight of hand. If we let $I = (x^2+1)$, then we can see that $x+I \in C$. Thus, we can evaluate $f(x) = x^2+1$ in C at $x+I$, which gives the following

$$f(x+I) = (x+I)^2 + (1+I) = (x^2+1) + I = 0$$

The root we've added to f is exactly $x+I$!. It's not too hard to see that $-x+I$ is also a (distinct) root of f in this new ring. In fact, it turns out that *all* the roots of *every* polynomial in $\mathbb{R}[x]$ are contained in C. The reason for this is the following very simple theorem.

Theorem 3.13.1. *$C \cong \mathbb{C}$.*

Proof. Note that since $x^2 = 1$ in C, any element in C has a unique representation of the form $a+bx$, where $a, b \in \mathbb{R}$. It is then a quick verification that $\varphi(a+bx) = a+bi$ is a well-defined isomorphism. □

One could choose to define $\mathbb{C}$ in this manner, as the field $\mathbb{R}[x]/(x^2+1)$. In an algebraic context, this is actually a quite intuitive way of defining $\mathbb{C}$. The statement above about finding roots for any polynomial also falls out of this result, as we have (from many different field) the following.

Theorem 3.13.2 (The Fundamental Theorem of Algebra)**.** *Every $f \in \mathbb{R}[x]$ has, including multiplicity[10], n roots in $\mathbb{C}$.*

Great, we've added all the solutions to polynomials we could ever want. But we don't stop here, because there's a new problem. As you would learn almost immediately in any class covering the complex numbers, $\mathbb{C}$ is just $\mathbb{R}^2$ endowed with a multiplication. So can we in turn endow $\mathbb{C}^2$, or equivalently $\mathbb{R}^4$, with a multiplication? The answer is yes, if we're willing to give up on that multiplication being commutative.

Definition 3.13.3. The *quaternions*, denoted $\mathbb{H}$, are $\mathbb{C}^2$ with the standard vector addition and a multiplication given by

$$(a,b)\cdot(c,d) = (ac - b^*d, da + bc^*)$$

There are many useful properties and representations of the quaternions. Proving them is mostly just very tedious symbol pushing, so we simply list them below.

Proposition 3.13.4. *$\mathbb{H}$ is a non-commutative division ring.*

Proposition 3.13.5. *The following spaces are all isomorphic.*

[10] This is defined in the usual manner from grade school.

1. $\mathbb{H}$
2. $\mathbb{R}[i,j,k]/I$, *where* $I = (i^2+1, j^2+1, k^2+1, ij+ji, k+ji, jk+kj, i+kj, ki+ik, j+ik)$
3. *The subset of* $\mathbb{C}^{2\times 2}$ *consisting of matrices of the form*

$$\begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix}$$

where $a, b \in \mathbb{C}$, *and* $*$ *is complex conjugation.*

Furthermore, in the standard form of these isomorphisms, the following elements are equivalent (where $a, b, c, d \in \mathbb{R}$*)*

1. $(a+bi, c+di)$
2. $a+bi+cj+dk$
3. $$\begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix}$$

The latter of these forms also gives us the inverse of any non-zero quaternion, using the formula for the inverse of a $2x2$ matrix.

Proposition 3.13.6. *If* $(a,b) \in \mathbb{H}$ *is non-zero, then* $(a,b)^{-1} = (|a|^2+|b|^2)^{-1}(a^*, -b)$.

We call this factor $|a|^2+|b|^2$ the *norm* of the quaternion, and denote it $N((a,b))$. One can note that this is the determinant of the matrix

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

Thus, N is a multiplicative homomorphism and $N(a+bi+cj+dk) = a^2+b^2+c^2+d^2$. We call $(a^*, -b)$ the *conjugation* of the quaternion, and denote it $(a,b)^*$. Thus, the above proposition could be compressed to saying that for $a \in \mathbb{H}$, $a^{-1} = \frac{a^*}{N(a)}$, just like the inverses in $\mathbb{C}$. One can also check that $aa^* = N(a)$, and that the complex conjugate is a field automorphism of $\mathbb{H}$.

All these results are well and good, but to justify doing all of this let's look at some useful applications of quaternions. All of classical physics can technically be formulated in terms of quaternions, it's just a bad way to do it. Except for exactly one area, **rotations**.

Proposition 3.13.7. *Every quaternion* $h \in \mathbb{H}$ *can be written uniquely in the form*

$$h = a(\cos(\theta) + \mathbf{n}\sin(\theta))$$

where $\theta, a \in \mathbb{R}$, $\mathbf{n} = n_1 i + n_2 j + n_3 k$, *and* $n_1^2 + n_2^2 + n_3^2 = 1$.

Note that $\mathbf{n}$ is essentially a unit vector in $\mathbb{R}^3$, and that a acts like a magnitude. Thus, numbers in $\mathbb{H}$ with real component 0, which we denote $\mathrm{Im}(\mathbb{H})$, encode $\mathbb{R}^3$. This connects with rotations in the following manner.

Proposition 3.13.8. *Let $q = \cos(\theta) + \mathbf{n}\sin(\theta) \in \mathbb{H}$, and let $\mathbf{h} \in Im(\mathbb{H})$. Then the map $\mathbf{h} \mapsto q\mathbf{h}q^*$ is the rotation of $\mathbf{h}$ by an angle 2θ about the axis collinear to $\mathbf{n}$ in $\mathbb{R}^3$.*

This gives us an efficient way to store and compute rotations! We can also, in a similar manner, use quaternions to compute the cross and dot product. In fact, looking at vectors in $\mathbb{R}^3$ as quaternions

$$\mathbf{u} \cdot \mathbf{v} = -\frac{\mathbf{uv} + \mathbf{vu}}{2} \qquad \mathbf{u} \times \mathbf{v} = \frac{\mathbf{uv} - \mathbf{vu}}{2}$$

Quaternions also have numerous uses in number theory, many of which are outlined in [Gub21].

There is, of course, one remaining question. Can we pull this trick again, and endow $\mathbb{H}^2$ with a multiplication? The answer is yes, but the result won't be a ring and will instead be something called a $\mathbb{R}$ (or $\mathbb{C}$)-algebra. Essentially, you can multiply numbers in $\mathbb{H}^2$, but that multiplication won't be associative. This trend continues, with each step up the ladder losing more and more nice properties that multiplication could have. If you'd like to learn more about this, see [Gub21].

3.14 Chinese Remainder Theorem*

The Chinese Remainder Theorem is one of the most fundamental theorems in ring theory, and yet fit nowhere anywhere else in this chapter. Nor is it used again for the remainder of this book, except perhaps in the final chapter. I put it here for lack of a better place, but despite its "optional" marking I would highly recommend going over this section. The content of this section is based on lectures given by Dr. Kalle Karu at UBC.

Let's start with some preliminaries.

Definition 3.14.1. Let R be a ring and $\{I_j\}_{j=1}^n$ a set of ideals of R. We define

1. $$\sum_{j=1}^{n} I_j = \Big\{ \sum_{j=1}^{n} f_j \mid f_j \in I_j \Big\}$$
2. $$\prod_{j=1}^{n} I_j = \Big\{ \sum_{k=1}^{m} \prod_{j=1}^{n} f_{j,k} \mid f_{j,k} \in I_j, m \in \mathbb{Z}^+ \Big\}$$

These are called the sum and product ideals respectively.

It is not too hard to show that, like the names suggest, the sum and product ideals are ideals in R (the latter is only guaranteed to be an ideal when R is commutative). It is also clear that $\prod_{j=1}^{n} I_j \subset \bigcap_{j=1}^{n} I_j$. The Chinese Remainder theorem first tells us sufficient conditions for these two expressions to be equal, namely the condition of ideals being coprime.

Definition 3.14.2. Two ideals $I, J \subset R$ are coprime if $I + J = R$.

Theorem 3.14.3 (Chinese Remainder Theorem (CRT) I)**.** *Let $I_1, \dots, I_n \subset R$ be a collection of pairwise coprime ideals in a commutative ring. Then*

$$\prod_{j=1}^{n} I_j = \bigcap_{j=1}^{n} I_j$$

Proof. It suffices to show that $\prod_{j=1}^{n} I_j \supset \bigcap_{j=1}^{n} I_j$. First, suppose that $n = 2$. Pick any $x \in I_1 \cap I_2$. Since $I_1 + I_2 = R$, there exist some $y_1, y_2 \in I_1, I_2$ such that $y_1 + y_2 = 1$. Therefore,

$$x = x(y_1 + y_2) = y_1 x + x y_2 \in I_1 I_2$$

as required. Now, suppose $n \geq 2$. Set $\prod_{j=1}^{n-1} I_j = I$. By the case $n = 2$, it suffices to show that I, I_n are coprime. Since I_k, I_n are pairwise coprime for each $1 \leq k \leq n-1$, we can find $x_k \in I_k, y_k \in I_n$ such that $x_k + y_k = 1$. Then $\prod_{k=1}^{n-1} x_k \in I$,

$$\sum_{k=1}^{n-1} y_k \prod_{j=k+1}^{n-1} x_j \in I_n$$

where we say that $\prod_{j=k+1}^{n-1} x_j = 1$ when $k = n-1$, and one can verify

$$\prod_{k=1}^{n-1} x_k + \sum_{k=1}^{n-1} y_k \prod_{j=k+1}^{n-1} x_j = 1$$

Thus, $I + I_n = (1) = R$ as claimed. □

Corollary 3.14.3.1 (Chinese Remainder Theorem II)**.** *Endow the cartesian product of rings with a ring structure via element-wise operations. Let $I_1, \dots, I_n \subset R$ be ideals in an arbitrary ring. Let $q : R \to R/I_1 \times \cdots \times R/I_n$ be the ring homomorphism induced by the quotient maps, that is the one given by*

$$q(x) = (x \bmod I_1, \dots, x \bmod I_n)$$

Then q is surjective if and only if the I_k are pairwise coprime.

Proof. Suppose q is surjective. Pick any $1 \leq k \leq n$, and $j \neq k$. Then there exists some $x \in R$ such that

$$q(x) = (\dots, 1, 0, 0, \dots)$$

where the 1 is in the kth position. Then $x \in I_j$ and $\exists y \in I_k$ such that $x = y + 1$, so $x - y = 1 \Rightarrow I_j + I_k = (1) = R$. Thus, all the ideals are pairwise coprime.

Now, suppose that all the ideals are pairwise coprime. Set $I = \prod_{j=1}^{n-1} I_j$. By the proof of CRT I, we can find $x \in I, y \in I_n$ such that $x + y = 1$. Then $q(x) = (0, 0, \dots, 1)$, and it follows by a symmetric argument to this that q is surjective. □

Corollary 3.14.3.2 (Chinese Remainder Theorem III)**.** *Let $I_1, \ldots, I_n \subset R$ be coprime ideals in an arbitrary ring. Then*

$$\frac{R}{\bigcap_{i=1}^n I_i} = R/I_1 \times \cdots \times R/I_n$$

Also, if R is commutative then

$$\frac{R}{\prod_{i=1}^n I_i} = R/I_1 \times \cdots \times R/I_n$$

Proof. Let $q : R \to R/I_1 \times \cdots \times R/I_n$ be the ring homomorphism induced by the quotient maps, as defined in CRT II. By CRT I/II, it suffices to show that $\ker(q) = \bigcap_{i=1}^n I_i$. But this is immediate. □

For an interesting example of CRT in action, I suggest you look into the history of the theorem, particularly its original form in terms of $\mathbb{Z}$.

Part II

Linear Algebra

Chapter 4

Modules

4.1 Basics Definitions

Modules, in a broad sense, are simply generalizations of vector spaces to be over arbitrary rings rather than fields. We begin their study here, following (loosely) similar explanations from [Jac09] and [Lan05].

Definition 4.1.1. Let R be a ring. A left R-module M is an Abelian group $(M, +, 0)$ together with a scalar multiplication operation $\cdot : R \times M \to M$ satisfying the following axioms for all $x, y \in R, \underline{v}, \underline{u} \in M$

1. $x \cdot (\underline{v} + \underline{u}) = x \cdot \underline{v} + x \cdot \underline{u}$
2. $(x + y) \cdot \underline{v} = x \cdot \underline{v} + y \cdot \underline{v}$
3. $x \cdot (y \cdot \underline{v}) = (xy) \cdot \underline{v}$
4. $1 \cdot \underline{v} = \underline{v}$

We also have right R-modules, which are defined similarly.

Definition 4.1.2. Let R be a ring. A right R-module M is an Abelian group $(M, +, 0)$ together with a scalar multiplication operation $\cdot : M \times R \to M$ satisfying the following axioms for all $x, y \in R, \underline{v}, \underline{u} \in M$

1. $(\underline{v} + \underline{u}) \cdot x = \underline{v} \cdot x + \underline{u} \cdot x$
2. $\underline{v} \cdot (x + y) = \underline{v} \cdot x + \underline{v} \cdot y$
3. $(\underline{v} \cdot x) \cdot y = \underline{v} \cdot (xy)$
4. $\underline{v} \cdot 1 = \underline{v}$

A quick note about notation before we move on : like with everything else in algebra, we generally drop the $\cdot$ from our scalar multiplication expressions and just write $x\underline{v}$. We will

also, in this text, use the convention of underlining symbols which represent module elements. This is not standard, and most other texts will have no particular convention in this regard.

At first glance, left and right R-modules seem like the exact same thing, and if R is commutative they in fact are the same thing. But when R is not commutative we get complications. To understand why, we need to take a diversion into *endomorphisms of Abelian groups* and *anti-morphisms*. Let's start by examining the structure of a left R-module M.

Proposition 4.1.3. *Suppose M is a left R-module. For each $x \in R$, let $\varphi_x : M \to M$ be the map given by $\varphi_x : \underline{v} \to x\underline{v}$. Then the map $f : x \mapsto \varphi_x$ is a ring homomorphism from R into* $\mathrm{End}(M)$*, the ring of endomorphisms of M as an Abelian group.*

Proof. That φ_x is an endomorphism of M is given by the first axiom in definition 4.1.1. To check that $f : x \mapsto \varphi_x$ is a ring homomorphism, we first need to give a ring structure to $\mathrm{End}(M)$. Pick any $\varphi, \psi \in \mathrm{End}(M)$ and $\underline{v} \in M$. It is not too hard to see that the operations

$$(\varphi + \psi)(\underline{v}) = \varphi(\underline{v}) + \psi(\underline{v}) (\varphi \cdot \psi)(\underline{v}) = (\varphi \circ \psi)(\underline{v})$$

put a ring structure on $\mathrm{End}(M)$. That f is a homomorphism is then guaranteed by axioms 2-4 of definition 4.1.1. □

Of course, the above result also tells us that any homomorphism from R to $\mathrm{End}(M)$ will give us a left R-module structure. Thus, we could have defined a left R-module M as a commutative group M with ring homomorphism $R \to \mathrm{End}(M)$. This is where right R-modules differ. The map $\underline{v} \mapsto \underline{v} \cdot x$ will still be an endomorphism of M. However, with the ring structure we've given to $\mathrm{End}(M)$, we would get that $f(xy) = f(y)f(x)$, not $f(x)f(y)$. This type of function is called an *anti-morphism*, and is actually quite similar to homomorphisms. In fact, an analogous result to proposition 4.1.3 holds for right R-modules using anti-morphisms. However, anti-morphisms are not homomorphisms (in general), so this shows that left and right R-modules are not necessarily the same. One may note that they are the same if R is commutative, hence the earlier statement that left and right R-modules are no different for commutative R.
For the rest of this chapter, we work with left R-modules. Keep in mind that analogous results will hold for right R-modules in almost all cases, if you wish to see these results explicitly see [Jac09]. With this in mind, let us give some basic definitions of module theory.

Definition 4.1.4. Let M be a left R-module. A submodule $N \subset M$ is a subset such that $RN \subset N$, where

$$RN = \left\{ \sum_{\alpha} r_\alpha \underline{v}_\alpha \mid r_\alpha \in R, \underline{v}_\alpha \in N \right\}$$

where all sums are finite. Put more simply, it's a subset of M which is itself an R-module using the structure of M as an R-module.

Like with ideals, we also have the following two constructions which are also subgroups.

Proposition 4.1.5. *Let $\{M_i\}_{i \in I}$ be a collection of submodules of a left R-module M. Then*

1. $\bigcap_{i \in I} M_i$

2.

$$\sum_{i\in I} M_i = \Big\{ \sum_{i,finite} \underline{v}_i \mid \underline{v}_i \in M_i \Big\}$$

are both sub-modules of M.

The former of these allows us to make the following definition.

Definition 4.1.6. Let $S \subset M$, where M is a left R-module. Then the submodule generated by S, denoted $\mathrm{Span}_R(S)$, is the intersection of all submodules of M containing S.

Again, like for subgroups and ideals, we have a simple way of explicitly writing out these generated sub-modules.

Proposition 4.1.7. *Let $S \subset M$, where M is a left R-module. Then*

$$\mathrm{Span}_R(S) = \Big\{ \sum_{i=1}^{n} r_i \underline{v}_i \mid r_i \in R, \underline{v}_i \in S, n \in \mathbb{Z}^+ \Big\}$$

Before moving on, let's list a pair of facts about modules that would be no fun to prove.

Proposition 4.1.8. *Suppose that M is a left R-module, and $N \subset M$ a submodule. Then*

1. *$0\underline{v} = \underline{0}$ and $(-1)\underline{v} = -\underline{v}$, for any $\underline{v} \in M$.*
2. *N is an additive subgroup of M.*

Next, we move to talking about module homomorphisms.

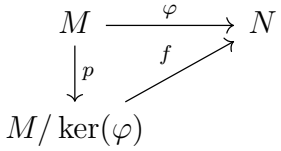
Definition 4.1.9. Let M, N be left R-modules. A module homomorphism $\varphi : M \to N$ is a homomorphism of additive groups satisfying, for any $x \in R, \underline{v} \in M$, $\varphi(r\underline{v}) = r\varphi(\underline{v})$. We denote the set of all module homomorphisms from M to N by $\mathrm{Hom}_R(M, N)$.

Note. $\mathrm{Hom}_R(M, N)$ is itself a left R-module. You will also here module homomorphisms be called R-linear maps.

The kernel of a module homomorphism is just the kernel of the underlying homomorphism of additive groups. It is easy to show that this kernel (and the image of a module homomorphism) is a submodule as well. One can also check that, if N is a submodule of M, then M/N is itself a left R-module, with operations inherited in the normal way from N.

At this point, I'm going to do something quite interesting. In the previous sections, I proved the fundamental theorems of homomorphisms. Here, I'm just going to state them.

Theorem 4.1.10 (First Fundamental Theorem of Module Homomorphisms)**.** *Let $\varphi \in \mathrm{Hom}_R(M, N)$ be a module homomorphism. Then the natural projection map $p : M \to M/\ker(\varphi)$ is a module homomorphism, and the map $f : M/\ker(\varphi) \to \mathrm{Im}(\varphi)$ given by $f : \underline{v} + \ker(\varphi) \to \varphi(\underline{v})$ is a well-defined module isomorphism. Finally, the following diagram commutes.*



Theorem 4.1.11 (Second Fundamental Theorem of Module Homomorphisms)**.** *Let* $\varphi \in \mathrm{Hom}_R(M, N)$ *be a surjective module homomorphism. Then*

1. *An additive subgroup* $S \subset M$ *containing* $\ker(\varphi)$ *is a submodule if and only if* $\varphi(S)$ *is a submodule.*
2. *The map* $S \mapsto \varphi(S)$ *of submodules of* M *containing* $\ker(\varphi)$ *is a bijection onto submodules of* N*.*
3. *If* $M' \subset M$ *is a submodule containing* $\ker(\varphi)$*, then* $M/M' \cong N/\varphi(M')$*.*

Corollary 4.1.11.1. *Suppose* $K \subset N$ *are both submodules of a left* R*-module* M*. Then*

$$M/N \cong \frac{M/K}{N/K}$$

Theorem 4.1.12 (Third Fundamental Theorem of Module Homomorphisms)**.** *Let* N, K *be submodules of a left* R*-module* N*. Then*

$$\frac{N}{N \cap K} \cong \frac{N + K}{K}$$

I don't bother with the proofs here for a good reason : these theorems are again essentially identical to those found in section 2.4 and section 3.4. At this point, you should be able to do them on your own, or at the very least believe them when you see them. If you do not feel that you've reached this point yet, then I would suggest reading those two sections again.

4.2 Free Modules and Bases

We take on now the vitally important task of generalizing bases from vector spaces to modules. To do so, we take a synthesis of similar sections in [Jac09] and [Lan05], along with some insights from course notes [Bad10] and linear algebra on vector spaces [Rom07].

Definition 4.2.1. Let M be a left R-module, and $S \subset M$. We call S

1. Linearly independent if
$$\sum_{\underline{v} \in S} a_{\underline{v}} \underline{v} = 0 \Rightarrow a_{\underline{v}} = 0, \forall \underline{v} \in S$$
2. Linearly dependent if it is not linearly independent.
3. A basis if it is linearly independent and $\mathrm{Span}_R(S) = M$.

These definitions are identical to those we use in vector spaces.

Note. In expressions such as

$$\sum_{\underline{v}\in S} a_{\underline{v}}\underline{v} = 0 \Rightarrow a_{\underline{v}} = 0, \forall \underline{v} \in S$$

we always assume that only finitely many terms have a non-zero $a_{\underline{v}}$, and ignore those with zero. Indeed, the expression is not well-defined otherwise. This just gives us a compact way to represent all finite linear combinations of elements in S.

Many of the nice properties you're used to from vector space bases carry over to modules as well.

Proposition 4.2.2. *Let M be a free left R-module with basis $V = \{\underline{v}_i\}_{i\in I}$. Then each $\underline{u} \in M$ can be written in a unique way in the form*

$$\underline{u} = \sum_{i\in I} a_i\underline{v}_i$$

where $a_i \in R$.

Proof. The existence is simply the statement that V is spanning. For uniqueness, suppose that

$$\sum_{i\in I} a_i\underline{v}_i = \sum_{i\in I} v_i\underline{v}_i$$

Rearranging, we get

$$\sum_{i\in I}(a_i - b_i)\underline{v}_i = \underline{0}$$

which by linear independent implies that $a_i = b_i$ for every $i \in I$. □

Proposition 4.2.3. *Let M be a free left R-module with basis $V = \{\underline{v}_i\}_{i\in I}$. Let N be any other left R-module, and for each $i \in I$ choose some $\underline{u}_i \in N$. Then there exists a unique $\varphi : \mathrm{Hom}_R(M, N)$ such that $\varphi(\underline{v}_i) = \underline{u}_i$, for all $i \in I$.*

Proof. By the previous proposition, and $\underline{v} \in V$ can be written uniquely in the form

$$\underline{v} = \sum_{i\in I} a_i\underline{v}_i$$

Thus, by linearity the only way φ could be defined is

$$\varphi(\underline{v}) = \sum_{i\in I} a_i\underline{u}_i$$

This gives us that φ is unique and well-defined. Checking that it is a module homomorphism is simple, and left to the reader. □

Corollary 4.2.3.1. *If M, N are left R-modules with bases V, U such that $|V| = |U|$, then $M \cong N$.*

In light of that last result, it is natural to ask whether modules have a unique basis cardinality (or for finite cases, basis size). The answer, it turns out, is sufficiently strange that we'll spend the rest of this section answering it.

Let's start by simplifying our terminology. Let I be an arbitrary set, and R a ring. We denote $\prod_{i\in I} R$ by R^I. We can endow a left R-module structure on this in the following way.

$$(a_i)_{i\in I} + (b_i)_{i\in I} = (a_i + b_i)_{i\in I} \qquad\qquad c \cdot (a_i)_{i\in I} = (ca_i)_{i\in I}$$

We denote the sequence with a one in the ith position and zeroes everywhere else by $\underline{e}_i \in R^I$. It is not hard to check that $\mathcal{B} = \{\underline{e}_i\}_{i\in I}$ is a basis for R^I, we call this the *standard basis*. By corollary 4.2.3.1, any free left R-module is isomorphic to R^I for some set I. Thus, we can reduce our study of free left R-modules to just the study of those R^I.

In light of this, we can carry over more results we know from basic linear algebra. In particular, in "finite-dimensional" spaces we can represent linear maps as matrices (relative to a pair of basis chosen), with the process being identical to that done for vector spaces. As such, the following result probably shouldn't be too surprising.

Theorem 4.2.4. *Let $n, m \in \mathbb{N}$. Then $R^n \cong R^m$ if and only if there exists a pair of matrices $A \in M_{n,m}(R), B \in M_{m,n}(R)$ such that $AB = \mathrm{Id}_n, BA = \mathrm{Id}_m$.*

Proof. First, suppose that there exists an isomorphism $\varphi : R^n \to R^m$. Let $\{\underline{e}_i\}_{i=1}^n, \{\underline{x}_j\}_{j=1}^m$ be the standard bases for R^n, R^m respectively. Then since φ is an isomorphism, $\varphi(\{\underline{e}_i\}_{i=1}^n)$ is a basis for R^m (this is not hard to check). We'll write $\underline{y}_i = \varphi(\underline{e}_i)$. Then for each $1 \le i \le n$, there exist unique $a_{i,j} \in R$ such that

$$\underline{y}_i = \sum_{j=1}^{m} a_{i,j}\underline{x}_j$$

Similarly, for each $1 \le j \le m$ we can find unique $b_{j,i}$ such that

$$\underline{x}_j = \sum_{i=1}^{n} b_{j,i}\underline{y}_i$$

We write $A = (a_{i,j}), B = (b_{j,i})$. We'll show that these are the desired matrices. First, we get

$$(AB)_{i,j} = \sum_{k=1}^{m} a_{i,k}b_{k,j}$$

Now, since

$$\underline{y}_i = \sum_{j=1}^{m} a_{i,j}x_j = \sum_{j=1}^{m}\sum_{k=1}^{n} a_{i,j}b_{j,k}\underline{y}_i = \sum_{k=1}^{n}\Big(\sum_{j=1}^{m} a_{i,j}b_{j,k}\Big)\underline{y}_k$$

By the uniqueness of the $a_{i,j}$ in the original expression, it follows that

$$(AB)_{i,j} = \sum_{k=1}^{m} a_{i,k}b_{k,j} = \delta_{ij} \Rightarrow AB = \mathrm{Id}_n$$

We also get $BA = \mathrm{Id}_m$ by the same argument.

Now, suppose that we have such matrices A, B. Then B is an R-linear map from R^n to R^m, which the existence of A shows is invertible and hence an isomorphism. □

In the case of commutative rings R, this result gives what we'd expect due to the following lemma.

Lemma 4.2.5. *Suppose R is a commutative ring and $A, B \in M_n(R)$. Then $AB = \mathrm{Id}_n \Rightarrow BA = \mathrm{Id}_n$.*

Proof. This follows from the properties of the determinant. Indeed, we get

$$\det(BA) = \det(B)\det(A) = \det(A)\det(B) = \det(AB) = 1$$

so BA is invertible. Then we note

$$BA = B(\mathrm{Id}_n)A = B(AB)A = (BA)^2$$

Applying $(BA)^{-1}$ on both sides, which we now know exists, gives the desired result. □

Note. This result **does not hold** if R is not commutative.

Corollary 4.2.5.1. *Suppose $n, m \in \mathbb{N}$ and R is commutative. Then $R^n \cong R^m$ if and only if $n = m$.*

Proof. The direction assuming $n = m$ is obvious, so instead assume that $R^n \cong R^m$. Assume, without loss of generality, that $n < m$. Using the same notation as in the proof of Theorem 4.2.4, we instead define $A, B \in M_m(R)$ by

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,m} \\ a_{2,1} & \cdots & a_{2,m} \\ \vdots & \vdots & \vdots \\ a_{n,1} & \cdots & a_{m,m} \\ 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \qquad B = \begin{pmatrix} b_{1,1} & \cdots & a_{1,n} & 0 & \cdots & 0 \\ b_{2,1} & \cdots & b_{2,n} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{m,1} & \cdots & b_{m,n} & 0 & \cdots & 0 \end{pmatrix}$$

Then by the calculation in Theorem 4.2.4, we'd still get $BA = \mathrm{Id}_n$, but $AB \neq \mathrm{Id}_n$, violating lemma 4.2.5. □

The above note is quite important here, as there are non-commutative rings R where $R^n \cong R^m$ and $n \neq m$. I'll leave it to the interested reader to look up counterexamples, as they don't tend to be one-line things. In general, we call any ring satisfying the above corollary an *invariant basis number* (IBN) ring.

Perhaps the most remarkable fact out of all of this is that modules with infinite bases are more well-behaved than finite ones! Indeed, if we replace n, m with infinite sets, then the above corollary becomes true for any ring. More precisely, we have the following.

Theorem 4.2.6. *Suppose I, J are sets such that I is infinite. Then $R^I \cong R^J$ if and only if $|I| = |J|$.*

Proof. Again, it suffices to show that $R^I \cong R^J \Rightarrow |I| = |J|$. First, we'll show that J is infinite. Indeed, suppose that J were finite. Let $\varphi : R^I \to R^J$ be a module isomorphism, $\underline{x}_i$ be the images of the standard basis elements of R^I under φ, and $\underline{y}_j$ be the standard basis elements of R^J. Since φ is an isomorphism, $\{\underline{x}_i\}_{i\in I}$ is a basis of R^J. Thus, each $\underline{y}_j$ can be written as a linear combination of finitely many $\underline{x}_i$. But since J is finite, this would imply that finitely many $\underline{x}_i$ span R^J, and hence that $\{\underline{x}_i\}_{i\in I}$ is not linearly independent. This is impossible, and hence J must be infinite.

Now, we can assume that J is also infinite. For each $j \in J$, let $U_j \subset I$ be a finite subset such that $\underline{y}_j$ can be expressed as a linear combination of vectors in $\{\underline{x}_i\}_{i\in U_j}$. Then since $\{\underline{y}_j\}_{j\in J}$ is a basis, we must get

$$\bigcup_{j\in J} U_j = I$$

Therefore, $|J| \leq |\bigcup_{j\in J} U_j| \leq |I|$. A symmetric argument shows that $|I| \leq |J|$, completing the proof. □

Note. If you're not comfortable playing around with set cardinalities like this, I suggest looking at the preliminaries in [Rom07].

If a module has a basis (this is not guaranteed), then we call the cardinality of this basis its *rank*. Note that finite ranks, when they exist, may not be unique for non-IBN rings.

4.3 Direct Sums and Products

This section primarily follows the notes of [Sil23], with some ideas from [Lan05] and [Rom07]. We start by defining *direct sums*.

Definition 4.3.1. Let $\{M_i\}_{i\in I}$ be a collection of left R-modules. The direct sum of these modules, denoted $\bigoplus_{i\in I} M_i$, is the set of sequences in $\prod_{i\in I} M_i$ with finite support (i.e. finitely many non-zero values) with addition defined element-wise and multiplication by

$$r(\underline{m}_i)_{i\in I} = (r\underline{m}_i)_{i\in I}$$

These have a couple properties which may be verified without too much difficulty.

Proposition 4.3.2. *Let $\{M_i\}_{i\in I}$ be a collection of left R-modules. Then*

1. *Each M_i can be embedded in $M = \bigoplus_{i\in I} M_i$ via the canonical set embedding. That is, calling our inclusions $\iota_i : M_i \to M$, we define that $\iota_i(\underline{m})$ to be the sequence with zeroes everywhere except having $\underline{m}$ in the ith position.*
2. *Suppose $\{B_i\}_{i\in I}$ is a collection of bases for each M_i. Then $\bigcup_{i\in I} \iota_i(B_i)$ is a basis for M.*
3. *If $n, m \in \mathbb{N}$, then $R^n \oplus R^m \cong R^{n+m}$.*

4. $(M_1 \oplus M_2) \oplus M_3 \cong M_1 \oplus (M_2 \oplus M_3)$

There is a (sometimes) closely related notion that we've encountered, called the *internal sum.*

Definition 4.3.3. Let $\{N_i\}_{i\in I} \subset M$ be sub-modules of a left R-module M. Then we define the internal sum of these to be

$$\sum_{i\in I} N_i = \mathrm{Span}_R\left(\bigcup_{i\in I} N_i\right) = \{\textit{finite sums of elements from the } N_i\}$$

Let $\varphi : \bigoplus_{i\in I} N_i \to \sum_{i\in I} N_i$ be the homomorphism defined by being the inclusion on each N_i. This sum is called *direct* if φ is an isomorphism.

Ideally, we'd like all our sums to be direct, as breaking a module down into a direct sum of simpler modules makes it much easier to work with. This, of course, isn't the case, but we do have simple rules for detecting when internal sums are and are not direct.

Theorem 4.3.4. *Let $\{N_i\}_{i\in I} \subset M$ be sub-modules of a left R-module M. Then the following are equivalent.*

1. *$\sum_{i\in I} N_i$ is direct and $\sum_{i\in I} N_i = M$.*
2. *For each $i \in I$, $N_i \cap \left(\sum_{j\neq i} N_j\right) = \{0\}$, and $\sum_{i\in I} N_i = M$.*
3. *Every $\underline{m} \in M$ has a unique representation as a finite sum of elements, each from a different N_i.*

Proof. First, suppose that (1) holds, and let $\varphi : \bigoplus_{i\in I} N_i \to M$ be the standard homomorphism. Pick $i \in I$, and suppose there exists $\underline{m_i} \in N_i$ and $\alpha_j \in R, \underline{m_j} \in N_j$ such that

$$\underline{m_i} = \sum_{j\neq i} \alpha_j \underline{m_j}$$

Defining $-\alpha_i = 1$, this implies in turn that

$$\varphi((\alpha_j \underline{m_j})_{j\in I}) = \underline{0}$$

But φ is an isomorphism, so it follows that $\underline{m_i} = \underline{0}$ and (2) holds.

Now, suppose that (2) holds, and that there exists two representations of some $\underline{m} \in M$ in the form of (3), say

$$\underline{m} = \sum_{i\in I} \underline{m_i} = \sum_{i\in I} \underline{x_i}$$

where $\underline{m_i}, \underline{x_i} \in N_i$. Without loss of generality, pick out some $i \in I$ such that $\underline{m_i} \neq \underline{x_i}$. Then we get

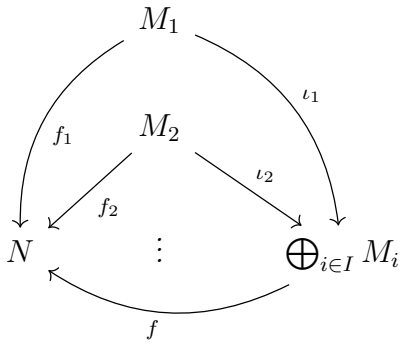
$$\underline{m_i} - \underline{x_i} \in N_i \backslash \{\underline{0}\}, \underline{m_i} - \underline{x_i} = \sum_{j\neq i} (\underline{x_j} - \underline{m_j}) \in \sum_{j\neq i} N_j$$

violating (2). Thus, (3) must hold.

Finally, suppose that (3) holds. That $\sum_{i\in I} N_i = M$ is immediate from this, so let $\varphi : \bigoplus_{i\in I} N_i \to M$ be the standard homomorphism. If $\ker(\varphi)$ is non-trivial, then $\underline{0}$ has two distinct representations in the form of (3). Thus, $\ker(\varphi)$ is trivial, making φ and isomorphism and (1) satisfied. □

The direct sum can also be characterized by the following universal property.

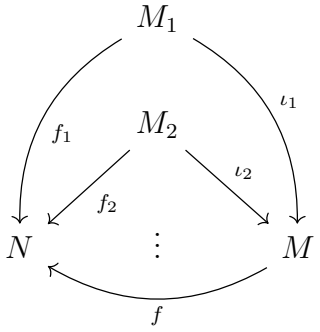
Theorem 4.3.5. *Let $\{M_i\}_{i\in I}$ be a collection of left R-modules with standard inclusions $\iota_i \in \mathrm{Hom}_R(M_i, \bigoplus_{i\in I} M_i)$, N some other R-module and $f_i \in \mathrm{Hom}_R(M_i, N)$ be linear maps. Then there exists a unique $f \in \mathrm{Hom}_R(\bigoplus_{i\in I} M_i, N)$ such that $f \circ \iota_i = f_i$. That is, the following diagram commutes.*



Proof. The condition $f \circ \iota_i = f_i$ forces the value of f on each $\iota_i(M_i)$, so since $\sum_{i\in I} \iota_i(M_i) = \bigoplus_{i\in M}$ this homomorphism is unique if it is well-define. To check that it is well-defined by the conditions $f \circ \iota_i = f_i$, we just need to check that it is single-valued, which is given by condition (2) in Theorem 4.3.4. □

In fact, we could have used this to define a direct sum. Indeed, starting at that point.

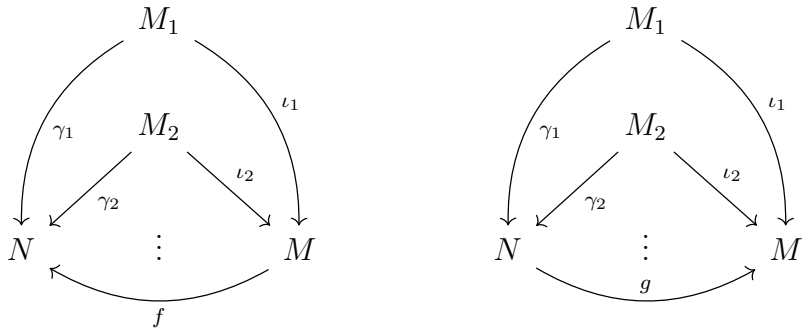
Definition 4.3.6. Let $\{M_i\}_{i\in I}$ be a collection of left R-modules. A direct sum of these modules is a left R-module M and collection of injective homomorphisms $\iota_i \in \mathrm{Hom}_R(M_i, M)$ satisfying the following property : If N is some other R-module and $f_i \in \mathrm{Hom}_R(M_i, N)$ linear maps, then there exists a unique $f \in \mathrm{Hom}_R(M, N)$ such that $f \circ \iota_i = f_i$. That is, the following diagram commutes.



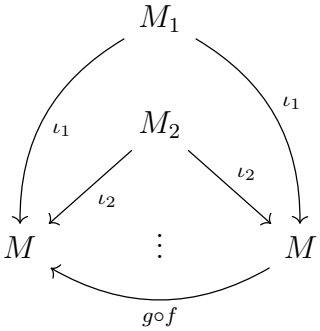
We can derive the following result.

Proposition 4.3.7. *Any pair of direct sums of a collection $\{M_i\}_{i\in I}$ of direct modules have a unique isomorphism between them satisfying the defining universal property.*

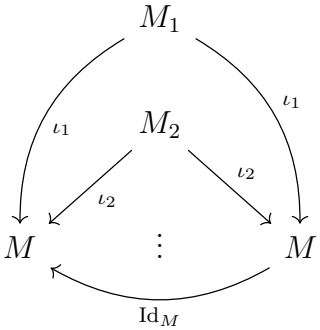
Proof. Suppose $(\{\iota_i\}_{i\in I}, M)$ and $(\{\gamma_i\}_{i\in I}, N)$ are two direct sums. Then there exists a unique homomorphism $f \in \mathrm{Hom}_R(M, N)$ and a unique homomorphism $g \in \mathrm{Hom}_R(N, M)$ such that the following two diagrams commute.



It thus follows that $(g \circ f) \circ \iota_i = g \circ (f \circ \iota_i) = g \circ \gamma_i = \iota_i$. Thus, the following diagram commutes.



But of course, the following diagram also commutes

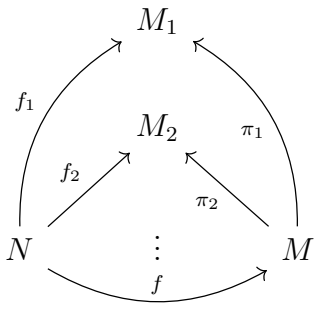


so by the uniqueness property of the direct sum we conclude that $g \circ f = \mathrm{Id}_M$. A similar argument shows that $f \circ g = \mathrm{Id}_N$, so f is the desired isomorphism. By the uniqueness property of the direct sum, f is the only such isomorphism. □

Essentially what we're saying here is that given any pair of direct sums, we can find a unique *compatible* isomorphism between them. Hence, there is in a sense only one direct sum construction.

We'll next do our first example of what's called *dualizing*. The premise is quite simple, what would happen if we reversed all the arrows in definition 4.3.6?

Definition 4.3.8. Let $\{M_i\}_{i\in I}$ be a collection of left R-modules. A direct product of these modules is a left R-module M and collection of surjective homomorphisms $\pi_i \in \mathrm{Hom}_R(M, M_i)$ satisfying the following property : If N is some other R-module and $f_i \in \mathrm{Hom}_R(N, M_i)$ linear maps, then there exists a unique $f \in \mathrm{Hom}_R(N, M)$ such that $\pi_i \circ f = f_i$. That is, the following diagram commutes.



The first thing to check is that something actually satisfies this definition. What turns out to work is essentially duplicating the original construction of the direct sum, but allowing all sequences instead of just those with finitely many non-zero elements. It will be left to the reader to check that this is the desired direct product, and that direct products, like direct sums, are unique up to a unique compatible isomorphism. In note of this, we denote the direct sum (rather confusingly) by $\prod_{i\in I} M_i$.

There is something very important to notice here. Our definitions of the direct product and sum depend only on *the properties of homomorphisms*. So there's nothing to stop us from taking these definitions and porting them over to rings or groups. Indeed, the direct products and sums that have shown up in previous chapters could be defined by the same universal property! We'll look into this more in Part III.

Note. In the case of left R-modules, *finite* direct sums and products are the same, but *infinite* ones are different. This does not necessarily carry over to other algebraic objects.

4.4 Free Modules over PIDs

From now on, we assume all our rings are PIDs and hence commutative. There will no longer be a distinction between left and right modules as a result, and we simply call both modules.
This, and the following three sections, are re-worked versions of similar topics as presented in [Jac09]. That being said, I hope that you'll find my explanations in this section much cleaner than his, which are often quite poor.

Our main goal for the last three sections of this chapter will be to derive the *Structure Theorem* for finitely generated modules over a PID. It makes sense then to consider free modules, both as a simple case and because they in fact relate to all finitely generated

modules. Indeed, suppose we have a module M over a PID R with generators $\underline{v}_1, \ldots, vv_n$. Then we get a homomorphism $\varphi : R^n \to M$ given by $\varphi(\underline{e}_i) = \underline{v}_i$, and $M \cong R^n/\ker(\varphi)$.

The above observations give us a blueprint for understanding the structure of modules : first understand free module, then the sub-modules $\ker(\varphi)$, and finally their quotient. Of course, we already understand free modules over commutative rings quite well, so we can move straight to step 2. This is completed via the following fundamental result.

Theorem 4.4.1. *Let R be a PID, and $n \in \mathbb{Z}_+$. Then any sub-module $M \subset R^n$ is free of rank $m \leq n$, where we define $R^0 = 0$.*

Proof. By induction on n. The case $n = 0$ is trivial. Now, suppose $n = 1$. Then any submodule of R^1 is an ideal in R, and hence since R is a PID if $N \subset R^1$ is a submodule then there exists some $\underline{v} \in R$ such that $N = (\underline{v})$. Now, since R is a domain, $a\underline{v} = 0 \Rightarrow a = 0$ or $\underline{v} = \underline{0}$. Thus, either $\underline{v} = 0 \Rightarrow N = 0$ or $N \cong R^1$, either way giving the desired result.

Now, consider some $n \geq 2$ and suppose the result holds for $n - 1$. Denote by R^{n-1} the submodule $(\underline{e}_2, \ldots, \underline{e}_n) \subset R^n$. Then R^{n-1} is free of rank $n - 1$, and R^n/R^{n-1} is free of rank one. Let $N \subset R^n$ be any sub-module, and let $q : R^n \to R^n/R^{n-1}$ be the standard quotient homomorphism. If $N \subset R^{n-1}$ or is generated by a multiple of $\underline{e}_1$, then it is free of rank $\leq n - 1$ by induction. Otherwise, note that by induction

1. $N \cap R^{n-1}$ is free of rank $\leq n - 1$.

2. $q(N)$ is free of rank 1.

Furthermore, since R is a domain, any $\underline{v} \in R^n$ can be written uniquely in the form $\sum_{i=1}^n a_i \underline{e}_i$, where $a_i \in R$ (this is easy to check). By our second observation and the linear independence of $\underline{e}_1, \ldots, \underline{e}_n$, there exists some $x \in R$ such that for every $\underline{v} \in N$, $a_1 = bx$ for some $b \in R$, where $b \in q(N)$. By our first observation, there exists a basis $\underline{u}_2, \ldots, \underline{u}_r$ of $N \cap R^{n-1}$, where $0 < r \leq n - 1$. Let $\underline{y} \in N \cap R^{n-1}$ be an arbitrary non-zero element. We'll show that $x\underline{e}_1 + \underline{y}, \underline{u}_2, \ldots, \underline{u}_r$ is a basis for N in R^n, hence making N free of rank $\leq n$ and completing the proof by corollary 4.2.3.1. That this set is linearly independent is immediate from the linear independence of $\underline{u}_2, \ldots, \underline{u}_r$ and of the standard basis. For spanning, note that there exists some $x \in R$ such that $a_1 = bx$. Then $bx\underline{e}_1 + \underline{y} \in N$, and hence there exists some $\underline{u} \in N$ such that $bx\underline{e}_1 + b\underline{y} + \underline{u} = \underline{v}$. But of course the coefficients of $bx\underline{e}_1 + b\underline{y}$ and $\underline{v}$ on $\underline{e}_1$ are equal, so by the uniqueness of that representation it follows that $\underline{u} \in N \cap R^{n-1}$. Hence, $\underline{u}$ is in the span of $\underline{u}_2, \ldots, \underline{u}_n$, completing the proof. □

This is quite a powerful result. Indeed, if we can find a nice way of relating the bases of R^n and any given submodule of R^n, then by our above observation that $M \cong R^n/\ker(\varphi)$ it should give us a thorough understanding of the structure of M. We'll therefore spend the rest of this section giving a quick overview of changing bases/generators in modules, followed by finding an algorithm to get simply-related bases in the next section.

Let $N \subset R^n$ be a submodule. Then by the above theorem N is finitely generated, say with generators $\underline{u}_1, \ldots, \underline{u}_m$, where potentially $m > n$. Writing out these generators in the unique

form

$$\begin{aligned}
\underline{u}_1 &= a_{11}\underline{e}_1 + a_{12}\underline{e}_2 + \cdots + a_{1n}\underline{e}_n \\
\underline{u}_2 &= a_{21}\underline{e}_1 + a_{22}\underline{e}_2 + \cdots + a_{2n}\underline{e}_n \\
\vdots &\qquad\qquad \vdots \\
\underline{u}_m &= a_{m1}\underline{e}_1 + a_{m2}\underline{e}_2 + \cdots + a_{mn}\underline{e}_n
\end{aligned}$$

where $a_{ij} \in R$ gives a matrix $A = (a_{ij}) \in M_{m\times n}(R)$, which we call the *relations matrix* of this basis and set of generators (in this particular order).

Now, suppose that $\underline{e}'_1, \ldots, \underline{e}'_n$ is another basis for R^n (since R is commutative the basis size of R^n is fixed), and $\underline{u}'_1, \ldots, \underline{u}'_r$ another set of generators for N. Let B be the companion matrix of this pair. We wish to relate A and B via matrix multiplication. To start off, we define the matrix $P = (p_{ij}) \in M_n(R)$ via the unique coefficients

$$\underline{e}'_i = \sum_{j=1}^{n} p_{ij}\underline{e}_j$$

Similarly, we define the matrix $Q = (q_{ij}) \in M_{r,m}(R)$ by the (possibly non-unique) coefficients

$$\underline{u}'_i = \sum_{j=1}^{r} q_{ij}\underline{u}_j$$

There are a few things to show here. First, we claim that P is invertible. Indeed, defining the matrix $L = (\ell_{ij}) \in M_n(R)$ via the unique coefficients

$$\underline{e}_i = \sum_{j=1}^{n} \ell_{ij}\underline{e}'_j$$

we get

$$\underline{e}_i = \sum_{j=1}^{n} \ell_{ij} \sum_{k=1}^{n} p_{jk}\underline{e}_k = \sum_{k=1}^{n} \Big(\sum_{j=1}^{n} \ell_{ij}p_{jk} \Big)\underline{e}_i$$

which implies that

$$\sum_{k=1}^{n} \Big(\sum_{j=1}^{n} \ell_{ij}p_{jk} \Big) = \delta_{ij}$$

Hence, $LP = \mathrm{Id}_n$ and since R is commutative $PL = \mathrm{Id}_n$, making P invertible. Next, we claim that $B = QAP^{-1}$, giving us our desired relation between companion matrices. Indeed,

one gets

$$\begin{aligned}\underline{u}'_i &= \sum_{j=1}^{m} q_{ij}\underline{u}_j = \sum_{j=1}^{m} q_{ij}\sum_{k=1}^{n} a_{jk}\underline{e}_k = \sum_{j=1}^{m} q_{ij}\sum_{k=1}^{n} a_{jk}\sum_{f=1}^{n} \ell_{kf}\underline{e}'_f \\ &= \sum_{f=1}^{n}\Big(\sum_{k=1}^{n}\sum_{j=1}^{m} q_{ij}a_{jk}\Big)\ell_{kf}\underline{e}'_f = \sum_{f=1}^{n}\Big(\sum_{k=1}^{n}(QA)_{ik}\Big)\ell_{kf}\underline{e}'_f \\ &= \sum_{f=1}^{n}\Big(\sum_{k=1}^{n}(QA)_{ik}\ell_{kf}\Big)\underline{e}'_f = \sum_{f=1}^{n}(QAL)_{if}\underline{e}'_f\end{aligned}$$

Therefore, $(QAL)_{if} = b_{if}$, so $QAP^{-1} = B$ as claimed. If we happen to have chosen bases for N, then Q will be invertible as well. This leads to A and B being what we call *similar*. More specifically, two matrices $A, B \in M_{m,n}(R)$ are similar (specifically that A is similar to B, denoted $A \sim B$) if there exists invertible matrices $Q \in M_m(R), P \in M_n(R)$ such that $B = QAP^{-1}$.

Note. Similarity is an equivalence relation.

4.5 Matrices Over PIDs

Let's cut straight to the chase, the *normal form* of matrices over PID.

Theorem 4.5.1. *If $A \in M_{m,n}(R)$, where R is a PID, then there exists a matrix $D \in M_{m,n}(R)$ similar to A of the form*

$$D = \begin{pmatrix} d_1 & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & d_2 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & d_r & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

where $d_i \neq 0$ and $d_i \mid d_j$ if $i \leq j$.

The proof of this is extremely tedious, so we'll separate all the setup here into a separate section. We'll start by defining E_{ij} to be the matrix in $M_{n,n}(R)$ with zeroes everywhere except for a 1 in position i, j. We define our four elementary matrices, for any $a, b, c, d \in R$, in the following manner.

1. $T_{ij}(a) = \mathrm{Id}_n + aE_{ij}$
2. $D_i(a) = \mathrm{Id}_n + (a-1)E_{ii}$
3. $P_{ij} = \mathrm{Id}_n - E_{ii} - E_{jj} + E_{ij} + E_{ji}$

4. $\chi(a,b,c,d) = \mathrm{Id}_n + (a-1)_E 11 + bE_{12} + cE_{21} + (d-1)E_{22}$

Note that the last of these is only well-defined when $n \geq 2$. The first three you may have seen before in a class on basic linear algebra, and have the following effects.

1. Any $X \in M_{m\times n}(R)$ multiplied by $T_{ij}(a)$ on the left will result in X with the jth row multiplied by a added to the ith row.
2. Any $X \in M_{m\times n}(R)$ multiplied by $D_i(a)$ on the left will result in X with the ith row multiplied by a.
3. Any $X \in M_{m\times n}(R)$ multiplied by P_{ij} on the left will result in X with the ith and jth rows switched.

Multiplication on the right does the same thing, just with columns instead of rows. These are the elementary row/column operations you'd do on matrices in normal linear algebra.

Note. Right multiplying would require the elementary matrix to be $n\times n$, and left multiplying $m \times m$. We do not distinguish the two in our notation, and assume that the exact dimension is clear from context.

Our goal will be to transform any matrix into its normal form (the one in Theorem 4.5.1) using these elementary matrices. We'd therefore like each of these elementary matrices to be reversible. This can be done by enforcing the following conditions.

1. $D_i(a)$ must have a be a unit.
2. $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ must be invertible to use $\chi(a,b,c,d)$.

It is not too difficult to check that these will suffice. Before doing our proof, we need one final definition.

Definition 4.5.2. Let R be a PID. The length of an element $a \in R$, denoted $\ell(a)$, is the number of prime factors (including multiplicities) in a prime factorization of a. If a is a unit, we say that $\ell(a) = 0$.

Note that since prime factorizations are unique up to multiplication by units, the above is a well-formed definition. Now, let's finally do this proof.

Proof of Theorem 4.5.1. Let $A \in M_{m,n}(R)$ be an arbitrary matrix. If $A = 0$, then we are done. Otherwise, suppose that $a_{11} \nmid a_{1k}$, for some $1 \leq k \leq n$. By swapping the kth and 2nd columns (multiplying by P_{2k} on the left), we may assume that $a_{11} \nmid a_{12}$. Define $x = a_{11}, y = a_{12}, z = \mathrm{GCD}(a,b)$. Then there exists some $a, c \in R$ such that $xa + yc = d$ (this is because Bézout's lemma holds in any PID[1]). Defining $b = yd^{-1}, d = -xd^{-1}$ (note that these are well-defined since $d \mid x, y$), we get that

$$\begin{pmatrix} -d & b \\ c & -a \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \mathrm{Id}_2$$

[1]I will provide a proof of this at the end of the section.

Thus, we can multiply by $\chi(a,b,c,d)$. Indeed, multiplying A by $\chi(a,b,c,d)$ on the right gives a matrix whose first row is $(z,0,a_{13},\ldots,a_{1n})$, where we note that $\ell(z)<\ell(a_{11})$. We may then repeat this process until the $(1,1)$ entry of our matrix divides every element of the first row, then apply the entire process again on the first column, giving us

$$\begin{pmatrix} f & a'_{12} & a'_{13} & a'_{14} & \cdots & a'_{1n} \\ a'_{21} & a'_{22} & a'_{23} & a'_{24} & \cdots & a'_{2n} \\ a'_{31} & a'_{32} & a'_{33} & a'_{34} & \cdots & a'_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a'_{m1} & a'_{m2} & a'_{m3} & a'_{m4} & \cdots & a'_{mn} \end{pmatrix}$$

where the a'_{ij} are not necessarily the same as a_{ij}, and $f \mid a'_{1j}, a'_{i1}$. Note that this process is only guaranteed to terminate since the length of the element in the (1,1) position strictly decreases every time. We can at this point use the $T_{j1}(x), T_{i1}(x)$ matrices (i.e. column and row addition) with the proper $x \in R$ to put our matrix in the form

$$\begin{pmatrix} f & 0 & 0 & 0 & \cdots & 0 \\ 0 & a'_{22} & a'_{23} & a'_{24} & \cdots & a'_{2n} \\ 0 & a'_{32} & a'_{33} & a'_{34} & \cdots & a'_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a'_{m2} & a'_{m3} & a'_{m4} & \cdots & a'_{mn} \end{pmatrix}$$

These new a'_{ij} have no particular relation to the old a'_{ij} (I'll continue using them as placeholder variables for the rest of this proof). Adding the second row of our matrix to the first, and repeating the process originally applied to the first row and column, we get a matrix of the same form. Note that if we apply any of the elementary matrices from the above processes to the sub-matrix (a'_{ij}), then it does not affect divisibility by f. Thus, we get a situation where $f \mid a'_{2j}$. We can then do this again with all the rest of the rows in the matrix (switching each of them to be the second row first), again knowing this process will terminate by the strict decreasing of $\ell(f)$. Note that since each new f divides the previous one, this results in a matrix of the above form where $f \mid a'_{ij}$. We can then apply the same process as above to the sub-matrix $(a_{ij})'$. Note that this, again, will not affect divisibility by f, so we get a matrix of the form

$$\begin{pmatrix} d_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & a'_{33} & a'_{34} & \cdots & a'_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a'_{m3} & a'_{m4} & \cdots & a'_{mn} \end{pmatrix}$$

Repeating this process continually to each new, smaller, sub-matrix, we get the desired result. □

The above proof is still quite dense and hard to follow, as if I had gone through and justified every little detail it would've run for far too many pages. It's worth going through and

proving any of the details which are unclear to you, or potentially running the algorithm yourself ([Jac09] has some practice problems if you really have too much time on your hands).

We call form of the matrix in Theorem 4.5.1 the *normal form* of the matrix, and the d_i its *invariant factors*. Of course, we have yet to prove that any of these things are unique. To do this, we'll need a pair of definitions.

Definition 4.5.3. Let $A \in M_{m,n}(R)$, where R is any ring. If $r \leq m, n$, then an r-rowed minor of A is the determinant of a sub-matrix of A given by only including elements from r rows and r columns from A. We say that A is of rank r if either

1. A has a non-zero $r = \min(m, n)$-rowed minor.
2. $r < \min(m, n)$, A has a non-zero r-rowed minor, and A has no non-zero $(r+1)$-rowed minors.

Note. These are just generalizations of the same concepts from linear algebra. Also, the recursive formula for the determinant implies that if A has rank r, then it has a non-zero $(r-k)$-rowed minor for $0 \leq k < r$, and all $(r+k)$-rowed minors are zero for $k \geq 1$ (indeed, this is required for each matrix to have a unique rank).

Theorem 4.5.4. *Suppose $A \in M_{m,n}(R)$ is of rank r, where R is a PID. Then any normal form of A will have r invariant factors $d_1, \ldots, d_n$. Furthermore, if Δ_i is a GCD of all the i-rowed minors of A, then the d_i differ by unit multipliers from*

$$d_1 = \Delta_1 \qquad d_2 = \Delta_2 \Delta_1^{-1} \qquad \cdots \qquad d_r = \Delta_r \Delta_{r-1}^{-1}$$

Proof. Let $Q \in M_m(R), P \in M_n(R)$ be invertible. Note that

$$(QA)_{ij} = \sum_{k=1}^{n} q_{ik} a_{kj}$$

Thus, any r-rowed minor of QA is a linear combination of r-rowed minors of A, making the rank of QA at most that of A and the Δ_i of QA an element which divides that of A. A similar result holds for AP, so it follows that QAP is at most of rank r and has minor GCDs dividing those of A. In particular, by Theorem 4.5.1, we can choose Q, P such that QAP is in normal form. But of course the same argument would apply to transforming QAP to A using Q^{-1}, P^{-1}, so it follows that the ranks of QAP and A are the same, as are their minor GCDs (up to unit multipliers). Since $\Delta_i \mid \Delta_{i+1}$ (by the recursive formula for the determinant), the desired result follows. □

Again, I cut some corners here in the proof for the sake of my sanity and the readability of the text. If there are any assertions which don't seem clearly true to you, take the time to work through them until they are.

The most important application of this uniqueness is in the following corollary.

Corollary 4.5.4.1. *Two matrices are similar if and only if they have the same invariant factors/normal forms (up to unit multipliers of the invariant factors).*

Finally, let's clean up a pair of loose ends. I'll first answer a question that may have popped into your head.

Wouldn't this imply that all matrices are diagonalizable?

The answer is, of course, no. Suppose A is our matrix, and P, Q invertible matrices such that $B = QAP^{-1}$ is in normal form. The matrix is diagonalizable only if we can choose $Q = P$. Indeed, otherwise when interpreting the action of our matrix A using B, we'd have to use different bases for the input and output. This isn't bad per se, but it's not diagonalization.

Second, I'll fulfill a promise I made before and prove the following.

Proposition 4.5.5. *Bézout's identity holds in any PID.*

Proof. Suppose R is a PID, and $a, b \in R$. Then there exists some $c \in R$ such that $(a, b) = (c)$. Since $c \mid a, b$, we conclude that c divides any GCD of a, b. Thus, any GCD d of a, b is in $(c) = (a, b)$, and hence there exist $x, y \in R$ such that $ax + by = d$. □

4.6 Structure Theorem

I'll come clean with you here, the normal form of a matrix is not (at least in my experience) particularly useful to you in most situations. We proved it instead in service of the following theorem of utmost importance and utility (after a quick definition).

Theorem 4.6.1 (Structure Theorem for Finitely Generated Modules Over a PID)**.** *Suppose $M \neq 0$ is a finitely generated module over a PID R. Then there exist elements $d_1, \ldots, d_r \in R$ such that*

1. $M \cong R/(d_1) \oplus R/(d_2) \oplus \cdots \oplus R/(d_r)$
2. $(d_1) \supset (d_2) \supset \cdots \supset (d_r)$

Proof. Suppose $\underline{v}_1, \ldots, \underline{v}_n$ is a set of generators for M. We can define a homomorphism $\varphi : R^n \to M$ by $\varphi : \underline{e}_i \mapsto \underline{v}_i$. This homomorphism is clearly surjective, so $M \cong R^n / \ker(\varphi)$. By Theorem 4.4.1, $\ker(\varphi)$ is a free submodule of R^n of rank $m \leq n$. Let $\underline{u}_1, \ldots, \underline{u}_m$ be a basis for $\ker(\varphi)$, and let $A \in M_{m,n}(R)$ be the relations matrix between this and the standard basis. By Theorem 4.5.1, there exist two invertible matrices $P \in M_n(R), Q \in M_m(R)$ such that QAP^{-1} is in normal form. Define

$$\underline{e}'_i = \sum_{j=1}^{n} p_{ij} \underline{e}_j$$
$$\underline{u}'_i = \sum_{j=1}^{m} q_{ij} \underline{u}_j$$

Then by our observations in section 4.4, $\{\underline{e}'_i\}_{i=1}^n$ is another basis for R^n, $\{\underline{u}'_i\}_{i=1}^m$ another basis for $\ker(\varphi)$ (the fact that this is a basis and not just a set of generators follows from Q being

invertible), and QAP^{-1} is the relations matrix between them. That is, allowing the abuse of notation that zeroes on the diagonal (or diagonals for rows past m) are still counted as invariant factors of A, and denoted said invariant factors $d_1, \dots, d_n$, then

$$\underline{u}'_i = d_i \underline{e}'_i$$

Using another notational trick here that $\underline{u}'_i = \underline{0}$ for $i > m$. But of course, we can also regard R^n as $R\underline{e}'_1 \oplus R\underline{e}'_2 \oplus \cdots \oplus R\underline{e}'_n$. In this view, it's then clear that $\ker(\varphi)$ is $(d_1) \oplus (d_2) \oplus \cdots \oplus (d_n)$, and hence $R^n/\ker(\varphi) \cong R/(d_1) \oplus R/(d_2) \oplus \cdots R/(d_n)$. By Theorem 4.5.1, and the fact that everything divides zero, we know that $d_1 \mid d_2 \mid \cdots \mid d_n$, and hence $(d_1) \supset (d_2) \supset \cdots \supset (d_n)$. □

Note. I've again, for the sake of my sanity, skipped over some technical details in this proof. [Jac09]'s proof, although admittedly of a slightly different variant of this theorem, goes into all those excruciating details if you really want them. They mostly amount to justifying treating changes of basis flippantly.

Of course, it'd be really nice to have some sort of uniqueness on this theorem. It turns out this does exist, and we'll spend the remainder of this section (mostly) building up to it. Before we do that, I'd like to present a slightly different form of the structure theorem, the one you might've seen if you looked in [Jac09].

Corollary 4.6.1.1. *Suppose $M \neq 0$ is a finitely generated module over a PID R. Then there exist elements $\underline{v}_1, \dots, \underline{v}_r \in M$ such that*

1. $M \cong R\underline{v}_1 \oplus R\underline{v}_2 \oplus \cdots \oplus R\underline{v}_r$
2. $\operatorname{ann}(\underline{v}_1) \supset \operatorname{ann}(\underline{v}_2) \supset \cdots \supset \operatorname{ann}(\underline{v}_r)$

There's a quick question to be answered here before we do the proof, namely what is ann?. Simply put, if $U \subset M$ is a subset of a module M, then

$$\operatorname{ann}(U) = \{r \in R \mid \forall \underline{u} \in U, r\underline{u} = \underline{0}\}$$

A quick check shows that these *annihilators*, as they're called, are always ideals in R. Anyway, on to the proof.

Proof. Consider the isomorphism $\varphi : R/(d_1) \oplus R/(d_2) \oplus \cdots \oplus R/(d_r) \to M$ from the structure theorem. Define $\underline{v}_i = \varphi(\underline{e}_i)$. Then clearly

$$M \cong R\underline{v}_1 \oplus \cdots \oplus R\underline{v}_r$$

and $\operatorname{ann}(\underline{v}_i) = (d_i)$, giving the desired result. □

The main thing we're going to focus on here is not annihilators, but instead a related concept.

Definition 4.6.2. Let M be a module. An element $\underline{v} \in M$ is called torsion if $\operatorname{ann}(\underline{v}) \neq 0$. The torsion submodule of M is the set of all torsion elements of M, and is denoted M^{tor}. If $M = M^{tor}$, we call M a torsion module.

As the name would suggest, the torsion submodule is in fact always a submodule. In order to understand the structure of modules further, we're going to have to work a bit more with it.

Theorem 4.6.3. *Suppose M is a finitely generated module over a PID R. Then M is the direct sum of its torsion submodule and a free submodule.*

Proof. By the above corollary of the structure theorem, we know that there exist elements $\underline{v}_1, \ldots, \underline{v}_r \in M$ such that

1. $M \cong R\underline{v}_1 \oplus R\underline{v}_2 \oplus \cdots \oplus R\underline{v}_r$
2. $\text{ann}(\underline{v}_1) \supset \text{ann}(\underline{v}_2) \supset \cdots \supset \text{ann}(\underline{v}_r)$

In particular, we may assume that none of these elements are zero (or equivalently none of them are annihilated by R), and that the first s are torsion while the rest are not. It is then a quick verification that

$$M^{tor} \cong R\underline{v}_1 \oplus \cdots \oplus R\underline{v}_s$$
$$R^{r-s} \cong R\underline{v}_{s+1} \oplus \cdots \oplus R\underline{v}_r$$

completing the proof. □

The free part of the above decomposition we understand the structure (and uniqueness) of quite well, so we're going to focus in on the torsion submodule.

Definition 4.6.4. Let $p \in R$ be a prime and M an R-module. Then $M_p \subset M$ is the submodule of M composed of all $\underline{v} \in M$ such that for some $k \in \mathbb{N}$, $p^k\underline{v} = \underline{0}$.

Note. I'll leave it to you to show that this is indeed always a submodule.

Lemma 4.6.5. *If $p_1, \ldots, p_r \in R$ are distinct primes, then $M_{p_1}, \ldots, M_{p_r}$ are linearly independent.*

Proof. It suffices to show that $M_{p_1} \cap (M_{p_2} + \cdots + M_{p_r}) = \emptyset$. To that end, suppose that $\underline{v} \in M_{p_1} \cap (M_{p_2} + \cdots + M_{p_r}) = \emptyset$, say $\underline{v} = \underline{u}_2 + \cdots + \underline{u}_r$, where $\underline{u}_i \in M_{p_i}$. For each $2 \leq i \leq r$, let k_i be the exponent required to annihilate $\underline{u}_i$, with k_1 that exponent for $\underline{v}$. By Bézout's identity, there exist $x, y \in R$ such that $xp_1^{k_1} + yp_2^{k_2} \cdots p_r^{k_r} = 1$. Thus, we get

$$\underline{v} = (xp_1^{k_1} + yp_2^{k_2} \cdots p_r^{k_r})(\underline{u}_2 + \cdots + \underline{u}_r) = xp_1^k\underline{v} + yp_2^{k_2} \cdots p_r^{k_r}(\underline{u}_2 + \cdots + \underline{u}_r) = \underline{0}$$

as claimed. □

Lemma 4.6.6. *The following hold for any module M over a PID R.*

1. *If $M = R\underline{v}$, where $\text{ann}(\underline{v}) = (d)$ and $d = gh$ such that $GCD(g, h) = 1$, then $\exists \underline{u}_1, \underline{u}_2 \in M$ such that $\text{ann}(\underline{u}_1) = (g)$, $\text{ann}(\underline{u}_2) = h$, and $M \cong R\underline{u}_1 \oplus R\underline{u}_2$.*
2. *If $M \cong R\underline{u}_1 \oplus R\underline{u}_2$, where $\text{ann}(\underline{u}_1) = (g)$, $\text{ann}(\underline{u}_2) = h$, $\underline{u}_1, \underline{u}_2 \in M$, and $GCD(g, h) = 1$, then there exists $\underline{v} \in M$ such that $M \cong R\underline{v}$, where $\text{ann}(\underline{v}) = (gh)$.*

Proof. We prove these one by one.

1. Set $\underline{u}_1 = h\underline{v}$, $\underline{u}_2 = g\underline{v}$. Note that $r\underline{u}_1 = \underline{0}$ only if $d \mid rh$, which since $\text{GCD}(g,h) = 1$ is only possible if $g \mid r$. Thus, $\text{ann}(\underline{u}_1) = (g)$, and similarly $\text{ann}(\underline{u}_2) = (h)$. Finally, we show that the desired isomorphism exists. Let $x, y \in R$ be such that $xg + yh = 1$. Then $\underline{v} = (xg + yh)\underline{v} = y\underline{u}_1 + x\underline{u}_2$, so $R\underline{v} = R\underline{u}_1 + R\underline{u}_2$. Suppose $\underline{u} \in R\underline{u}_1 \cap R\underline{u}_2$. Then $\text{ann}(\underline{u}) \supset (g) + (h) = R$, so $\underline{u} = \underline{0}$. Thus, by Theorem 4.3.4, $R\underline{v} \cong R\underline{u}_1 \oplus R\underline{u}_2$.

2. Set $\underline{v} = \underline{u}_1 + \underline{u}_2$. Suppose that $r \in \text{ann}(\underline{v})$. Then $r\underline{u}_1 + r\underline{u}_2 = \underline{0} \Rightarrow r\underline{u}_1 = -r\underline{u}_2$. Thus, $h \in \text{ann}(r\underline{u}_1)$, so $g, h \mid r$. Since $\text{GCD}(g,h) = 1$, it follows that $gh \mid r$. Therefore, $\text{ann}(\underline{v}) = (gh)$. Since $M = R\underline{u}_1 + R\underline{u}_2$, it suffices to show that $R\underline{v} = R\underline{u}_1 + R\underline{u}_2$, as we already know the sum to be direct. $R\underline{v} \subset R\underline{u}_1 + R\underline{u}_2$ is clear. Let $a, b \in R$ be such that $ag + bh = 1$. Then $\underline{u}_1 = ah\underline{u}_1 = ah(\underline{v} - \underline{u}_2) = ah\underline{v}$, so $\underline{u}_1 \in R\underline{v}$. Similarly, $\underline{u}_2 \in R\underline{v}$, so $R\underline{v} \supset R\underline{u}_1 + R\underline{u}_2$.

□

Note. These results may look a little more intuitive in the following form.

1. If $M = R/(d)$, where $d = gh$ such that $\text{GCD}(g,h) = 1$, then $M \cong R/(g) \oplus R/(h)$.
2. If $M \cong R/(g) \oplus R/(h)$, where $\text{GCD}(g,h) = 1$, then $M \cong R/(gh)$.

Using these results, we can re-phrase part of the structure theorem in terms of primary components.

Theorem 4.6.7. *Suppose M is a finitely generated torsion module over a PID R. Then there exist (possibly non-distinct) prime elements $p_1, \ldots, p_n \in R$ and exponents $k_1, \ldots, k_n \in \mathbb{N}$ such that*

$$M \cong R/(p_1^{k_1}) \oplus \cdots \oplus R/(p_n^{k_n})$$

Proof. By the structure theorem, there exist $d_1, \ldots, d_r \in R$ such that

$$M \cong R/(d_1) \oplus \cdots \oplus R/(d_r)$$

The result follows by applying the first point of lemma 4.6.6 repeatedly to each of the terms in the direct sum. □

Note. If you look carefully, you can see that this proof actually gives us a procedure for converting between the two forms of the structure theorem. To see it written out, one can take a look at the corresponding section in [Jac09].

Note that, again, we can re-phrase this in the following way.

Corollary 4.6.7.1. *Suppose M is a finitely generated torsion module over a PID R. Then there exist (possibly non-distinct) prime elements $p_1, \ldots, p_n \in R$, exponents $k_1, \ldots, k_n \in \mathbb{N}$, and elements $\underline{v}_1, \ldots, \underline{v}_n \in M$ such that $\text{ann}(\underline{v}_i) = (p_i^{k_i})$ and*

$$M \cong R\underline{v}_1 \oplus \cdots \oplus R\underline{v}_n$$

That's not all either, we can actually derive a more surprising result using the following.

Lemma 4.6.8. *Let M be a finitely generated torsion module over a PID R. Then there exist finitely many primes $p_1, \dots, p_n$ such that $M \cong M_{p_1} \oplus \cdots \oplus M_{p_n}$.*

Proof. That there are only finitely many primes $p \in R$ such that $M_p \neq 0$ follows from corollary 4.6.7.1, and that the sum is direct was proven in lemma 4.6.5. □

Corollary 4.6.8.1. *Suppose $p \in R$ is a prime from the decomposition given by corollary 4.6.7.1. Then the direct sum of all $R\underline{v}_i$ such that $p \in \text{ann}(\underline{v}_i)$ is isomorphic to M_p.*

Is the above result useful? Not really for our purposes, but I just think it's interesting. Anyhow, we can now (finally) prove the invariance of the structure theorem.

Theorem 4.6.9. *Let $M \cong R\underline{v}_1 \oplus \cdots R\underline{v}_n$, $M \cong R\underline{u}_1 \oplus \cdots \oplus R\underline{u}_m$ be two decompositions in the form of Theorem 4.6.1. Then $n = m$ and $\text{ann}(\underline{v}_i) = \text{ann}(\underline{u}_i)$.*

Proof. This proof follows that given in [Jac09]. We start by reducing to the case of M being torsion. Indeed, we can group the terms in each decomposition by putting all the elements with non-zero annihilators first. Then the direct sums of those elements are M^{tor}, and modding out by M^{tor} we see that the remaining number of terms must be the same for both. Thus, we may assume that M is torsion. In fact, consider the decomposition given in corollary 4.6.7.1. Note that the $\text{ann}(\underline{v}_i), \text{ann}(\underline{u}_i)$ uniquely determine the resulting primary decomposition, and vice-versa. Thus, the above proof is equivalent to showing that our primary decomposition is uniquely determined. Since we can group the factors in that decomposition by base prime, and it's clear that the primes given in lemma 4.6.8 are unique, we can further assume that $M = M_p$, for some prime p. Now, after all of this let's prove the statement for the primary decomposition. Assume that $\text{ann}(\underline{v}_i) = (p^{k_i}), \text{ann}(\underline{u}_i) = (p^{f_i})$. Note that by assumption, $k_i \leq k_{i+1}$ and $f_i \leq f_{i+1}$. For any $x \in \mathbb{N}$, let $p^x M = \{p^x \underline{v} \mid \underline{v} \in M\}$. Note that each $p^x M$ is a submodule, and $M \supset pM \supset \cdots \supset p^{k_n} M \supset p^{k_{n+1}} M = 0$. A similar result holds using p^{f_m}. Define $M^{(r)} = p^r M / p^{r+1} M$. Note that any element of this has the form $p^r \underline{v} + p^{r+1} M$, so $\text{ann}(M^{(r)}) = (p)$ (or R, but we'll assume that $p^r M \neq 0$ here). Thus, $M^{(r)}$ with the same operations is also an $R/(p)$ module. But since p is prime $R/(p)$ is maximal, so $R/(p)$ is a field and hence $M^{(r)}$ is a finite-dimensional vector space over $R/(p)$. Note that the dimension of $M^{(r)}$ is the number of terms of the form p^x, with $x > r$, in each decomposition. Thus, each must have an equal number of these terms, so $n = m$ and $k_i = f_i$. □

Note that this proof also showed the uniqueness of our primary decomposition (which we can extend to non-torsion modules by adding allowing $p_i = 0$)! We call the $\text{ann}(\underline{v}_i)$ *invariant factors ideals* in a decomposition of the form corollary 4.6.1.1, and *elementary divisor ideals* in the other case. The generating elements for these ideals are often called *invariant factors/elementary divisors* We also get two final, quite useful results from all this.

Corollary 4.6.9.1. *Suppose M, N are finitely generated modules over a PID R. Then they are isomorphic if and only if they have the same invariant factor ideals/elementary divisor ideals.*

This one should've been immediate, the second is less so.

Corollary 4.6.9.2. *Let G be a finitely generated Abelian group. Then there exist some $p_i, e_i \in \mathbb{Z}^+$ such that*

$$G \cong \mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_n^{e^n}\mathbb{Z}$$

where each p_i is a prime or zero such that $p_i^{e_i}$. Furthermore, this decomposition is unique.

Proof. This is just the applying the structure theorem by regarding Abelian groups as $\mathbb{Z}$-modules, with the action of $\mathbb{Z}$ being defined by repeated addition. □

Note. This is just an extension of our Theorem 2.8.11 from all the way back in chapter 2.

Chapter 5

Free Commutative Modules

In this chapter, we assume that all rings are commutative.

5.1 Basic Results

This might be the simplest section in the book, it's just establishing a couple of results.

Theorem 5.1.1. *Every vector space is free, and any linearly independent set in a vector space is contained in some basis of that vector space.*

Proof. We start by proving that every vector space is free. Let V be a vector space over a field F, and let Σ be the set of all finite linearly independent subsets of V. Note that this has a partial order given by set inclusion, and that a maximal element of Σ is by definition a basis for V. Thus, by Zorn's lemma it suffices to show that every totally ordered subset of Σ has an upper bound. Indeed, suppose that $\Omega \subset \Sigma$ is totally ordered. $Y = \bigcup_{X \in \Omega} X$ is linearly independent, and hence is the desired upper bound. But note also that we could've restricted Σ to sets containing a particular linearly independent set, and gotten the same result. Hence, we can for any linearly independent set find a basis containing it. □

Note. This is an equivalent result to the statement that every vector space has a basis, which is what we actually proved. It also turns out to be equivalent to the axiom of choice. If that makes you uncomfortable, then I apologize in advance for your attempts to prove things about vector spaces without bases.

If you want to see a more thorough proof specific to vector spaces, see [Rom07]. Since all of our rings are assumed to be commutative, and all modules free, all of our modules have a well-defined rank, which we'll denote $\text{rank}_R(M)$.

Theorem 5.1.2 (Rank-Nullity Theorem)**.** *Let V, W be vector spaces over a field F, and $\varphi \in \text{Hom}_F(V, W)$ a linear map. Then*

$$\text{rank}_F(V) = \text{rank}_F(\ker(\varphi)) + \text{rank}_F(\text{Im}(\varphi))$$

Proof. Let $\{\underline{v}_i\}_{i\in I}$ be a basis for $\ker(\varphi)$, and $\{\underline{v}_j\}_{j\in J}$ an extension of that to a basis of V. We claim that $\{\underline{v}_j + \ker(\varphi)\}_{j\in J\setminus I}$ is a basis for $V/\ker(\varphi)$, which immediately implies the desired result. That it is spanning is clear. Suppose there existed some $\alpha_j \in F$ such that

$$\underline{0} + \ker(\varphi) = \sum_{j\in J\setminus I} \alpha_j(\underline{v}_j + \ker(\varphi))$$

Then in particular, there exists some $\underline{u} \in \ker(\varphi)$ such that

$$\underline{0} = \sum_{j\in J\setminus I} \alpha_j \underline{v}_j + \underline{u}$$

Furthermore, there exist unique $\alpha_i \in F$ such that $\underline{u} = \sum_{i\in I} \alpha_i \underline{v}_i$. Thus, we get

$$\underline{0} = \sum_{j\in J} \alpha_j \underline{v}_j \Rightarrow \alpha_j = 0$$

so the set linearly independent as claimed. □

Corollary 5.1.2.1. *Suppose V is a vector space over F with subspace U. Then*

$$\operatorname{rank}_F(V) = \operatorname{rank}_F(U) + \operatorname{rank}_F(V/U)$$

Proof. Apply Theorem 5.1.2 to the quotient map $q : V \to V/U$. □

Corollary 5.1.2.2. *Suppose V is a vector space over F with subspace U. Then* $\operatorname{rank}_F(U) \leq \operatorname{rank}_F(V)$*, and if* $\operatorname{rank}_F(V) < \infty$ *then* $\operatorname{rank}_F(V) = \operatorname{rank}_F(U) \Rightarrow V = U$.

5.2 Dual Modules

This section is based on lecture notes taken from [Sil23], along with a similar section in [Lan05]. Let's start by defining our objects of study.

Definition 5.2.1. Let M be a module over a commutative ring R. The dual module of M, denoted $\overleftarrow{M}$, is the R-module $\operatorname{Hom}_R(M, R)$. Elements of the dual module are called linear functionals.

Dual modules aren't so much interesting by themselves so much as in how they're applied. The rest of this section will be dedicated to establishing the basic facts about dual modules to allows us to study these applications effectively.

Proposition 5.2.2. *Let $\{M_i\}_{i\in I}$ be a collection of R-modules, M an R-module, and $N \subset M$ a submodule. Then the following hold.*

1. $$\overleftarrow{\left(\bigoplus_{i\in I} M_i\right)} \cong \prod_{i\in I} \overleftarrow{M_i}$$

2.

$$\overleftarrow{M/N} \cong \{\varphi \in \overleftarrow{M} \mid \varphi|_N = 0\}$$

Proof. 1. It suffices to show that $\overleftarrow{\left(\bigoplus_{i\in I} M_i\right)}$ satisfies the universal property of direct products. To that end, define $\pi_i : \overleftarrow{\left(\bigoplus_{i\in I} M_i\right)} \to \overleftarrow{M_i}$ by the action of $\varphi \in \overleftarrow{\left(\bigoplus_{i\in I} M_i\right)}$ on the first component in the direct sum. Let M' be some other R-module and $f_i \in \mathrm{Hom}_R(N, \overleftarrow{M_i})$ be linear maps. Define $f : M' \to \overleftarrow{\left(\bigoplus_{i\in I} M_i\right)}$ by $f(\underline{v})$ acting on the ith component of the direct sum in the same way as $f_i(\underline{v})$. Then it is clear that f is the desired linear map, and that any other such map must also have the defining property of f, making f unique.

2. This follows quickly by noting that any $\varphi \in \overleftarrow{M}$ which is zero on N induces a linear functional on M/N, and that any linear functional on M/N can be extended to M by making it zero on N.

□

Corollary 5.2.2.1. *The dual of any finite-rank free module is free.*

Proof. This follows from point (1) of the above proposition as long as $\mathrm{Hom}_R(R, R) \cong R$. But of course any $\varphi \in \mathrm{Hom}_R(R, R)$ is entirely defined by its action on $1 \in R$, so this is clear. □

Note. This result does not hold in general for the dual of infinite-rank free modules.

We can explore the connection between the "freeness" of a module and its dual further by introducing the concept of *dual bases*.

Definition 5.2.3. Let M be a free R-module with basis $B = \{\underline{v}_i\}_{i\in I}$. The dual basis of B is $\overleftarrow{B} = \{\overleftarrow{\underline{v}_i}\}_{i\in I}$, where $\overleftarrow{\underline{v}_i} \in \overleftarrow{M}$ is defined by

$$\overleftarrow{\underline{v}_i}(\underline{v}_j) = \delta_{ij}$$

Note. The notation of $\overleftarrow{\underline{v}_i}$ is perhaps a bit poor, as the exact definition of that functional depends not only on $\underline{v}_i$ but also the basis that it's contained in.

Proposition 5.2.4. *Let M be a free R-module with basis $B = \{\underline{v}_i\}_{i\in I}$. Then $\overleftarrow{B}$ is linearly independent in $\overleftarrow{M}$.*

Proof. Suppose that $\alpha_i \in R$ are such that $\underline{0} = \sum_{i\in I} \alpha_i \overleftarrow{\underline{v}_i}$. Then in particular, we get that for any $j \in I$

$$0 = \sum_{i\in I} \alpha_i \overleftarrow{\underline{v}_i}(\underline{v}_j) = \alpha_j$$

Thus, $\alpha_j = 0$ for all $j \in I$, and $\overleftarrow{B}$ is linearly independent. □

Proposition 5.2.5. *Let M be a free R-module of rank $n \in \mathbb{N}$ with basis $B = \{\underline{v}_i\}_{i\in I}$. Then $\overleftarrow{B}$ spans $\overleftarrow{M}$.*

Proof. Pick any $\varphi \in \overleftarrow{M}$, and let $\alpha_i = \varphi(\underline{v}_i)$. We'll show that $\varphi = \sum_{i=1}^n \alpha_i \overleftarrow{\underline{v}_i}$, completing the proof. Indeed, pick any $\underline{u} = \sum_{i=1}^n \beta_i \underline{v}_i \in M$. Then

$$\Big(\sum_{i=1}^n \alpha_i \overleftarrow{\underline{v}_i}\Big)(\underline{u}) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \beta_j \delta_{ij} = \sum_{i=1}^n \alpha_i \beta_i = \sum_{i=1}^n \beta_i \varphi(\underline{v}_i) = \varphi(\underline{u})$$

giving us the claimed equality. □

Note. The above proof works only in the finite-rank case, and the result will often fail to hold for infinite-rank modules!

Corollary 5.2.5.1. *If M is a finite-rank R-module, then $M \cong \overleftarrow{M}$.*

We can go even further than the dual module and look at the dual of the dual module, which we call the *double dual* and denote $\overleftarrow{\overleftarrow{M}}$. This, it turns out, has a decent resemblance to the original module M. More explicitly, like how we have the dual basis, we also have a double dual basis, which is given by

$$\overleftarrow{\underline{v}_i}(\varphi) = \varphi(\underline{v}_i)$$

I'll leave it to the reader to show that this is, in fact, linearly independent, and is spanning when M is finite-rank.

Note. In the double dual basis, unlike the dual basis, the double dual element depends only on the original basis element, and is independent of the rest of the basis. Hence, the double dual of any element is well-defined, giving us a natural linear inclusion $\iota : M \to \overleftarrow{\overleftarrow{M}}$ by $\iota(\underline{v}) = \overleftarrow{\underline{v}}$.

We can also dualize[1] the linear maps that go with modules.

Definition 5.2.6. Let M, N be R-modules and $\varphi \in \mathrm{Hom}_R(M, N)$ a linear map. The dual map of φ, $\overleftarrow{\varphi} \in \mathrm{Hom}_R(\overleftarrow{N}, \overleftarrow{M})$, is the map defined by

$$(\overleftarrow{\varphi}(\psi))(\underline{v}) = \psi(\varphi(\underline{v}))$$

for any $\psi \in \overleftarrow{N}, \underline{v} \in M$.

Again, I'll leave it to you to show that this is, in fact, a well-defined linear map. There are some other interesting properties of dual maps, which I'll list below.

Proposition 5.2.7. *Let M, N, L be R-modules and $\varphi \in \mathrm{Hom}_R(M, N), \psi \in \mathrm{Hom}_R(N, L)$ linear maps. Then*

1. $\overleftarrow{\psi \circ \varphi} = \overleftarrow{\varphi} \circ \overleftarrow{\psi}$.
2. *If A is the matrix of φ with respect to some pair of bases for M and N, then A^t is the matrix of $\overleftarrow{\varphi}$ with respect to the dual bases.*

The proofs of these are tedious and not particularly enlightening, so I'll leave it to the reader to prove them if they wish (this is question 1 in homework 4 of [Sil23]).

[1] This notion of dualizing will be defined formally in chapter 7, and used heavily in the latter chapters of Part IV.

5.3 Pairings and Tensor Products

This section is based on a similar section in [Lan05], lectures from [Sil23], and lectures by Kalle Karu.

Definition 5.3.1. Let M, N, L be R-modules. A bilinear map/pairing is a map $\varphi : M \times N \to L$ which is linear in each argument. We call a pairing non-degenerate if

1. $\forall \underline{u} \in M$, if $\forall \underline{v} \in N, \varphi(\underline{u}, \underline{v}) = \underline{0}$ then $\underline{u} = \underline{0}$.
2. $\forall \underline{v} \in N$, if $\forall \underline{u} \in M, \varphi(\underline{u}, \underline{v}) = \underline{0}$ then $\underline{v} = \underline{0}$.

Otherwise, we call a pairing degenerate. A pairing is called alternating if $N = M$ and for all $\underline{v} \in M, \varphi(\underline{v}, \underline{v}) = \underline{0}$.

Example 5.3.1. The evaluation map $\psi : \overleftarrow{M} \times M \to R$ defined by $\psi(\varphi, \underline{v}) = \varphi(\underline{v})$ is bilinear.

That above example is not a one-off, and indeed indicates the strong relationship between pairings and dual spaces. For example, we have the following result.

Proposition 5.3.2. *Let M be an R-module and ψ the evaluation map from above. Let ψ' be the evaluation map for another R-module N. Let $\varphi \in \mathrm{Hom}_R(M, N)$. Then $\overleftarrow{\varphi}$ is the unique linear map such that for any $\gamma \in \overleftarrow{N}, \underline{v} \in M$*

$$\psi'(\gamma, \varphi(\underline{v})) = \psi(\overleftarrow{\varphi}(\gamma), \underline{v})$$

In fact, the above result allows us to define "dual maps" in that manner for any non-degenerate pairings.

Proposition 5.3.3. *Let M, N be an R-modules, $\psi : M \times N \to R$ a non-degenerate pairing, and $\varphi \in \mathrm{End}_R(N)$ a linear map. Define $\tilde{\varphi} \in \mathrm{End}_R(M)$ by, for any $\underline{u} \in M, \underline{v} \in N$*

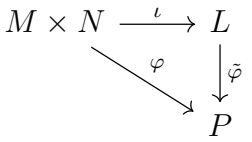
$$\psi(\underline{u}, \varphi(\underline{v})) = \psi(\tilde{\varphi}(\underline{u}), \underline{v})$$

This map is unique if well-defined.

This result isn't important for our purposes, but does come up (along with related results) quite a bit in functional analysis. If you're interested in seeing it developed in an algebraic context see [Sil23], or if you prefer analysis see [Fol99].

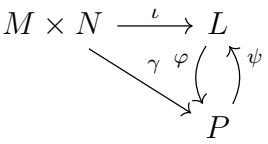
Our goal, for the rest of this section, will be to turn pairings into linear maps. That is, we're looking for modules satisfying the following universal property.

Definition 5.3.4. Let M, N be R-modules. A tensor product of M, N is an R-module L equipped with a bilinear map $\iota : M \times N \to L$ satisfying the following universal property: Given any R-module P and bilinear $\varphi : M \times N \to P$, there exists a unique linear map $\tilde{\varphi} : L \to P$ such that the following diagram commutes.

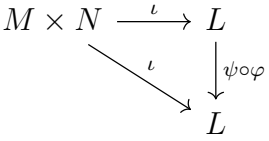


Proposition 5.3.5. *Suppose $(L, \iota), (P, \gamma)$ are two tensor products of M, N. Then there exists a unique isomorphism $\varphi : L \to P$ such that $\varphi \circ \iota = \gamma$.*

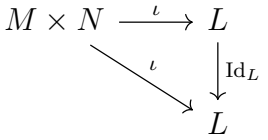
Proof. By the universal property of tensor products, there exists a unique homomorphism $\varphi : L \to P$ and unique homomorphism $\psi : P \to L$ making the following diagram commute.



Hence, the following diagram commutes



as does this one



But by the universal property, the homomorphism $f : L \to L$ making those two diagrams commute is unique, so $\psi \circ \varphi = \mathrm{Id}_L$. A similar proof shows that $\varphi \circ \psi = \mathrm{Id}_P$, making φ bijective, and the uniqueness of φ follows from the uniqueness in the universal property. □

Of course, I have not yet proven to you that tensor products exist. I'll do so now.

Proposition 5.3.6. *Let M, N be R-modules. Then there exists a tensor product of M, N.*

Proof. Start by defining the module $F = \bigoplus_{i \in M \times N} R$ as formal R-coefficient sums of elements of $M \times N$. That is, we denote by $(\underline{u}, \underline{v})$ the element with zeroes everywhere except in the ith position. Let H be the submodule of F generated by all expressions of the form

1. $(\underline{u}_1 + \underline{u}_2, \underline{v}) - (\underline{u}_1, \underline{v}) - (\underline{u}_2, \underline{v})$
2. $(\underline{u}, \underline{v}_1 + \underline{v}_2) - (\underline{u}, \underline{v}_1) - (\underline{u}, \underline{v}_2)$
3. $(x\underline{u}, \underline{v}) - x(\underline{u}, \underline{v})$
4. $(\underline{u}, x\underline{v}) - x(\underline{u}, \underline{v})$

where $\underline{u}_1, \underline{u}_2, \underline{u} \in M, \underline{v}_1, \underline{v}_2, \underline{v} \in N, x \in R$ (we'll keep these variables for the rest of the proof). We'll show that $L = F/H$ with bilinear map $\varphi : M \times N \to L$ given by $\varphi(\underline{u}, \underline{v}) = [(\underline{u}, \underline{v})]$ is a tensor product. First, we show that φ is bilinear. Indeed, we get

$$\varphi(x\underline{u}_1 + \underline{u}_2, \underline{v}) = [(x\underline{u}_1 + \underline{u}_2, \underline{v})] = x[(\underline{u}_1, \underline{v})] + [(\underline{u}_2, \underline{v})] = x\varphi(\underline{u}_1, \underline{v}) + \varphi(\underline{u}_2, \underline{v})$$
$$\varphi(\underline{u}, x\underline{v}_1 + \underline{v}_2) = [(\underline{u}, x\underline{v}_1 + \underline{v}_2)] = x[(\underline{u}, \underline{v}_1)] + [(\underline{u}, \underline{v}_2)] = x\varphi(\underline{u}, \underline{v}_1) + \varphi(\underline{u}, \underline{v}_2)$$

as required. Next, we show that (L, φ) satisfy the required universal properties. Suppose P is some other R-module, and $\psi : M \times N \to P$ a bilinear map. We'd like to define $f : L \to P$ by being linear with $f([(\underline{u}, \underline{v})]) = \psi(\underline{u}, \underline{v})$ (note that any homomorphism making the diagram commute must satisfy these properties, so f is unique if this is well-defined, and that f by definition makes our diagram commute). We need only check that this is well-defined. Indeed, we can see that this map would be inherited from the well-defined linear map $g : (\underline{u}, \underline{v}) \mapsto \psi(\underline{u}, \underline{v})$ from F to P as long as $H \subset \ker(g)$. To prove this, we just need to check that g is zero on the generators of H. We'll do so for generators of the type (1) and (3), the other two are substantially similar.

$$\begin{aligned} g((\underline{u}_1 + \underline{u}_2, \underline{v}) - (\underline{u}_1, \underline{v}) - (\underline{u}_2, \underline{v})) &= g((\underline{u}_1 + \underline{u}_2, \underline{v})) - g((\underline{u}_1, \underline{v})) - g((\underline{u}_2, \underline{v})) \\ &= \psi(\underline{u}_1 + \underline{u}_2, \underline{v}) - \psi(\underline{u}_1, \underline{v}) - \psi(\underline{u}_2, \underline{v}) = \underline{0} \\ g((x\underline{u}, \underline{v}) - x(\underline{u}, \underline{v})) &= g((x\underline{u}, \underline{v})) - xg((\underline{u}, \underline{v})) = \psi(x\underline{u}, \underline{v}) - x\psi(\underline{u}, \underline{v}) = \underline{0} \end{aligned}$$

□

Given that tensor products always exist and are unique, we generally refer to the tensor product from the above proposition as **the** tensor product, and denote it $M \otimes_R N$. We also take the convention of denoting $[(\underline{u}, \underline{v})]$ in this space by $\underline{u} \otimes_R \underline{v}$. We will often times drop the subscript R when the underlying ring is clear from context.

Proposition 5.3.7. *Let M, N, P be R-modules. Then the following maps are canonical isomorphisms (i.e. isomorphisms that play nice with the relevant universal properties).*

1. *The map $M \otimes N \to N \otimes M$ given by $\underline{u} \otimes \underline{v} \mapsto \underline{v} \otimes \underline{u}$.*
2. *The map $R \otimes M \to M$ given by $x \otimes \underline{v} \mapsto x\underline{u}$.*
3. *The map $(M \otimes N) \otimes P \to M \otimes (N \otimes P)$ given by $(\underline{u} \otimes \underline{v}) \otimes \underline{q} \mapsto \underline{u} \otimes (\underline{v} \otimes \underline{q})$.*
4. *The map $(M \oplus N) \otimes P \to (M \otimes P) \oplus (N \otimes P)$ given by $(\underline{u}, \underline{v}) \otimes \underline{q} \mapsto (\underline{u} \otimes \underline{q}, \underline{v} \otimes \underline{q})$.*

Proving this would be extremely tedious, so we won't do so. The main thing to take out of this is that the tensor product is, in a sense, commutative, associative, and distributive, and R acts like a unit for the tensor product. The associativity suggests that we should be able to take the tensor product of more than two modules, which is in fact the case using multilinear maps!

Definition 5.3.8. Let $\{M_i\}_{i \in I}, N$ be R-modules. A multilinear map is a map $f : \times_{i \in I} M_i \to N$ which is linear in each argument.

The definition of the tensor product, at this point, is just an extension of the usual universal property.

Definition 5.3.9. Let $\{M_i\}_{i\in I}$ be a collection of R-modules. A tensor product of the M_i is an R-module N along with a multilinear map $\iota : \times_{i\in I} M_i \to N$ satisfying the following universal property: For any R-module P and multilinear map $f : \times_{i\in I} \to P$, there exists a unique $\varphi \in \mathrm{Hom}_R(M, N)$ such that the following diagram commutes.

$$\begin{array}{ccc} \times_{i\in I} M_i & \xrightarrow{\iota} & N \\ & \searrow^{f} & \downarrow{\scriptstyle \varphi} \\ & & P \end{array}$$

Like our previous definition, this tensor product always exists and is unique up to a unique isomorphism. It also plays well with our previous definition in the way you'd expect, as

$$(M \otimes N) \otimes P \cong M \otimes N \otimes P \cong M \otimes (N \otimes P)$$

where that middle term is the tensor product of M, N, P, and all the isomorphisms are canonical as in proposition 5.3.7.

I'd like to mention one more thing before we move on to bases, namely that linear maps on the underlying spaces naturally induce linear maps on the tensor product. Indeed, suppose that M, N are R-modules and $f : M \to M, g : N \to N$ endomorphisms. Then the map $f \times g : M \times N \to M \otimes N$ given by $(\underline{u}, \underline{v}) \mapsto f(\underline{u}) \otimes f(\underline{v})$ is bilinear, and hence induces an endomorphism $f \otimes g : M \otimes N \to M \otimes N$ which is given by $(f \otimes g)(\underline{u} \otimes \underline{v}) = f(\underline{u}) \otimes g(\underline{v})$.

Now, on to bases. Our main result is the following.

Theorem 5.3.10. *Suppose M, N are free R-modules with bases $\{\underline{u}_i\}_{i\in I}, \{\underline{v}_j\}_{j\in J}$. Then $B = \{\underline{u}_i \otimes \underline{v}_j \mid i \in I, j \in J\}$ is a basis for $M \otimes N$.*

Proof. First, we show that B is linearly independent. Indeed, suppose that $\sum_{i\in I}\sum_{j\in J} \alpha_{ij}\underline{u}_i \otimes \underline{v}_j = \underline{0}$, where $\alpha_{ij} \in R$. Consider the bilinear map $\overleftarrow{\underline{u}_i} \times \overleftarrow{\underline{v}_j}$ given by $(\overleftarrow{\underline{u}_i} \times \overleftarrow{\underline{v}_j})(\underline{u}, \underline{v}) = \overleftarrow{\underline{u}_i}(\underline{u})\overleftarrow{\underline{v}_j}(\underline{v})$. This induces a linear map $\overleftarrow{\underline{u}_i} \otimes \overleftarrow{\underline{v}_j} : M \otimes N \to R$ given by $(\overleftarrow{\underline{u}_i} \otimes \overleftarrow{\underline{v}_j})(\underline{u} \otimes \underline{v}) = \overleftarrow{\underline{u}_i}(\underline{u})\overleftarrow{\underline{v}_j}(\underline{v})$. Applying this map to both sides of our expression, we get

$$\begin{aligned} 0 = (\overleftarrow{\underline{u}_i} \otimes \overleftarrow{\underline{v}_j})(\underline{0}) = (\overleftarrow{\underline{u}_i} \otimes \overleftarrow{\underline{v}_j})\Big(\sum_{k\in I,\ell\in J} \alpha_{k\ell}\underline{u}_k \otimes \underline{v}_\ell\Big) = \sum_{k\in I,\ell\in J} \alpha_{k\ell}(\overleftarrow{\underline{u}_i} \otimes \overleftarrow{\underline{v}_j})(\underline{u}_k \otimes \underline{v}_\ell) \\ = \sum_{k\in I,\ell\in J} \alpha_{k\ell}\overleftarrow{\underline{u}_i}(\underline{u}_k)\overleftarrow{\underline{v}_j}(\underline{v}_\ell) = \sum_{k\in I,\ell\in J} \alpha_{k\ell}\delta_{ik}\delta_{j\ell} = \alpha_{ij} \end{aligned}$$

Thus, B is linearly independent. To check that B is spanning, we just need to show that any element of the form $\underline{u} \otimes \underline{v}$ can be reached. But indeed, there exist $\alpha_i \in R, \beta_j \in R$ such that

$$\underline{u} = \sum_{i\in I} \alpha_i\underline{u}_i \qquad\qquad \underline{v} = \sum_{j\in J} \beta_j\underline{v}_j$$

so

$$\underline{u} \otimes \underline{v} = \sum_{i \in I, j \in J} \alpha_i \beta_j \underline{u}_i \otimes \underline{v}_j$$

as required. □

Corollary 5.3.10.1. *If M, N are free R-modules, then* $\operatorname{rank}_R(M \otimes N) = \operatorname{rank}_R(M) \operatorname{rank}_R(N)$ *(when this expression is well-defined).*

Note. Both of these results extend in the obvious way to tensor products of more than two modules.

Note. The second of these results shows that, in the case of finite-rank free modules, the tensor product and direct sum are almost never isomorphic. Indeed, one can check that $\operatorname{rank}_R(M \oplus N) = \operatorname{rank}_R(M) + \operatorname{rank}_R(N)$.

Note. Non-free modules can be much nastier than this, with the tensor product actually shrinking the modules. For example, $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z} = 0$ for any $n \in \mathbb{N}$.

There's one final concept I'd like to at least mention before we move on: the extension of scalars. You may remember from a basic course in linear algebra that if we have a vector space over $\mathbb{R}$, there are no problems in suddenly treating it like it's a vector space over $\mathbb{C}$ to find eigenvalues. Formally, this amounts to moving from V to $\mathbb{C} \otimes_{\mathbb{R}} V$, and defining this new space as a $\mathbb{C}$ vector space by always defining multiplication in the first argument. This same trick will work with any field and subfield, for more details see [Sil23].

5.4 Symmetric and Antisymmetric Products

This section is based on lecture notes by [Sil23]. In it, we focus in on the action of permutations on tensor products.

Definition 5.4.1. Let M be an R-module, $n \in \mathbb{N}$, and $\sigma \in S_n$. We define the action of σ on $\bigotimes_{k=1}^n M$ by the rule

$$\sigma(\underline{v}_1 \otimes \underline{v}_2 \otimes \cdots \otimes \underline{v}_n) = \underline{v}_{\sigma(1)} \otimes \underline{v}_{\sigma(2)} \otimes \cdots \otimes \underline{v}_{\sigma(n)}$$

and linearity.

Note. This is essentially a map from S_n to $\operatorname{End}_R(M)$, with the "action" of a permutation being applying its corresponding endomorphism.

Definition 5.4.2. Let M be an R-module, $n \in \mathbb{N}$. We call $\underline{v} \in \bigotimes_{k=1}^n M$ symmetric/even if, for every $\sigma \in S_n$, $\sigma(\underline{v}) = \underline{v}$. We call it antisymmetric/odd if, for every $\sigma \in S_n$, $\sigma(\underline{v}) = \operatorname{sgn}(\sigma)\underline{v}$.

Note. If $2 = 0$ in our ring, then symmetric and antisymmetric are the same thing.

For the rest of this, and the next, section, we will only care about (anti)symmetric vectors. Indeed, if $n \in \mathbb{N}$, we define the submodules of $\bigotimes_{k=1}^{n} M$

$$\mathrm{Sym}^n(M) = \left\{ \underline{v} \in \bigotimes_{k=1}^{n} M \mid \forall \sigma \in S_n, \sigma(\underline{v}) = \underline{v} \right\}$$
$$\bigwedge^n M = \left\{ \underline{v} \in \bigotimes_{k=1}^{n} M \mid \forall \sigma \in S_n, \sigma(\underline{v}) = \mathrm{sgn}(\sigma)\underline{v} \right\}$$

That is, the submodules of even and odd vectors. The interesting note here is that if M is free of finite rank, then we can explicitly construct a basis for these submodules.

Theorem 5.4.3. *Suppose M is a free R-module of rank $n \geq 2$, with basis $\{\underline{v}_i\}_{i=1}^n$. For any $k \in \mathbb{N}$, define the sets*

$$B_k^+ = \left\{ \sum_{\sigma \in S_k} \sigma(\underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k}) \mid \{i_1, \cdots, i_k\} \subset [n] \right\}$$
$$B_k^- = \left\{ \sum_{\sigma \in S_k} \mathrm{sgn}(\sigma)\sigma(\underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k}) \mid 1 \leq i_1 < i_2 < \cdots < i_k \leq n \right\}$$

where $[n] = \{1, 2, \cdots, n\}$. Then B_k^p is a basis for $\mathrm{Sym}^k(M)$, and B_k^- is a basis for $\bigwedge^k M$.

Proof. That these subsets are contained in their claimed submodules is clear. By induction on Theorem 5.3.10, the following set is a basis for $\bigotimes_{i=1}^{k} M$.

$$B_k = \left\{ \underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k} \mid i_1, \ldots, i_k \in [n] \right\}$$

It follows from this that B_k^+, B_k^- are linearly independent. Indeed, one can note that the $\sigma(\underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k})$ are always linearly independent for any choice of $\{i_1, \cdots, i_k\} \subset [n]$, and that each choice of $\{i_1, \cdots, i_k\} \subset [n]$ leads to a different set of basis vectors from applying σ. Next, we show that B_k^+, B_k^- are spanning. First, we'll look at B_k^+. Pick any $\underline{v} = \sum_{\{i_1,\ldots,i_k\} \subset [n]} \alpha_{i_1,\ldots,i_k} \underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k}$ in $\mathrm{Sym}^k(M)$, where $\alpha_{i_1,\ldots,i_k} \in R$. For any $\{i_1, \ldots, i_k\} \subset [n]$, the symmetry condition requires that

$$\alpha_{i_1,\ldots,i_k} = \alpha_{\sigma(i_1),\ldots,\sigma(i_k)}$$

for each $\sigma \in S_k$. Thus, we get

$$\underline{v} = \sum_{\{i_1,\ldots,i_k\} \subset [n]} \sum_{\sigma \in S_k} \alpha_{i_1,\ldots,i_k} \sigma(\underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k})$$

as required. Second, we look at B_k^-. Pick any $\{i_1, \ldots, i_k\} \subset [n]$. Note that the antisymmetry condition forces

$$\alpha_{i_1,\ldots,i_k} = \mathrm{sgn}(\sigma)\alpha_{\sigma(i_1),\ldots,\sigma(i_k)}$$

Assume, without loss of generality, that $i_1 = i_2$. Pick any ordering of $i_1, \ldots, i_k$, and let σ be the permutation which results in that ordering. Note that switching the position of i_1 and i_2

in this order does nothing, the two corresponding elements of B will be the same, but does flip the sign of the permutation. Thus, we get these two terms cancelling, resulting in the coefficients for all such basis elements being zero. Therefore, we may assume without loss of generality that $i_1, \dots, i_k$ are all distinct. Thus, we can assume (by applying the correct permutation and subsequent signs to the coefficients) that $i_1 < i_2 < \cdots < i_n$. We therefore get

$$\underline{v} = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \sum_{\sigma \in S_k} \alpha_{i_1,\dots,i_k} \operatorname{sgn}(\sigma) \sigma(\underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k})$$

as required. □

Corollary 5.4.3.1. *Suppose M is a free module of rank $n \geq 2$, and $k \in \mathbb{N}$. Then*

1. $\operatorname{rank}_R(\operatorname{Sym}^k(M)) = |\{k\text{-multisets of } [n]\}|$
2. $\operatorname{rank}_R(\bigwedge^k M) = \binom{n}{k}$

where we denote that $\binom{n}{k} = 0$ if $k > n$.

The second result in that corollary is the most interesting, as it demonstrates a pair of very interesting behaviours.

1. $\binom{3}{2} = 3$
2. $\binom{n}{n} = 1$

The first of these actually relates to the *cross product*, and more generally *exterior products*. For more details on this, see [Rom07]. The second, when combined with the tensor product of linear operators, gives us another way to approach determinants. Indeed, suppose that M is a rank n free R-module, and $T \in \operatorname{End}_R(M)$. If we were to represent T in matrix form relative to some basis, say this matrix is A, we could then calculate the determinant of that matrix to determine if T is invertible. But of course a change of basis amounts to changing our matrix to QAQ^{-1}, for some invertible matrix Q. This suggests that there should be some way of defining the determinant of T without appealing to bases, which is in fact the case.

Proposition 5.4.4. *Suppose M is a free R-module of rank n, and $k \in \mathbb{N}$. Then for any $T \in \operatorname{End}_R(M)$, $\bigotimes_{i=1}^k T \in \operatorname{End}_R(\bigotimes_{i=1}^k M)$ restricts to linear maps on the symmetric and antisymmetric submodules. We denote these restricted maps by $\operatorname{Sym}^k(T)$ and $\bigwedge^k T$.*

Proof. It suffices to show that these restrictions map basis elements to elements of the submodule. Indeed, suppose that $\{\underline{v}_i\}_{i=1}^n$ is a basis for M. Pick any $\{i_1, \dots, i_k\} \subset [n]$. Then by definition

$$\begin{aligned}\Big(\bigotimes_{j=1}^k T\Big)\Big(\sum_{\sigma \in S_k} \sigma(\underline{v}_{i_1} \otimes \cdots \otimes \underline{v}_{i_k})\Big) &= \sum_{\sigma \in S_k} T(\underline{v}_{i_{\sigma(1)}}) \otimes \cdots \otimes T(\underline{v}_{i_{\sigma(k)}}) \\ &= \sum_{\sigma \in S_k} \sigma(T(\underline{v}_{i_1}) \otimes \cdots \otimes T(\underline{v}_{i_k}))\end{aligned}$$

and

$$\Big(\bigotimes_{j=1}^{k} T\Big)\Big(\sum_{\sigma\in S_k}\operatorname{sgn}(\sigma)\sigma(\underline{v}_{i_1}\otimes\cdots\otimes\underline{v}_{i_k})\Big) = \sum_{\sigma\in S_k}\operatorname{sgn}(\sigma)T(\underline{v}_{i_{\sigma(1)}})\otimes\cdots\otimes T(\underline{v}_{i_{\sigma(k)}})$$
$$= \sum_{\sigma\in S_k}\operatorname{sgn}(\sigma)\sigma(T(\underline{v}_{i_1})\otimes\cdots\otimes T(\underline{v}_{i_k}))$$

giving the desired results. □

Definition 5.4.5. Let M be an R-module, $n = \operatorname{rank}_R(M)$, and $T \in \operatorname{End}_R(M)$. The determinant of T, denoted $\det(T)$, is the scalar multiple that $\bigwedge^n T$ is in $\bigwedge^n M$.

Note. This is well-defined since $\operatorname{rank}_R(\bigwedge^n M) = 1$, so every linear map on that space is just scalar multiplication.

Theorem 5.4.6. *Let M be an R-module, $n = \operatorname{rank}_R(M)$, $B = \{\underline{v}_i\}_{i=1}^n$ a basis for M, $T \in \operatorname{End}_R(M)$, and $A \in M_n(R)$ the matrix for T relative to B. Then* $\det(T) = \det(A)$.

Proof. By Theorem 5.4.3, one of the choices for the basis vector of $\bigwedge^n M$ is

$$\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\underline{v}_{\sigma(1)}\otimes\cdots\otimes\underline{v}_{\sigma(n)}$$

Writing $A = (a_{ij})$, we know by definition that

$$T(\underline{v}_j) = \sum_{i=1}^{n} a_{ij}\underline{v}_i$$

and

$$\det(A) = \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{1\sigma(1)}\cdots a_{n\sigma(n)}$$

Finally, we get

$$\Big(\bigwedge^n T\Big)\Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\underline{v}_{\sigma(1)}\otimes\cdots\otimes\underline{v}_{\sigma(n)}\Big) = \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)(T\underline{v}_{\sigma(1)})\otimes\cdots\otimes(T\underline{v}_{\sigma(n)})$$
$$= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\Big(\sum_{i=1}^{n} a_{i\sigma(1)}\underline{v}_i\Big)\otimes\cdots\otimes\Big(\sum_{i=1}^{n} a_{i\sigma(n)}\underline{v}_i\Big)$$
$$= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n} a_{i_1\sigma(1)}\cdots a_{i_n\sigma(n)}\underline{v}_{i_1}\otimes\cdots\otimes\underline{v}_{i_n}$$
$$= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\sum_{\omega\in S_n} a_{\omega(1)\sigma(1)}\cdots a_{\omega(n)\sigma(n)}\underline{v}_{\omega(1)}\otimes\cdots\otimes\underline{v}_{\omega(n)}$$
$$= \sum_{\omega\in S_n}\operatorname{sgn}(\omega)\Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\omega)\operatorname{sgn}(\sigma)a_{\omega(1)\sigma(1)}\cdots a_{\omega(n)\sigma(n)}\Big)\underline{v}_{\omega(1)}\otimes\cdots\otimes\underline{v}_{\omega(n)}$$

But since this is a multiple of $\sum_{\omega\in S_n}\operatorname{sgn}(\omega)\underline{v}_{\omega(1)}\otimes\cdots\otimes\underline{v}_{\omega(n)}$, it follows that

$$\sum_{\sigma\in S_n}\operatorname{sgn}(\omega)\operatorname{sgn}(\sigma)a_{\omega(1)\sigma(1)}\cdots a_{\omega(n)\sigma(n)}$$

is independent of ω. Thus, we can say that

$$\sum_{\sigma\in S_n}\operatorname{sgn}(\omega)\operatorname{sgn}(\sigma)a_{\omega(1)\sigma(1)}\cdots a_{\omega(n)\sigma(n)}=\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{1\sigma(1)}\cdots a_{n\sigma(n)}$$

Thus, we conclude

$$\begin{aligned}&\sum_{\omega\in S_n}\operatorname{sgn}(\omega)\Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\omega)\operatorname{sgn}(\sigma)a_{\omega(1)\sigma(1)}\cdots a_{\omega(n)\sigma(n)}\Big)\underline{v}_{\omega(1)}\otimes\cdots\otimes\underline{v}_{\omega(n)}\\&=\Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{1\sigma(1)}\cdots a_{n\sigma(n)}\Big)\sum_{\omega\in S_n}\operatorname{sgn}(\omega)\underline{v}_{\omega(1)}\otimes\cdots\otimes\underline{v}_{\omega(n)}\end{aligned}$$

giving us

$$\det(T)=\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{1\sigma(1)}\cdots a_{n\sigma(n)}=\det(A)$$

as claimed. □

There you go, a completely basis-free way of defining the determinant! Of course, we have one final (obvious) result that this implies.

Corollary 5.4.6.1. *Let M be an R-module, $n=\operatorname{rank}_R(M)$, and $T\in\operatorname{End}_R(M)$. Then T is invertible if and only if $\det(T)$ is a unit in R.*

5.5 Rational and Jordan Canonical Form

This section is based on similar ones from [Jac09] and [Sil23]. For this section, fix F to be a field, V an n-dimensional F-vector space, and $T\in\operatorname{End}_F(V)$. We first note that V can be regarded as an $F[\lambda]$ module, with scalar multiplication given by $\lambda\underline{v}=T\underline{v}$. We wish to study the structure of V as an $F[\lambda]$ module. It's clear that V is finitely generated over $F[\lambda]$. Our first non-trivial result is the following.

Proposition 5.5.1. *V is a torsion $F[\lambda]$ module.*

Proof. Pick any non-zero $\underline{v}\in V$. Then since $\dim_F(V)=n$, there exists some $1\leq m\leq n$ such that $\lambda^m\underline{v}\in\operatorname{Span}_F(\underline{v},\lambda\underline{v},\ldots,\lambda^{m-1}\underline{v})$. In particular then, there exist $a_i\in F$ such that

$$\lambda^m\underline{v}=\sum_{k=0}^{m-1}a_k\lambda^k\underline{v}$$

Thus, setting $f(\lambda)=\lambda^m-\sum_{k=0}^{m-1}a_k\lambda^k$, we get $f(\lambda)\underline{v}=\underline{0}$ and $f\neq 0$. □

Let $\{\underline{u}_i\}_{i=1}^n$ be a basis for V as an F-vector space. We get a natural homomorphism of $F[\lambda]$-modules $q : F[\lambda]^n \to V$ given by $q : \underline{e}_i \mapsto \underline{u}_i$. This map is surjective by construction, so we get that as $F[\lambda]$-modules

$$V \cong F[\lambda]/\ker(q)$$

Our next task then will be to study the structure of $\ker(q)$. Note that $F[\lambda]$ is a PID, so $\ker(q)$ is free. Thus, we search for a basis of $\ker(q)$.

Proposition 5.5.2. *Let $A = (a_{ij}) \in M_n(F)$ be the matrix of T relative to the basis $\{\underline{u}_i\}_{i=1}^n$. Then the set*

$$B = \left\{\underline{v}_i = \lambda\underline{e}_i - \sum_{j=1}^n a_{ji}\underline{e}_j \mid 1 \leq i \leq n\right\}$$

is a basis for $\ker(q)$.

Proof. That each of these is in the kernel of q is clear, as by definition in V

$$\lambda\underline{u}_i = \sum_{j=1}^n a_{ji}\underline{u}_j$$

Thus, we just need to check that these are spanning and linearly independent. Note that

$$\lambda\underline{e}_i = \underline{v}_i + \sum_{j=1}^n a_{ji}\underline{e}_j$$

Thus, any element of the form $\sum_{i=1}^n f_i(\lambda)\underline{e}_i$, where $f_i \in F[\lambda]$, can be written in the form

$$\sum_{i=1}^n g_i(\lambda)\underline{v}_i + \sum_{i=1}^n b_i\underline{e}_i$$

for some $g_i \in F[\lambda], b_i \in F$. If this element is in $\ker(q)$, then since each $\underline{v}_i \in \ker(q)$ we must get $\sum_{i=1}^n b_i\underline{e}_i \in \ker(q) \Rightarrow \sum_{i=1}^n b_i\underline{u}_i = \underline{0} \Rightarrow b_i = 0$. Thus, B is spanning. For linear independence, suppose that $\sum_{i=1}^n g_i(\lambda)\underline{v}_i = \underline{0}$. Then we'd get

$$\sum_{i=1}^n \left(g_i(\lambda)\lambda\underline{e}_i - \sum_{j=1}^n a_{ji}g_i(\lambda)\underline{e}_j\right) = \underline{0} \Rightarrow \sum_{i=1}^n \left(g_i(\lambda)\lambda - \sum_{j=1}^n a_{ij}g_j(\lambda)\right)\underline{e}_i = \underline{0}$$

This, of course, implies that

$$g_i(\lambda)\lambda = \sum_{j=1}^n a_{ij}g_j(\lambda)$$

Suppose, without loss of generality, that $g_i \neq 0$ is a polynomial g_j of maximal degree. Then this relation is clearly impossible unless $g_i = 0$. Thus, all the $g_i = 0$, making B linearly independent. □

We can note something further from this, namely that

$$\underline{v}_i = (\lambda \operatorname{Id}_n - A)\underline{e}_i$$

That is, $\lambda \operatorname{Id}_n - A$ is the matrix form of an isomorphism from $F[\lambda]^n$ onto $\ker(q)$ relative to the bases $\{\underline{e}_i\}_{i=1}^n, \{\underline{v}_i\}_{i=1}^n$. But of course $\lambda \operatorname{Id}_n - A \in M_n(F[\lambda])$, and $F[\lambda]$ is a PID. Thus, we can find some other pair of bases $\{\underline{e}'_i\}_{i=1}^n, \{\underline{v}'_i\}_{i=1}^n$ for $F[\lambda]^n, \ker(q)$ and invertible matrices $P, Q \in M_n(F[\lambda])$ such that

1. $P\underline{e}'_i = \sum_{j=1}^n p_{ji}\underline{e}_j$.
2. $Q\underline{v}'_i = \sum_{j=1}^n q_{ji}\underline{v}_j$.
3. $\lambda \operatorname{Id}_n - A = QDP^{-1}$, where D is a matrix in normal form.

In particular, since F is a field we can choose the polynomials $d_1, \ldots, d_n$ on the diagonal in D to all be monic if they are non-zero.

We can actually say more about this "diagonalization" by looking at the *characteristic polynomial*.

Definition 5.5.3. Let $A \in M_n(F)$. The characteristic polynomial of A, denoted $f_A \in F[\lambda]$, is $\det(\lambda \operatorname{Id}_n - A)$.

We can note here that f_A and $\det(PDQ)$ differ only by units, as P, Q are invertible. Furthermore, $\det(D) = d_1(\lambda) \cdots d_n(\lambda)$. Thus, since $F[\lambda]$ is a UFD, the d_i are uniquely determined by the characteristic polynomial. Furthermore, since $f_A \neq 0$, it also follows from this that none of the d_i are zero.

This is where we drag the structure theorem back into everything. Remember, we have by construction that

$$\underline{v}'_i = d_i(\lambda)\underline{e}'_i$$

Furthermore, by the proof of the structure theorem, we then get that as $F[\lambda]$-modules

$$V \cong F[\lambda]/(d_1(\lambda)) \oplus \cdots \oplus F[\lambda]/(d_n(\lambda))$$

There's another way of looking at this which may be more useful. We note that the image of each $\underline{e}'_i$ in M is $\underline{u}'_i = \sum_{j=1}^n p_{ji}\underline{u}_i$. Thus, we get

$$V \cong F[\lambda]\underline{u}'_1 \oplus \cdots \oplus F[\lambda]\underline{u}'_n$$

where $\operatorname{ann}_{F[\lambda]}(\underline{u}'_i) = (d_i(\lambda))$.

Note. It is very important to notice here that this notation is allowing for $d_i = 0$, in which case $F[\lambda]\underline{u}'_i \cong 0$.

Note what this implies : the linear transformation T acts on each component of this direct sum separately. Thus, considering these $F[\lambda]\underline{u}'_i$ as F-subspaces of V, we get a matrix for T in "block diagonal" form, with one block for each non-zero $F[\lambda]\underline{u}'_i$. The final thing we wish to do is study these blocks. To that end, we have the following proposition.

Proposition 5.5.4. *Let $A = (a_{ij}) \in M_n(F)$ be the matrix of T relative to the basis $\{\underline{u}_i\}_{i=1}^n$, and suppose that A only has one non-unit invariant factor, call it $d_n(\lambda)$ and suppose it is of degree $m \geq 1$. Then $V \cong F[\lambda]\underline{u}'_n$ as $F[\lambda]$-modules, $m = n$, $B = \{\underline{u}'_n, \lambda\underline{u}'_n, \ldots, \lambda^{m-1}\underline{u}'_n\}$ is a basis for V as an F-vector space, and relative to this basis T has the matrix form*

$$\begin{pmatrix} 0 & 0 & \cdots & 0 & \alpha_n \\ 1 & 0 & \cdots & 0 & \alpha_{n-1} \\ 0 & 1 & \cdots & 0 & \alpha_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha_1 \end{pmatrix}$$

where $d_n(\lambda) = \lambda^m - \sum_{i=1}^m \alpha_i \lambda^{n-i}$.

Proof. First, we show that B is spanning. Pick any $\underline{v} \in V$. Since $F[\lambda]\underline{u}'_n = V$, there exists some polynomial $g(\lambda) \in F[\lambda]$ such that $g(\lambda)\underline{u}'_n = \underline{v}$. Thus, it suffices to show that $\lambda^m \underline{u}'_n \in \mathrm{Span}_F(B)$. Indeed, we know that $d_n(\lambda)\underline{u}'_n = \underline{0}$. Thus, setting $d_n(\lambda) = \lambda^m - \sum_{i=1}^m \alpha_i \lambda^{m-i}$, we get

$$\lambda^m \underline{u}'_n = \sum_{i=1}^m \alpha_i \lambda^{m-i} \underline{u}'_n$$

as required. Note that $n = m$ would then immediately imply that B is a basis. We cannot have $m < n$ as the dimension of V is at most m. Suppose $m > n$. Then B would be linearly dependent, so there would exist $b_i \in F$ such that $\sum_{i=1}^m b_i \lambda^{i-1} \underline{u}'_n = \underline{0}$. But then a polynomial of degree $m-1$ would annihilate $\underline{u}'_n$, so $\mathrm{ann}_{F[\lambda]}(\underline{u}'_n) \neq (d_n)$. Thus, we cannot have $m > n$, so $m = n$ as required. Finally, we consider the form of the matrix for T relative to the basis B. If $k < m - 1$, we get

$$T(\lambda^k \underline{u}'_n) = \lambda^{k+1} \underline{u}'_n$$

If $k = m - 1$, we instead get

$$T(\lambda^k \underline{u}'_n) = \lambda^m \underline{u}'_n = \sum_{i=1}^m \alpha_i \lambda^{n-i} \underline{u}'_n$$

Thus, our matrix is of the form

$$\begin{pmatrix} 0 & 0 & \cdots & 0 & \alpha_n \\ 1 & 0 & \cdots & 0 & \alpha_{n-1} \\ 0 & 1 & \cdots & 0 & \alpha_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha_1 \end{pmatrix}$$

as claimed. □

A matrix of the above form is often called the *companion matrix* for the polynomial $d_n(\lambda)$. Finally, we can bring these observations all together into the following.

Theorem 5.5.5. *Suppose V is an F-vector space of finite dimension n, and $T \in \mathrm{End}_F(V)$. Then there exists a basis in which the matrix for T is in block diagonal form*

$$\begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_m \end{pmatrix}$$

where each D_i is the companion matrix for a polynomial $d_i \in F[\lambda]$ of degree at least one, $m \leq n$, and $d_1 \mid d_2 \mid \cdots \mid d_m$.

If A is the matrix for T in any other basis, then we call this block diagonal form the *rational canonical form* of A. Note that the d_i are uniquely determined by the characteristic polynomial of the matrix, which leads to the following result.

Theorem 5.5.6. *Suppose $A, B \in M_n(F)$. Then they differ by a change of basis, i.e. there exists an invertible $Q \in M_n(F)$ such that $B = QAQ^{-1}$, if and only if they have the same rational canonical form.*

This is all quite dense, so I don't blame you if you're having trouble following everything at this point. I did as well when I first did this, and even now I had to refer to [Jac09] frequently to remember how all this worked. I ask that you stick with me just a bit longer!

The astute amongst you might have noticed that we only used one form of the structure theorem above, namely the invariant factors formulation, and asked whether we can get a similar result using the elementary divisors formulation. The answer turns out to be yes, but only in some specific circumstances. To start off, we can note that since the invariant factors of $\lambda \operatorname{Id}_n - A$ (our matrix from way back when in this construction) were uniquely determined by the characteristic polynomial, so are the elementary divisors. Call these elementary divisors $p_1(\lambda)^k_1, \ldots, p_m(\lambda)^{k_m}$, and assume that none of them are units (note that since none of the invariant factors are zero, none of the elementary divisors are zero either). Then by the structure theorem, there exist elements $\underline{u}'_1, \ldots, \underline{u}'_m \in V$ such that $\mathrm{ann}_{F[\lambda]}(\underline{u}'_i) = (p_i^{k_i})$, and as $F[\lambda]$-modules

$$V \cong F[\lambda]\underline{u}'_1 \oplus \cdots \oplus F[\lambda]\underline{u}'_m$$

Again, this gives us that the action of T on V splits over this direct sum, giving a block diagonal matrix, and we study the structure of each of these blocks. In general, this is not particularly interesting, but if p_i is of the form $p_i(\lambda) = \lambda - x$, where $x \in F$, we do get an interesting result.

Proposition 5.5.7. *Let $A = (a_{ij}) \in M_n(F)$ be the matrix of T relative to the basis $\{\underline{u}_i\}_{i=1}^n$, and suppose that A only has one non-unit elementary factor, call it $p_1(\lambda)^{k_1}$ and suppose p_1 is of degree 1. Then $V \cong F[\lambda]\underline{u}'_1$ as $F[\lambda]$-modules, $k_1 = n$, $B = \{p_1(\lambda)^{k_1-1}\underline{u}'_1, p_1(\lambda)^{k_1-2}\underline{u}'_1, \ldots, \underline{u}'_1\}$*

is a basis for V as an F-vector space, and relative to this basis T has the matrix form

$$\begin{pmatrix} x_1 & 1 & 0 & 0 & \cdots 0 & 0 & 0 \\ 0 & x_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & x_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \ddots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & \cdots & \ddots & x_1 & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & x_1 \end{pmatrix}$$

where $p_1(\lambda) = \lambda - x_1$.

Proof. First, we show that B is spanning. Pick any $\underline{v} \in V$. Since $F[\lambda]\underline{u}'_1 = V$, there exists some polynomial $g(\lambda) \in F[\lambda]$ such that $g(\lambda)\underline{u}'_1 = \underline{v}$. Thus, it suffices to show that $p_1(\lambda)^{k_1}\underline{u}'_1 \in \mathrm{Span}_F(B)$. Indeed, we know that $p_1(\lambda)^{k_1}\underline{u}'_1 = \underline{0}$. Thus, setting $p_1(\lambda) = \lambda - x_1$, and $p_1(\lambda)^{k_1} = \lambda^{k_1} - \sum_{i=1}^{k_1} \alpha_i \lambda^{k_1 - i}$, we get

$$\lambda^{k_1}\underline{u}'_1 = \sum_{i=1}^{k_1} \alpha_i \lambda^{k_i - i}\underline{u}'_1$$

as required. Note that $n = k_1$ would then immediately imply that B is a basis. We cannot have $k_1 < n$ as the dimension of V is at most k_1. Suppose $k_1 > n$. Then B would be linearly dependent, so there would exist $b_i \in F$ such that $\sum_{i=1}^{k_1} b_i \lambda^{i-1}\underline{u}'_1 = \underline{0}$. But then a polynomial of degree $k_1 - 1$ would annihilate $\underline{u}'_1$, so $\mathrm{ann}_{F[\lambda]}(\underline{u}'_1) \neq (p_1^{k_1})$. Thus, we cannot have $k_1 > n$, so $k_1 = n$ as required. Finally, we consider the form of the matrix for T relative to the basis B. If $k < k_1 - 1$, we get

$$T((\lambda - x_1)^k \underline{u}'_1) = \lambda(\lambda - x_1)^k \underline{u}'_1 = (\lambda - x_1)^{k+1}\underline{u}'_1 + x_1(\lambda - x_1)^k \underline{u}'_1$$

If $k = k_1 - 1$, we instead get

$$T((\lambda - x_1)^k \underline{u}'_1) = \lambda(\lambda - x_1)^k \underline{u}'_1 = (\lambda - x_1)^{k_1}\underline{u}'_1 + x_1(\lambda - x_1)^{k_1 - 1}\underline{u}'_1 = x_1(\lambda - x_1)^{k_1 - 1}\underline{u}'_1$$

Thus, our matrix is of the form

$$\begin{pmatrix} x_1 & 1 & 0 & 0 & \cdots 0 & 0 & 0 \\ 0 & x_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & x_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \ddots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & \cdots & \ddots & x_1 & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & x_1 \end{pmatrix}$$

as claimed. □

Note. The characteristic polynomial of matrix representing a linear map is independent of the basis chosen. Thus, we can refer to the *characteristic polynomial* of the linear map T, which we denote $f_T(\lambda) \in F[\lambda]$.

We call matrices of the above form a *Jordan block* for the elementary factor $\lambda - x_1$. Again, we can bring all these observations together in the following theorem.

Theorem 5.5.8. *Suppose V is an F-vector space of finite dimension n, and $T \in \mathrm{End}_F(V)$ is a linear map whose characteristic polynomial factors into a product of the form $f_T(\lambda) = \prod_{i=1}^{m}(\lambda - x_i)^{k_i}$ in $F[\lambda]$. Then there exists a basis in which the matrix for T is in block diagonal form*

$$\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{pmatrix}$$

where each J_i is a Jordan block for some x_i.

Note. The above theorem allows, and in fact it is quite common, for one x_i to correspond to multiple Jordan blocks. It also requires that every x_i have at least one Jordan block, since each x_i appears at least once as an elementary factor (this follows from the structure theorem).

The trick to this, of course, is that it doesn't always work. To guarantee that it does, we can insist that F be *algebraically complete*, that is that every polynomial in $F[\lambda]$ of degree at least one factors into a product of linear factors in $F[\lambda]$ (or equivalently that every such polynomial has a root in $F[\lambda]$).

Putting a matrix into this form via a change of basis is called putting it into *Jordan canonical form*. We again get a uniqueness result out of this, assuming that F is algebraically complete.

Theorem 5.5.9. *Suppose $A, B \in M_n(F)$, where F is algebraically complete. Then they differ by a change of basis, i.e. there exists an invertible $Q \in M_n(F)$ such that $B = QAQ^{-1}$, if and only if they have the same Jordan canonical form (up to the order of Jordan blocks).*

There's more to talk about with Jordan canonical form, as it has a strong connection to the notion of eigenvalues and eigenspaces. To start off, we generalize these notions.

Definition 5.5.10. Let V be an F-vector space, $T \in \mathrm{End}_F(V)$. We call $\lambda \in F$ a generalized eigenvalue of T if there exists some non-zero $\underline{v} \in V$ and $n \in \mathbb{N}$ such that

$$(\lambda - T)^n \underline{v} = \underline{0}$$

In such a case, we call $\underline{v}$ a generalized eigenvector.

The thing about this definition is that, in a way, I haven't really defined much new. Indeed, we get the following.

Proposition 5.5.11. *Let V be an F-vector space, $T \in \mathrm{End}_F(V)$. Then $\lambda \in F$ a generalized eigenvalue of T if and only if it is an eigenvalue of T.*

Proof. That every eigenvalue is a generalized eigenvalue is immediate. Now, suppose that $\lambda \in F$ is a generalized eigenvalue. Let $n \in \mathbb{N}$ be the minimal number such that there exists non-zero $\underline{v} \in V$ such that $(\lambda - T)^n \underline{v} = \underline{0}$. If $n = 1$ then we're done. Otherwise, we'd get that $(\lambda - T)^{n-1}\underline{v} \neq \underline{0}$ and $(\lambda - T)(\lambda - T)^{n-1}\underline{v} = \underline{0}$, contradicting the minimality of n. Thus, $n = 1$, completing the proof. □

Note. Despite this result, generalized eigenvectors need not be eigenvectors.

Definition 5.5.12. Let V be an F-vector space, $T \in \mathrm{End}_F(V)$, and $\lambda \in F$ an eigenvalue. We denote by V_λ the subspace of V consisting of $\underline{0}$ and all the generalized eigenvectors of T with generalized eigenvalue λ.

I'll leave it to the reader to show that V_λ is indeed a subspace, again it is not too difficult to do. It's at this point though that we can start to connect things to Jordan canonical form. Indeed, notice that every vector in the basis given to us by Jordan canonical form consists entirely of generalized eigenvalues, and that those with different corresponding eigenvalues correspond to different blocks! We only need one more result than before our final conclusion.

Proposition 5.5.13. *Let V be an F-vector space, $T \in \mathrm{End}_F(V)$. Every generalized eigenvector of T has a unique corresponding generalized eigenvalue.*

Proof. Suppose $\lambda \in F$ is a generalized eigenvalue of T. It suffices to show that $(\mu - T)$ is invertible for any generalized eigenvalue $\mu \neq \lambda$ on V_μ. Indeed, suppose that weren't the case. Then there would exist some non-zero $\underline{u} \in V_\mu$ and $n \in \mathbb{N}$ such that $T\underline{u} = \mu\underline{u}$ and $(\lambda - T)^n \underline{u} = \underline{0}$. But $(\lambda - T)^n \underline{u} = (\lambda - \mu)^n \underline{u}$, so this implies that $\lambda = \mu$. □

Now, we bring everything together.

Theorem 5.5.14. *Let V be a finite-dimensional F-vector space, where F is algebraically complete, and $T \in \mathrm{End}_F(V)$ be a linear map with generalized eigenvalues $\lambda_1, \ldots, \lambda_n \in F$. Then $V \cong V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_n}$ as F-vector spaces, and by choosing the correct basis for each V_{λ_i} we can put the matrix representation of T into Jordan canonical form.*

Proof. Note that the x_i from the Jordan blocks in Theorem 5.5.8 are the roots of the characteristic polynomial of T, and hence precisely the eigenvalues of T (this should have been covered in an elementary linear algebra class). Hence, we can group the blocks and corresponding subspaces by eigenvalue to get the desired result. □

This approach is actually a much more intuitive way of thinking about Jordan canonical form, and is the one taken by [Sil23]. I recommend reading those notes, as it should hopefully help clear up Jordan canonical form.

All of these results can be quite intimidating at first, so it can be incredibly useful to work through some examples of both canonical forms. I will not do so in this text, as I would prefer to keep it free of long examples, but I'll provide some resources and tips for using them.

1. [Sil23] has some excellent practice problems for Jordan canonical form (and a few for Rational as well), approached in the manner I mentioned above.

2. [Jac09] has some mediocre to bad (and very difficult) practice problems/examples, following the approach taken here. However, Jacobson phrases everything in terms of row instead of column vector[2], so everything in there is what we've done with a transpose applied to the matrices.

3. [Rom07] seems to have some examples that take an approach in-between the previous two, and may be worth checking out. It also, however, takes a different notation and order of bases for things, causing its results to be slightly different.

4. Many computer algebra systems and tutorials exist for these topics online as well, and are likely just a Google away.

[2]This is terrible practice, don't do it.

Part III

The Abstract View

Chapter 6

Universal Algebras

6.1 Universal Algebras

Let us start this chapter with a construction which will seem rather abstract now, but becomes powerful in practice. Much of this section is based of the results in [BS12].

Definition 6.1.1. A universal algebra $\mathcal{A} = \langle \mathcal{U}, \mathcal{F} \rangle$ is a set (called the universe) $\mathcal{U}$ along with a family of operations $\mathcal{F}$ from $\mathcal{U}^n$ to $\mathcal{U}$ of finite arity.

Example 6.1.1. A group is a universal algebra $\langle G, \cdot, ^{-1}, 1 \rangle$ with 2-ary, 1-ary, and 0-ary operations respectively satisfying the following axioms, for all $x, y, z \in G$

1. $x \cdot (y \cdot z) = (x \cdot y) \cdot z$
2. $x \cdot 1 = 1 \cdot x = x$
3. $x \cdot x^{-1} = x^{-1} \cdot x = 1$

A group is called commutative (or Abelian) if the additional axiom $x \cdot y = y \cdot x$ is satisfied.

Example 6.1.2. A monoid is a group without the 1-ary operation.

Example 6.1.3. A ring is a universal algebra $\langle R, \cdot, 1, +, -, 0 \rangle$ such that

1. $\langle R, +, -, 0 \rangle$ is a commutative group
2. $\langle R, \cdot, 1 \rangle$ is a monoid
3. $\forall x, y, z \in R$,

$$x \cdot (y + z) = x \cdot y + x \cdot z \qquad (x + y) \cdot z = x \cdot z + y \cdot z$$

Example 6.1.4. A left monoid over a ring R is a universal algebra $\langle M, +, -, 0, \{f_r\}_{r \in R} \rangle$ such that

1. $\langle M, +, -, 0 \rangle$ is a commutative group

2. $\forall x, y \in M, r \in R,\ f_r(x+y) = f_r(x) + f_r(y)$

3. $\forall x \in M, r, s \in R,\ f_{r+s}(x) = f_r(x) + f_s(x)$ and $f_{r\cdot s}(x) = f_r(f_s(x))$

Definition 6.1.2. A sub-universe of a universal algebra $\mathcal{A} = \langle \mathcal{U}, \mathcal{F} \rangle$ is a subset $\mathcal{S} \subseteq \mathcal{U}$ such that $\langle \mathcal{S}, \mathcal{F}' \rangle$ is a universal algebra, where $\mathcal{F}' = \{f_{|\mathcal{S}} \mid f \in \mathcal{F}\}$. $\langle \mathcal{S}, \mathcal{F}' \rangle$ is a sub-algebra of $\mathcal{A}$.

Definition 6.1.3. A congruence $\equiv$ on a universal algebra $\mathcal{A} = \langle \mathcal{U}, \mathcal{F} \rangle$ is an equivalence relation on $\mathcal{U}$ such that for any n-ary operation $f \in \mathcal{F}$ and $a_i, a_i' \in \mathcal{U}$

$$a_i \equiv a_i' \Rightarrow f(a_1, \ldots, a_n) \equiv f(a_1', \ldots, a_n')$$

The notion of congruence allows us to define a particular type of universal algebra, which we call the quotient algebra.

Theorem 6.1.4. *Let $\mathcal{A} = \langle A, \mathcal{F} \rangle$ be a universal algebra and $\equiv$ a congruence on $\mathcal{A}$. Then ${}^{\mathcal{A}}/_{\equiv} = \langle {}^{A}/_{\equiv}, \mathcal{F} \rangle$ is a well-defined universal algebra, where an n-ary $f \in \mathcal{F}$ acts by, for any $[a_1], \ldots, [a_n] \in {}^{A}/_{\equiv}$*

$$f([a_1], \ldots, [a_n]) = [f(a_1, \ldots, a_n)]$$

Proof. It suffices to show that the action of each n-ary $f \in \mathcal{F}$ is well-defined. Indeed, suppose that $[a_i] = [b_i] \in {}^{A}/_{\equiv}$. Then since $\equiv$ is a congruence and $a_i \equiv b_i$,

$$f(a_1, \ldots, a_n) \equiv f(b_1, \ldots, b_n) \Rightarrow [f(a_1, \ldots, a_n)] = [f(b_1, \ldots, b_n)]$$

as required. □

This ${}^{\mathcal{A}}/_{\equiv}$ is called the *quotient* of $\mathcal{A}$ by $\equiv$. Finally, let $A' \subseteq A$ be an arbitrary subset. We define the *universal algebra generated by* A', denoted $\langle A' \rangle$ to be the intersection of all universes B containing A' such that $\langle B, \mathcal{F} \rangle$ is a universal algebra. This is well-defined as long as the intersection of universes with operations $\mathcal{F}$ is a universe, the verification of which is left to the reader.

At this point, we begin to play a bit fast and loose with definitions to avoid some more abstract concepts. We'll say that two universal algebras are of the same *type* if they have a corresponding set of n-ary operations, that is their n-ary operations can be put in bijective correspondence for each $n \in \mathbb{N} \cup \{0\}$. The most basic example of this would be to note that any sub-algebra is of the same type as its parent algebra. For a more precise notion than this, see [BS12].

Definition 6.1.5. Let $\mathcal{A} = \langle A, \mathcal{F} \rangle, \mathcal{B} = \langle B, \mathcal{F} \rangle$ be universal algebras of the same type. A map $\varphi : A \to B$ is called a homomorphism if, for any n-ary operation $f \in \mathcal{F}$ and $a_i \in A$

$$f(\varphi(a_1), \ldots, \varphi(a_n)) = \varphi(f(a_1, \ldots, a_n))$$

Note. There's a bit of an abuse of notation here, as f is technically two different operations on A and B. We regard it as acting on both by the pairing between n-ary operations of the two universal algebras. This notation will be used for the remainder of this section.

Note. A nice property of homomorphisms, the proof of which is left to the reader, is that the composition of two homomorphisms is a homomorphism.

An injective homomorphism is called a *monomorphism*, a surjective homomorphism an *epimorphism*, and a bijective homomorphism an *isomorphism*. An isomorphism from a universe to itself is called an *automorphism*, and two universal algebras are called *isomorphic* if there exists an isomorphism between them, which is denoted by the symbol $\cong$. Homomorphism are of vital importance of algebra, as their "preservation" of operations allows us to determine when two universal algebras are essentially identical. In order to make this more precise, we need the following two lemmas.

Lemma 6.1.6. *Suppose $\mathcal{A}, \mathcal{B}$ are universal algebras of the same type, and $\varphi : A \to B$ a homomorphism. Then $\varphi(\mathcal{A}) = \langle \varphi(A), \mathcal{F} \rangle$ is a sub-algebra of $\mathcal{B}$, and $\varphi^{-1}(\mathcal{B}) = \langle \varphi^{-1}(B), \mathcal{F} \rangle$ is a sub-algebra of $\mathcal{A}$.*

Proof. Let $f \in \mathcal{F}$ be an n-ary operation. If $b_1, \ldots, b_n \in \varphi(A)$, then $\exists a_i \in A$ such that $\varphi(a_i) = b_i$. Thus,

$$f(b_1, \ldots, b_n) = f(\varphi(a_1), \ldots, \varphi(a_n)) = \varphi(f(a_1, \ldots, a_n)) \in \varphi(A)$$

so $\varphi(\mathcal{B})$ is a sub-algebra, as claimed. The proof of the second statement is essentially identical. □

Lemma 6.1.7. *Suppose $\mathcal{A}, \mathcal{B}$ are universal algebras of the same type, and $\varphi : A \to B$ a homomorphism. Then the equivalence relation $\equiv$ on A given by $a_1 \equiv a_2$ if $\varphi(a_1) = \varphi(a_2)$ is a congruence.*

Proof. Suppose $f \in \mathcal{F}$ is an n-ary operation, and $a_i \equiv a_i' \in A$. Then

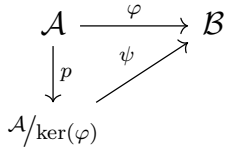
$$\varphi(f(a_1, \ldots, a_n)) = f(\varphi(a_1), \ldots, \varphi(a_n)) = f(\varphi(a_1'), \ldots, \varphi(a_n')) = \varphi(f(a_1', \ldots, a_n'))$$

so $f(a_1, \ldots, a_n) \equiv f(a_1', \ldots, a_n')$, as required. □

In the case of the above lemma, we denote that ${}^{\mathcal{A}}\!/_{\equiv} = {}^{\mathcal{A}}\!/_{\ker(\varphi)}$, and ${}^{A}\!/_{\equiv} = {}^{A}\!/_{\ker(\varphi)}$.

With these, we finally reach the fundamental theorems of homomorphisms.

Theorem 6.1.8 (The First Fundamental Theorem of Homomorphisms)**.** *Suppose $\mathcal{A}, \mathcal{B}$ are universal algebras of the same type, and $\varphi : A \to B$ a homomorphism. Then the projection map $p : A \to {}^{A}\!/_{\ker(A)}$ is a homomorphism, and there exists an isomorphism $\psi : {}^{\mathcal{A}}\!/_{\ker(\varphi)} \to \varphi(\mathcal{A})$ such that the following diagram commutes*



In particular, ${}^{\mathcal{A}}\!/_{\ker(\varphi)} \cong \varphi(\mathcal{B})$.

Proof. The first statement follows from the projection map onto equivalence classes for a congruence always being a homomorphism, which is easily verified and left to the reader. We define $\psi : {}^{A}\!/_{\ker(A)} \to \varphi(A)$ by, for any $[a] \in {}^{A}\!/_{\ker(A)}$, $\psi([a]) = \psi(a)$. It remains to show that this is in fact well-defined, a homomorphism, bijective, and satisfies the commutative property. For the first, we note that if $[a] = [a'] \in {}^{A}\!/_{\ker(A)}$, then $\varphi(a) = \varphi(a')$, making ψ well-defined. Pick any n-ary $f \in \mathcal{F}$, and $[a_1], \ldots, [a_n] \in {}^{A}\!/_{\ker(A)}$. Then

$$f(\psi([a_1]), \ldots, \psi([a_n])) = f(\varphi(a_1), \ldots, \varphi(a_n)) = \varphi(f(a_1, \ldots, a_n)) = \varphi(f([a_1], \ldots, [a_n]))$$

so ψ is a homomorphism. Pick any $b \in \varphi(B)$. Then $\exists a \in A$ such that $\varphi(a) = b$, so $\psi([a]) = b$ and hence ψ is surjective. Suppose $\psi([a_1]) = \psi([a_2])$, where $[a_1], [a_2] \in {}^{A}\!/_{\ker(A)}$ Then $\varphi(a_1) = \varphi(a_2) \Rightarrow [a_1] = [a_2]$, so ψ is injective and hence bijective. Finally, we check the commutativity property. Let $a \in A$, then by definition

$$\varphi(a) = \psi([a]) = (\psi \circ p)(a)$$

as claimed. □

For the next two isomorphism theorems, we need the notion of a sub-congruence.

Definition 6.1.9. Let $\mathcal{A}$ be a universal algebra, and $\equiv$ a congruence on $\mathcal{A}$. A congruence $\sim$ on $\mathcal{A}$ is called a sub-congruence of $\equiv$ if $a \equiv a' \Rightarrow a \sim a'$, for all $a, a' \in A$.

Lemma 6.1.10. *Let $\equiv$ be a congruence on a universal algebra $\mathcal{A}$, and $\sim$ a sub-congruence of $\equiv$. Let ${}^{\equiv}\!/_{\sim}$ be the equivalence relation on ${}^{\mathcal{A}}\!/_{\sim}$ defined by*

$$[a]_{\sim} {}^{\equiv}\!/_{\sim} [a']_{\sim} \iff a \equiv a'$$

This is a well-defined equivalence relation, and a congruence on ${}^{\mathcal{A}}\!/_{\sim}$

Proof. We start by show that it is an equivalence relation. Reflexivity and symmetry are clear, so it suffices to show transitivity. Suppose that $[a], [a'], [a''] \in {}^{\mathcal{A}}\!/_{\sim}$ are such that $[a] {}^{\equiv}\!/_{\sim} [a']$ and $[a'] {}^{\equiv}\!/_{\sim} [a'']$. Then $a \equiv a'$ and $a' \equiv a''$, so $a \equiv a''$ and hence $[a] {}^{\equiv}\!/_{\sim} [a'']$, as required. Now, let $f \in \mathcal{F}$ be an n-ary operation. Suppose $[a_i] {}^{\equiv}\!/_{\sim} [b_i] \in {}^{\mathcal{A}}\!/_{\sim}$. Then $a_i \equiv b_i$, so

$$f([a_1], \ldots, [a_n]) = [f(a_1, \ldots, a_n)] {}^{\equiv}\!/_{\sim} [f(b_1, \ldots, b_n)] = f([b_1], \ldots, [b_n])$$

making ${}^{\equiv}\!/_{\sim}$ a congruence as claimed. □

Theorem 6.1.11 (The Second Fundamental Theorem of Homomorphisms)**.** *Let $\equiv$ be a congruence on a universal algebra $\mathcal{A}$, and $\sim$ a sub-congruence of $\equiv$. Then there exists an isomorphism $\varphi : {}^{({}^{\mathcal{A}}\!/_{\sim})}\!/_{({}^{\equiv}\!/_{\sim})} \to {}^{\mathcal{A}}\!/_{\equiv}$ such that the following diagram commutes.*

$$\begin{array}{ccc} \mathcal{A} & \longrightarrow & {}^{\mathcal{A}}\!/_{\equiv} \\ \downarrow & & \uparrow{\scriptstyle \varphi} \\ {}^{\mathcal{A}}\!/_{\sim} & \longrightarrow & {}^{({}^{\mathcal{A}}\!/_{\sim})}\!/_{({}^{\equiv}\!/_{\sim})} \end{array}$$

where all the unlabelled arrows are the natural projection maps. In particular, $\mathcal{A}/\equiv \cong (\mathcal{A}/\sim)/(\equiv/\sim)$.

Proof. We define φ by the rule, for any $a \in A$, $\varphi([[a]_\sim]_{\equiv/\sim}) = [a]_\equiv$. We first show that it is a homomorphism. Pick any n-ary $f \in \mathcal{F}$, $a_i \in A$. Then

$$\begin{aligned} f(\varphi([[a_1]_\sim]_{\equiv/\sim}), \ldots, \varphi([[a_n]_\sim]_{\equiv/\sim})) &= f([a_1]_\equiv, \ldots, [a_n]_\equiv) = [f(a_1, \ldots, a_n)]_\equiv \\ &= \varphi([[f(a_1, \ldots, a_n)]_\sim]_{\equiv/\sim}) = \varphi(f([[a_1]_\sim]_{\equiv/\sim}, \ldots, [[a_n]_\sim]_{\equiv/\sim})) \end{aligned}$$

as required. Next, we show that it's injective. Suppose that $\varphi([[a]_\sim]_{\equiv/\sim}) = \varphi([[a']_\sim]_{\equiv/\sim})$ for some $a, a' \in A$. Then $[a]_\equiv = [a']_\equiv \Rightarrow a \equiv a' \Rightarrow [a]_\sim {}^{\equiv}\!/_\sim [a']_\sim \Rightarrow [[a]_\sim]_{\equiv/\sim} = [[a']_\sim]_{\equiv/\sim}$ as required. Surjectivity is clear, so φ is an isomorphism. Finally, we show that the commutativity property is satisfied. Pick any $a \in A$. Then

$$(\varphi \circ p_{\equiv/\sim} \circ p_\sim)(a) = (\varphi \circ p_{\equiv/\sim})([a]_\sim) = \varphi([[a]_\sim]_{\equiv/\sim}) = [a]_\equiv = p_\equiv(a)$$

as required. □

Theorem 6.1.12 (The Third Fundamental Theorem of Homomorphisms)**.** *Suppose* $\mathcal{B}$ *is a subalgebra of* $\mathcal{A}$*, and* $\equiv$ *is a congruence on* $\mathcal{A}$*. Set* $B^\equiv = \{a \in A \mid B \cap [a]_\equiv \neq \emptyset\}$*. Then*

1. $\equiv_{|_B}$ *is a congruence*
2. $\mathcal{B}^\equiv$ *is a subalgebra of* $\mathcal{A}$
3. $\mathcal{B}/\equiv_{|_B} \cong \mathcal{B}^\equiv/\equiv_{|_{B^\equiv}}$

Proof. (1) is fairly clear, and its proof will be left to the reader. For (2), pick any n-ary $f \in \mathcal{F}$ and $a_1, \ldots, a_n \in B^\equiv$. It suffices to show that $f(a_1, \ldots, a_n) \in B^\equiv$. Indeed, we can note that for each a_i, there exists some $b_i \in B$ such that $a_i \equiv b_i$, and hence

$$[f(a_1, \ldots, a_n)]_\equiv = f([a_1]_\equiv, \ldots, [a_n]_\equiv) = f([b_1]_\equiv, \ldots, [b_n]_\equiv) = [f(b_1, \ldots, b_n)]_\equiv$$

Since $\mathcal{B}$ is a subalgebra, $f(b_1, \ldots, b_n) \in B$, so it follows that $[f(a_1, \ldots, a_n)]_\equiv \cap B \neq \emptyset$ and hence $f(a_1, \ldots, a_n) \in B^\equiv$ as required. It is also good to note that this further shows that $\mathcal{B}$ is a subalgebra of $\mathcal{B}^\equiv$. For (3), we first note that (1) implies that $\equiv_{|_{B^\equiv}}$ is a congruence, so this statement is well-defined. We define our isomorphism $\varphi : B/\equiv_{|_B} \to B^\equiv/\equiv_{|_{B^\equiv}}$ by, for any $b \in B$, $\varphi([b]_{\equiv_{|_B}}) = [b]_{\equiv_{|_{B^\equiv}}}$. We first show that this is a homomorphism. Pick any n-ary $f \in \mathcal{F}$ and $b_1, \ldots, b_n \in B$. Then

$$\begin{aligned} f(\varphi([b_1]_{\equiv_{|_B}}), \ldots, \varphi([b_n]_{\equiv_{|_B}})) &= f([b_1]_{\equiv_{|_{B^\equiv}}}, \ldots, [b_n]_{\equiv_{|_{B^\equiv}}}) = [f(b_1, \ldots, b_n)]_{\equiv_{|_{B^\equiv}}} \\ &= \varphi([f(b_1, \ldots, b_n)]_{\equiv_{|_B}}) = \varphi(f([b_1]_{\equiv_{|_B}}, \ldots, [b_n]_{\equiv_{|_B}})) \end{aligned}$$

as required. Surjectivity is clear. To show that φ is injective, suppose that $b, b' \in B$ are such that $\varphi([b]_{B_{|_\equiv}}) = \varphi([b']_{B_{|_\equiv}})$. Then $[b]_{\equiv_{|_{B^\equiv}}} = [b']_{\equiv_{|_{B^\equiv}}} \Rightarrow [b]_{B_{|_\equiv}} = [b']_{B_{|_\equiv}}$ as required. □

These theorems appear in different guises for groups in section 2.4, rings in section 3.4, and modules in section 4.1. It's worth taking the time now to go back through and look at each of these chapters, and figure out for yourself how the homomorphism theorems for universal algebras connect to the corresponding theorems in each chapter.

6.2 Direct Products

This section is based on results from [BS12] and [Sil23]. Some of it may seem familiar to you from the section on direct sums/products of modules, and if so that's good! These ideas are a generalization of those constructions.

In this section, we cover the one of the most common constructions on spaces in algebra, the direct product. Before doing this, we have a quick bit of notation : we denote the collection of all homomorphisms between universal algebras $\mathcal{A}, \mathcal{B}$ by $\mathrm{Hom}(\mathcal{A}, \mathcal{B})$.

Definition 6.2.1. Let $\{\mathcal{A}_i\}_{i\in I}$ be an indexed family of universal algebras of the same type. We call another universal algebra of the same type $\mathcal{B}$ a direct product of $\{\mathcal{A}_i\}_{i\in I}$ if there exist homomorphisms $\pi_i : B \to A_i$ such that for any other universal algebra of the same type $\mathcal{C}$ and homomorphisms $\{\psi_i : C \to A_i\}_{i\in I}$, there exists a unique homomorphism $\varphi \in \mathrm{Hom}(\mathcal{C}, \mathcal{B})$ such that $\pi_i \circ \varphi = \psi_i$ for all $i \in I$.

This of course is quite an abstract definition. Luckily, for universal algebras we have the following result to make things more tangible.

Theorem 6.2.2. *For any indexed family $\{\mathcal{A}_i\}_{i\in I}$ of universal algebras of the same type, there exists a direct product unique up to a unique isomorphism.*

Proof. We start with uniqueness. Suppose $\mathcal{B}, \mathcal{C}$ are direct products of $\{\mathcal{A}_i\}_{i\in I}$. Then there exists a unique homomorphism $\varphi \in \mathrm{Hom}(\mathcal{C}, \mathcal{B})$ such that $\pi_{i,B} \circ \varphi = \pi_{i,C}$, and there exists a unique homomorphism $\zeta \in \mathrm{Hom}(\mathcal{B}, \mathcal{C})$ such that $\pi_{i,C} \circ \zeta = \pi_{i,B}$. It follows that $\pi_{i,B} \circ \varphi \circ \zeta = \pi_{i,B}$, so $\varphi \circ \zeta = \mathrm{Id}_{\mathcal{C}}$. Similarly, $\zeta \circ \varphi = \mathrm{Id}_{\mathcal{B}}$, so φ is an isomorphism. By the uniqueness of φ, ζ, it is the unique isomorphism between $\mathcal{C}, \mathcal{B}$, as claimed.

Next, we show existence. We define the universal algebra $\prod_{i\in I} \mathcal{A}_i$ as having the universe $\prod_{i\in I} A_i$, and operations defined by, for any $f \in \mathcal{F}$

$$f((a_{i1})_{i\in I}, \ldots, (a_{in})_{i\in I}) = (f(a_{i1}, \ldots, a_{in}))_{i\in I}$$

where $a_{ij} \in \mathcal{A}_i$. That this is a universal algebra is clear, and since projection maps between universal algebras are homomorphisms, we have our required homomorphisms $\pi_j \in \mathrm{Hom}(\prod_{i\in I} \mathcal{A}_i, \mathcal{A}_j)$. It suffices then to show that these satisfy the desired property. Suppose $\mathcal{C}$ is another universal algebra of the same type, and $\psi_j \in \mathrm{Hom}(\mathcal{C}, \mathcal{A}_j)$ homomorphisms. If we try to define $\varphi : C \to B$ by $\pi_i \circ \varphi = \psi_i$, we can note that this requires $\varphi(c) = \prod_{i\in I} \psi_i(c)$ for any $c \in C$, giving the uniqueness of φ. It suffices then to prove that φ is a homomorphism. Pick any $f \in \mathcal{F}$ and $c_i \in C$. Then

$$\begin{aligned} f(\varphi(c_1), \ldots, \varphi(c_n)) &= f\Big(\prod_{i\in I} \psi_i(c_1), \ldots, \prod_{i\in I} \psi_i(c_n)\Big) = \prod_{i\in I} f(\psi_i(c_1), \ldots, \psi_i(c_n)) \\ &= \prod_{i\in I} \psi_i(f(c_1, \ldots, c_n)) = \varphi(f(c_1, \ldots, c_n)) \end{aligned}$$

as required. □

Note. It is not too difficult to show, using the above theorem, that the direct product is associative and commutative up to isomorphism.

There's another important construction to mention here, called the *direct sum*.

Definition 6.2.3. Let $\{\mathcal{A}_i\}_{i\in I}$ be an indexed family of universal algebras of the same type. We call another universal algebra of the same type $\mathcal{B}$ a direct sum of $\{\mathcal{A}_i\}_{i\in I}$ if there exist homomorphisms $\iota_i : A_i \to B$ such that any other universal algebra of the same type $\mathcal{C}$ and homomorphisms $\{\psi_i : A_i \to C\}_{i\in I}$, there exists a unique homomorphism $\varphi \in \mathrm{Hom}(\mathcal{B}, \mathcal{C})$ such that $\varphi \circ \iota_i = \psi_i$ for all $i \in I$.

This is essentially just the definition of the direct product with the arrows reversed. Similarly to the direct product, one could show that the direct sum is unique up to unique isomorphism when it exists. Existence is a much trickier proposition, as the structure of a direct sum can get exceedingly complicated[1]. As such, we will treat direct sums in the subsequent chapters on a case-by-case basis. A more unified description of direct sums and when they exist is possible though category theory, which we cover next.

[1] I suspect that it's not guaranteed for all universal algebras, but I have not yet thought of a counterexample

Chapter 7

Categories

Mathematics, like physics, has a bit of an obsession with unifying itself under one "theory of everything". Of course, this takes a much different form in mathematics, a subject where what unifying means isn't as clear[1]. For our purposes, unifying is the process of not only formulating a basic logical system/set of axioms for mathematics, but formulating one which can actually be used directly in modern research and provides a framework for comparing similar mathematical objects. To my knowledge, no single concept has been more successful in this endeavour than *category theory*, which we now explore. Everything in this section will be based on similar ones from [Lan10], with other references mentioned when relevant.

It is also quite possible, in fact likely, that the contents of this chapter will seem rather impenetrable to you right now. This is perfectly okay! It's hard to jump to this level of abstraction, and the hope is that chapters 8 and 9 will ease you into a lot of these concepts before chapter 10 jumps in the deep end.

7.1 Basic Definitions

Definition 7.1.1. A category $\mathcal{C} = (\mathcal{O}, \mathcal{A}, \text{dom}, \text{codom}, \circ)$ is a collection of objects $\mathcal{O}$, arrows $\mathcal{A}$, two maps $\text{dom}, \text{codom} : \mathcal{A} \to \mathcal{O}$, and one partially defined binary "composition" map $\circ : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ satisfying the following axioms

1. If $f, g \in \mathcal{A}$, then $f \circ g$ is defined if and only if $\text{codom}(g) = \text{dom}(f)$, and in this case $\text{dom}(f \circ g) = \text{dom}(g), \text{codom}(f \circ g) = \text{codom}(f)$.
2. For each $a \in \mathcal{O}$, there exists a map $\text{Id}_a \in \mathcal{A}$ such that for all $f, g \in \mathcal{A}$ with $\text{dom}(f) = \text{codom}(g) = a$
$$f \circ \text{Id}_a = f \qquad\qquad \text{Id}_a \circ g = g$$
3. Composition, when defined, is associative.

There's a ton to unpack here, so let's start with some examples.

[1] But has proven far more fruitful than string theory.

Example 7.1.1. The category **Set** has the collection of all sets as its objects and maps between sets as its arrows, with standard function composition.

Example 7.1.2. The category **Grp** has the collection of all groups as its objects and the collection of all homomorphisms between groups as its arrows, with standard function composition.

Note. There's an important distinction to be made in both of the above examples about what are and are not identical objects. Namely, two sets can be the same size, but unless we explicitly identify all of their objects they are not the same set. For example, $\{a, b, c\} \neq \{a, b, d\}$. With groups it's similar, we can have the same group structure on two different sets, but unless those underlying sets have the same objects we consider the groups to be different. They are, however, isomorphic (we'll define that shortly).

Example 7.1.3. A monoid is a (small) category with one object.

You may have noticed in these examples that the word collection is doing some heavy lifting. Indeed, it is clear from the first that collections cannot just include sets, as the collection of all sets is itself not a set. The solution to this, or at least one of them, is to allow our collections to be what are called *classes*. These originate from the Gödel-Bernays axioms of set theory, and for our purposes can be thought of as an extension of the idea of sets. That is, every set is a class, but not every class is a set. We call a class *proper* if it is not a set. A category is called *small* if its collections of objects and arrows are sets, and *large* otherwise.

We will not delve too deeply into foundational issues here, (a slightly more thorough examination of the topic can be found in [Lan10], and a full view on the topic belongs in a course on logic) and will instead begin to familiarize ourselves with the myriad of terminologies used in category theory.

1. The category $\mathcal{C} = (\mathcal{O}, \mathcal{A}, \text{dom}, \text{codom}, \circ)$ is usually just denoted by $\mathcal{C}$. In this notation, we denote $\mathcal{O}$ by $\text{ob}(\mathcal{C})$ and $\mathcal{A}$ by $\text{hom}(\mathcal{C})$.

2. The arrows of a category are often called *morphisms*, hence the notation $\text{hom}(\mathcal{C})$. We usually denote the set of all arrows between two objects $a, b \in \text{ob}(\mathcal{C})$ by $\text{hom}(a, b)$, and show that $f \in \text{hom}(a, b)$ by writing $f : a \to b$.

3. A morphism $f : a \to b$ is called

   (a) A *monomorphism* if, for any two morphisms $g, h \in \text{hom}(c, a)$, we get $f \circ g = f \circ h \Rightarrow g = h$.

   (b) An *epimorphism* if, for any two morphisms $g, h \in \text{hom}(b, c)$, we get $g \circ f = h \circ f \Rightarrow g = h$.

   (c) An *isomorphism* (or invertible) if there exists $g : b \to a$ such that $f \circ g = \text{Id}_b, g \circ f = \text{Id}_a$. Such a g is called an inverse.

   (d) An *endomorphism* if $a = b$. The set of all such endomorphisms is denoted $\text{end}(a)$.

   (e) An *automorphism* if it is an isomorphism and endomorphism. The set of all such automorphisms is denoted $\text{aut}(a)$.

4. Two objects $a, b \in \text{ob}(\mathcal{C})$ are called isomorphic, denoted $a \cong b$, if there exists an isomorphism between them.

Let's take a break at this point to do a short proof.

Proposition 7.1.2. *Every isomorphism $f : a \to b$ has a unique inverse $g : b \to a$. Furthermore, every isomorphism is a mono and epimorphism.*

Proof. We begin with the first statement. Suppose $g, h : b \to a$ were two inverses. Then we get

$$g = g \circ \text{Id}_b = g \circ (f \circ h) = (g \circ f) \circ h = \text{Id}_a \circ h = h$$

We denote this unique inverse by f^{-1}. Now for the second statement. Suppose $g, h \in \hom(c, a)$ and $f \circ g = f \circ h$. Then composing with f^{-1} on the left we get $g = h$, making f monomorphic. The argument for epimorphic is essentially the same. □

It's important to note here that epi and monomorphisms may not behave exactly as you'd expect. Specifically, they need not have left (right) inverses. We give such morphisms a different name.

Definition 7.1.3. A morphism is split monic (i.e. a split monomorphism) if it has a left inverse. We call such a left inverse a retraction of the morphism. Similarly, a morphism is split epi (i.e. a split epimorphism) if it has a right inverse, and we call such a right inverse a section of the morphism.

Note. We do not require that our sections or retraction be unique, and indeed they may not be.

It should be clear from the definition that split monomorphisms are necessarily monomorphisms, and split epimorphisms necessarily epimorphisms, although the converse does not hold in general.

There is one final piece of terminology for morphisms we'll need before moving on to objects.

Definition 7.1.4. A morphism $f : a \to a$ is called idempotent if $f^2 = \text{Id}_a$. It is called split idempotent if there exist morphisms g, h such that $f = g \circ h, h \circ g = \text{Id}$, where we do not specify here the object that $h \circ g$ is the identity on.

Finally, we can start looking at terminology for objects in categories. An object $a \in \text{ob}(\mathcal{C})$ is called

1. Terminal if for each $b \in \text{ob}(\mathcal{C})$ there exists exactly one morphism in $\hom(b, a)$.
2. Initial if for each $b \in \text{ob}(\mathcal{C})$ there exists exactly one morphism in $\hom(a, b)$.
3. Null if it is initial and terminal.

There's a couple of things to notice here about the above definitions.

1. If $a \in \text{ob}(\mathcal{C})$ is initial or terminal, then $\hom(a, a) = \{\text{Id}_a\}$.

2. If $a \in \text{ob}(\mathcal{C})$ is null, then for each $b, c \in \text{ob}(\mathcal{C})$ we can find unique $g \in \hom(b, a), h \in \hom(a, c)$. We call $h \circ g \in \hom(b, c)$ the *zero arrow* from b to c. Note that the composition of two zero arrows is itself a zero arrow.

Before moving on to talking about functors, I would like to mention one more important type of category.

Definition 7.1.5. A groupoid is a category in which all morphisms are invertible.

Note. A group is just a small groupoid with a single object.

We will not talk about groupoids in this text, but I felt them worth mentioning as they're of utmost importance in algebraic topology, which originated category theory.

I've given you a lot of information in this section, but I've so far omitted a pretty important concept. Namely, I've given you no way to compare categories. Let's fix that.

Definition 7.1.6. Let $\mathcal{C}, \mathcal{D}$ be categories. A (covariant) functor $f : \mathcal{C} \to \mathcal{D}$ is a pair of maps $f_o : \text{ob}(\mathcal{C}) \to \text{ob}(\mathcal{D}), f_a : \hom(\mathcal{C}) \to \hom(\mathcal{D})$ satisfying the following axioms.

1. If $\varphi \in \hom(a, b)$, then $f_a(\varphi) \in \hom(f_o(a), f_o(b))$.
2. If $a \in \text{ob}(\mathcal{C})$, then $f_a(\text{Id}_a) = \text{Id}_{f_o(a)}$.
3. If $\varphi \in \hom(a, b)$ and $\psi \in \hom(b, c)$, then $f_a(\psi \circ \varphi) = f_a(\psi) \circ f_a(\varphi)$.

Note. We normally denote both f_o and f_a by f, as the particular map being used is clear from context.

This is where I'm going to make a big leap, so don't be afraid to take time with this. We can construct a category **Cat**, whose objects consist of all (small[2]) categories and morphisms consist of all functors (the composition being standard function composition applied to the object and arrow functions). This allows us to bring over all our terminology for morphisms and apply it to functors, in particular giving us the notion of isomorphic categories.

There are a couple more terms we use when describing functors. We call a functor $f : \mathcal{C} \to \mathcal{D}$

1. Faithful if for any pair of objects $a, b \in \mathcal{C}$, $f_a : \hom(a, b) \to \hom(f_o(a), f_o(b))$ is injective.
2. Full if for any pair of objects $a, b \in \mathcal{C}$, $f_a : \hom(a, b) \to \hom(f_o(a), f_o(b))$ is surjective.
3. Fully faithful if it is full and faithful.
4. An endofunctor if $\mathcal{C} = \mathcal{D}$.

Note. Faithful functors need not be monomorphisms, full functors need not be epimorphisms, and fully faithful functors need not be isomorphisms.

[2]Proof that this caviot is required can be found here [rus99], and my thanks to Oakley Edens for pointing this out to me.

Functors also allow us to work a bit more with sub-categories (which are defined in the exact way you'd expect). These come with an inclusion functor, and we call a sub-category *full* if that inclusion functor is full.

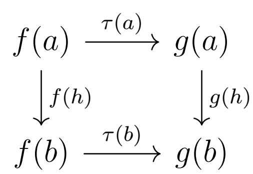
I'll also mention a classic example of a functor here, the *forgetful functor*. Generally, this is a functor that takes a category with a "richer" structure to one with a "weaker" structure. Let's look at some examples.

Example 7.1.4. The forgetful functor $f : \mathbf{Grp} \to \mathbf{Set}$, which takes groups to their underlying sets and homomorphisms to their underlying set maps.

Example 7.1.5. The forgetful functor $f : \mathbf{Rng} \to \mathbf{AbGrp}$ (rings to Abelian groups), which forgets the multiplication structure on the ring and preserves the addition structure.

Finally, we wish to find a way to define morphisms between functors. This comes from the following definition.

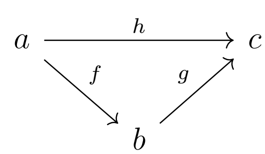
Definition 7.1.7. Take two functors $f, g : \mathcal{C} \to \mathcal{D}$. A natural transformation $\tau : \text{ob}(\mathcal{C}) \to \text{hom}(\mathcal{D})$ from f to g is a map such that for any $a, b \in \text{ob}(\mathcal{C})$ and $h \in \text{hom}(a, b)$, the following diagram commutes.



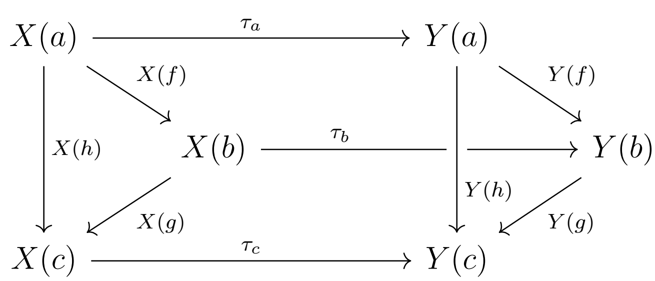
We call $\tau(a)$ a component of the natural transformation, and say that $\tau(a)$ is natural in a.

It will become clear in practice that the above notation is not ideal for natural transformations. We usually instead regard τ as a collection of maps $\{\tau_a = \tau(a) : f(a) \to g(a)\}_{a \in \text{ob}(\mathcal{C})}$ indexed by the objects of $\mathcal{C}$, and denote that τ is a natural transformation from f to g by $\tau : f \dot{\to} g$.

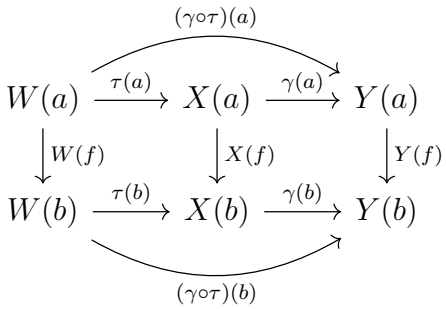
What natural transformations are really doing is taking commutative diagrams to commutative diagrams. Say, for example, that a, b, c are objects and $f : a \to b, g : b \to c, h : a \to c$ morphisms such that the following diagram commutes.



Then if X, Y are two functors and τ a natural transformation between them, we get the following commutative diagram.



In this way, natural transformations start to look like morphisms between functors. Indeed, fix four functors $W, X, Y : \mathcal{C} \to \mathcal{D}$, and let τ, γ be natural transformations from W to X, X to Y, and Y to Z. We can define $\gamma \circ \tau$ by $(\gamma \circ \tau)(a) = \gamma(a) \circ \tau(a)$. We first check that this is indeed a natural transformation from W to Y. To do this, we note that the following diagram commutes



It is not too hard to check that this "composition" is associative, and that for each functor X we have an "identity" natural transformation given by $\mathrm{Id}_X(a) = \mathrm{Id}_{X(a)}$. Thus, we can form the category $\mathrm{Func}(\mathcal{C}, \mathcal{D})$, the category of all functors between these two categories with natural transformations as morphisms. If $X, Y : \mathcal{C} \to \mathcal{D}$ are two functors, we will often denote the set of all natural transformations from X to Y by $\mathrm{Nat}(X, Y)$ rather than $\hom(X, Y)$.

Again, this further allows us to transfer over all our terminology for morphisms to natural transformations, in particular defining an isomorphic natural transformation (usually called a natural isomorphism) and hence a notion of equivalent functors. This notion of equivalent functors, strangely, also turns out to be quite important to the study of objects. Indeed, there's a sense in which our notion of isomorphic categories is a little too strong. To fix this, we weaken our notion of "equivalent" categories, and say that two categories are equivalent if there exist functors $F : \mathcal{C} \to \mathcal{D}, G : \mathcal{D} \to \mathcal{C}$ such that $G \circ F \cong \mathrm{Id}_{\mathcal{C}}, F \circ G \cong \mathrm{Id}_{\mathcal{D}}$[3].

7.2 Dual and Product Categories

Let $\mathcal{C}$ be a category. It's *opposite/dual category*, denoted $\mathcal{C}^*$, is the category with the same collection of objects and arrows as $\mathcal{C}$, but with the dom and codom functions switched. If $f \in \hom(a, b)$, we denote the same arrow in $\mathcal{C}^*$ by $f^* \in \hom(b, a)$. Composition is defined by, if $f^* \in \hom(a, b), g^* \in \hom(b, c)$

$$g^* \circ f^* = (f \circ g)^*$$

It is not too difficult to show that this composition is associative, and that Id_a^* is an identity for a under this composition, thus making the opposite category, in fact, a category.

Now, consider the functor-like map $\varphi : \mathcal{C} \to \mathcal{D}$ given by $\varphi(a) = a, \varphi(f) = f^*$. This satisfies all the axioms of a functor, except one. Namely, we get

$$\varphi(g \circ f) = \varphi(f) \circ \varphi(g)$$

We have a specific name for maps of this type.

[3]Yes, this does look like a homotopy. I'm not expanding on that here.

Definition 7.2.1. Let $\mathcal{C}, \mathcal{D}$ be categories. A contravariant functor $f : \mathcal{C} \to \mathcal{D}$ is a pair of maps $f_o : \text{ob}(\mathcal{C}) \to \text{ob}(\mathcal{D}), f_a : \hom(\mathcal{C}) \to \hom(\mathcal{D})$ satisfying the following axioms.

1. If $\varphi \in \hom(a, b)$, then $f_a(\varphi) \in \hom(f_o(b), f_o(a))$.
2. If $a \in \text{ob}(\mathcal{C})$, then $f_a(\text{Id}_a) = \text{Id}_{f_o(a)}$.
3. If $\varphi \in \hom(a, b)$ and $\psi \in \hom(b, c)$, then $f_a(\psi \circ \varphi) = f_a(\varphi) \circ f_a(\psi)$.

Like with covariant functors (which we usually just call functors), we drop the subscripts when working with contravariant functors and assume that the particular map being used is clear from context. Our terminology of faithful, full, fully faithful, and endofunctors also carries over to contravariant functors.

Before continuing, let's prove two quick statements about opposite categories.

Proposition 7.2.2. $\mathcal{C}^{**} \cong \mathcal{C}$ *in the category* **Cat**.

Proof. This follows immediately by noting that the functor-like maps above are contravariant functors, that the map $\varphi(a) = a, \varphi(f) = f^{**}$ is the composition of these maps, and that the composition of two contravariant functors is a covariant functor. The inverse of this covariant functor is clear. □

Proposition 7.2.3. *Suppose $\varphi : \mathcal{C} \to \mathcal{D}$ is a functor. Then we get an induced covariant functor $\varphi^* : \mathcal{C}^* \to \mathcal{D}^*$ given by $\varphi^*(a) = a, \varphi^*(f^*) = \varphi(f)^*$.*

Proof. Suppose $f^* \in \hom(b, a)$. Then $\varphi^*(f^*) = \varphi(f)^* \in \hom(\varphi(b), \varphi(a))$, as required. The second functor axiom is clearly satisfied, as the identity for a in $\mathcal{C}^*$ is Id_a^*. For the third axiom, suppose that $f^* \in \hom(a, b), g^* \in \hom(b, c)$. Then

$$\varphi^*(g^* \circ f^*) = \varphi^*((f \circ g)^*) = \varphi(f \circ g)^* = (\varphi(f) \circ \varphi(g))^* = \varphi(g)^* \circ \varphi(f)^* = \varphi^*(g^*) \circ \varphi^*(f^*)$$

as required. □

This functor is, unsurprisingly, often called the *dual functor*. Note that the map taking a category to its dual and a functor to its dual is then an endofunctor of **Cat**, in particular an invertible one.

One may ask, of course, why we would bother with contravariant functors. One motivation is that it allows us to freely switch between a category and its dual. Say, for example, that we have a functor $f : \mathcal{C}^* \to \mathcal{D}$. Then we can define a corresponding contravariant functor $f_* : \mathcal{C} \to \mathcal{D}$ by

$$f_*(a) = f(a), f_*(\varphi) = f(\varphi^*)$$

This is just part of a larger trend, called duality. In order to introduce this, we will unfortunately[4] need to delve a little deeper into logic.

[4]Depending on your viewpoint I suppose.

Definition 7.2.4. Let a, b be arbitrary symbols for objects in a category, and f, g, h ones for arrows. An atomic statement is then one of the following.

1. $a = \text{dom}(f)$.
2. $a = \text{codom}(f)$.
3. f is the identity arrow for a.
4. g can be composed with f to get h.
5. $a = b$.
6. $f = g$.

A statement is any well-formed phrase[5] built from atomic statements, quantifiers (for all and there exists), and connectives (and,or,if and only if, etc.). A sentence, in turn, is a statement where every variable (symbol representing an object, arrow, etc.) is bound to a quantifier.

Let's take a look at some examples. The following phrase would be a sentence.

Example 7.2.1. For all arrows $f : a \to b$, there exists an object c and arrows $g : b \to c, h : c \to a$ such that $h \circ g \circ f$ is the identity arrow for a.

Note. In the above example, $f : a \to b$ is an abbreviation of f has domain a and codomain b.

The next phrase would be a statement, but not a sentence.

Example 7.2.2. a is the domain of f.

We can then, from these, construct *dual statement*.

Definition 7.2.5. Let Σ be a statement. The dual of the statement, denoted Σ^*, is the original statement with the following replacements for the component atomic statements

$$a = \text{dom}(f) \leftrightarrow a = \text{codom}(f).$$
$$g \text{ can be composed with } f \text{ to get } h \leftrightarrow f \text{ can be composed with } g \text{ to get } h.$$

To phrase the above much more simply, all arrows and compositions are reversed. The interesting thing is what happens to our category axioms when they're dualized. Namely, absolutely nothing. All that happens is the domain and codomain functions being switched. Thus, we get the following fundamental result.

Theorem 7.2.6 (Duality Principle)**.** *A statement Σ holds in a category $\mathcal{C}$ following from the axioms of category theory if and only if Σ^* holds in $\mathcal{C}^*$.*

For examples of this, see section 2.1 in [Lan10], which also goes into far more detail about the concept of the duality principle. We will not get much deeper into it here, as it won't be relevant to us for a while.

Another basic construction in category theory is the *product category*.

[5]This has an official definition, but for our purposes it just means that it makes sense.

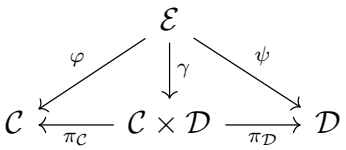
Definition 7.2.7. Let $\mathcal{C}, \mathcal{D}$ be categories. The product category $\mathcal{C} \times \mathcal{D}$ has objects $\text{ob}(\mathcal{C}) \times \text{ob}(\mathcal{D})$, arrows $\hom(\mathcal{C}) \times \hom(\mathcal{D})$, and element-wise composition.

In essence, it's exactly what you'd think the product of two categories would be. These product categories also come with obvious projection functors, defined in the following manner.

$$\pi_{\mathcal{C}}((a,b)) = a, \pi_{\mathcal{C}}((f,g)) = f, \pi_{\mathcal{D}}((a,b)) = b, \pi_{\mathcal{D}}((f,g)) = g$$

It turns out that these projections satisfy a very nice property.

Proposition 7.2.8. *For any category $\mathcal{E}$ and functors $\varphi : \mathcal{E} \to \mathcal{C}, \psi : \mathcal{E} \to \mathcal{D}$, there exists a unique functor $\gamma : \mathcal{E} \to \mathcal{C} \times \mathcal{D}$ making the following diagram commute.*



Proof. We define the functor γ by

$$\gamma(\epsilon) = (\varphi(\epsilon), \psi(\epsilon))$$

where $\epsilon \in \mathcal{E}$ is an object or arrow. The verification that this is our desired (unique) functor is then trivial. □

In the construction above, we denote $\gamma = \varphi \times \psi$ and call it the product of the two functors. One can note that

$$(f \times g) \circ (h \times q) = (f \circ h) \times (g \circ q)$$

Thus, $\times : \mathbf{Cat} \times \mathbf{Cat} \to \mathbf{Cat}$ is a functor. For this reason we often call functors from product categories *bifunctors*, as they can be looked at as maps on the product which are functors in each argument.

We finish by proving two facts about bifunctors.

Proposition 7.2.9. *Let $\mathcal{C}, \mathcal{D}, \mathcal{E}$ be categories. For all objects $c \in \text{ob}(\mathcal{C}), d \in \text{ob}(\mathcal{D})$, let $L_c : \mathcal{D} \to \mathcal{E}, M_d : \mathcal{C} \to \mathcal{E}$ be functors such that $L_c(d) = M_d(c)$. Then there exists a bifunctor $S : \mathcal{C} \times \mathcal{D} \to \mathcal{E}$ such that $S(\cdot, d) = M_d, S(c, \cdot) = L_c$ (where $S(\cdot, d)$ denotes S with the second argument fixed to d for the object and Id_d for the arrow, and similar for $S(c, \cdot)$) if and only if for any pair of arrows $f : c \to c', g : d \to d'$ we get*

$$M_{d'}(f) \circ L_c(g) = L_{c'}(g) \circ M_d(f)$$

Furthermore, in this case we get $S(f, g) = M_{d'}(f) \circ L_c(g)$.

Proof. It is clear that

$$(\text{Id}_{c'}, g) \circ (f, \text{Id}_d) = (f, d) = (f, \text{Id}_{d'}) \circ (\text{Id}_c, g)$$

Applying S to both sides (assuming it existed), we'd get the following commutative diagram.

$$\begin{array}{ccc} S(c,d) & \xrightarrow{S(f,\mathrm{Id}_d)} & S(c',d) \\ {\scriptstyle S(\mathrm{Id}_c,g)}\downarrow & \searrow^{S(f,g)} & \downarrow{\scriptstyle S(\mathrm{Id}_{c'},g)} \\ S(c,d') & \xrightarrow[S(f,\mathrm{Id}_{d'})]{} & S(c',d') \end{array}$$

This is equivalent to the diagram

$$\begin{array}{ccc} S(c,d) & \xrightarrow{M_d(f)} & S(c',d) \\ {\scriptstyle L_c(g)}\downarrow & \searrow^{S(f,g)} & \downarrow{\scriptstyle L_{c'}(g)} \\ S(c,d') & \xrightarrow[M_{d'}(f)]{} & S(c',d') \end{array}$$

which is exactly the claimed relations above. This proves necessity, so all that remains is to show that a map defined in this manner is, in fact, a functor. This is left to the reader. □

Theorem 7.2.10. *Let $S, S' : \mathcal{C} \times \mathcal{D} \to \mathcal{E}$ be two bifunctors. Let $\alpha : \mathrm{ob}(\mathcal{C} \times \mathcal{D}) \to \hom(\mathcal{E})$ be a map such that $\alpha(c,d) : S(c,d) \to S'(c,d)$. Then α is a natural transformation from S to S' if and only if α is natural in each argument.*

Proof. First, suppose that α is natural. Then for any $(f,g) \in \hom((c,d),(c',d'))$, the following diagram commutes.

$$\begin{array}{ccc} S(c,d) & \xrightarrow{\alpha(c,d)} & S'(c,d) \\ {\scriptstyle S(f,g)}\downarrow & & \downarrow{\scriptstyle S'(f,g)} \\ S(c',d') & \xrightarrow[\alpha(c',d')]{} & S'(c',d') \end{array}$$

In particular, this holds fixing either c or d. Now, suppose that α is natural in each argument. Then for any $f \in \hom(c,c'), g \in \hom(d,d')$, the following diagrams commute.

$$\begin{array}{ccc} S(c,d) & \xrightarrow{\alpha(c,d)} & S'(c,d) \\ {\scriptstyle S(f,\mathrm{Id}_d)}\downarrow & & \downarrow{\scriptstyle S'(f,\mathrm{Id}_d)} \\ S(c',d) & \xrightarrow[\alpha(c',d)]{} & S'(c',d) \end{array} \qquad \begin{array}{ccc} S(c',d) & \xrightarrow{\alpha(c',d)} & S'(c',d) \\ {\scriptstyle S(\mathrm{Id}_{c'},g)}\downarrow & & \downarrow{\scriptstyle S'(\mathrm{Id}_{c'},g)} \\ S(c',d') & \xrightarrow[\alpha(c',d')]{} & S'(c',d') \end{array}$$

We combine these into the following diagram.

$$\begin{array}{ccc} S(c,d) & \xrightarrow{\alpha(c,d)} & S'(c,d) \\ {\scriptstyle S(f,\mathrm{Id}_d)}\downarrow & & \downarrow{\scriptstyle S'(f,\mathrm{Id}_d)} \\ S(c',d) & \xrightarrow[\alpha(c',d)]{} & S'(c',d) \\ {\scriptstyle S(\mathrm{Id}_{c'},g)}\downarrow & & \downarrow{\scriptstyle S'(\mathrm{Id}_{c'},g)} \\ S(c',d') & \xrightarrow[\alpha(c',d')]{} & S'(c',d') \end{array}$$

Since $S(\mathrm{Id}_{c'}, g) \circ S(f, \mathrm{Id}_d) = S(f,g)$ and $S'(\mathrm{Id}_{c'}, g) \circ S'(f, \mathrm{Id}_d) = S'(f,g)$, this proves that α is natural. □

7.3 The Yoneda Lemma

The Yoneda lemma might be the most important but sneakily complicated results to come from category theory. It represents a shift in perspective from looking at objects to looking at maps between them. In particular, what the lemma will say is that there is nothing lost in this shift!

I would like to give credit to [m3m17] for providing the perspective necessary for me to understand the importance of this result. [Lan10] is strangely dismissive of it, and [Lan05] doesn't cover it at all[6]!

Without further ado (except to say that we call the collections $\hom(a, b)$ hom-sets), let's jump right into the thick of things.

Theorem 7.3.1 (Yoneda Lemma)**.** *Suppose $\mathcal{D}$ has small hom-sets, $\varphi : \mathcal{D} \to \mathbf{Set}$ is a functor, and $d \in \mathrm{ob}(\mathcal{D})$. Then there exists a bijection*

$$\gamma : \mathrm{Nat}(\hom(d, \cdot), \varphi) \leftrightarrow \varphi(d)$$

which sends a natural transformation $\alpha \in \mathrm{Nat}(\hom(d, \cdot), \varphi)$ to $\alpha_d(\mathrm{Id}_d) \in \varphi(d)$.

Proof. It suffices to show that any $\alpha \in \mathrm{Nat}(\hom(d, \cdot), \varphi)$ is uniquely defined by $\alpha_d(\mathrm{Id}_d)$, and that any choice of $\alpha_d(\mathrm{Id}_d)$ will cause α to be a well-defined natural transformation. First, suppose that α is a natural transformation. Then for any $f \in \hom(d, d')$, the following diagram commutes.

$$\begin{array}{ccc} \hom(d,d) & \xrightarrow{\alpha_d} & \varphi(d) \\ {\scriptstyle \hom(d,f)}\downarrow & & \downarrow{\scriptstyle \varphi(f)} \\ \hom(d,d') & \xrightarrow[\alpha_{d'}]{} & \varphi(d') \end{array}$$

In particular, we get

$$(\alpha_{d'} \circ \hom(d, f))(\mathrm{Id}_d) = (\varphi(f) \circ \alpha_d)(\mathrm{Id}_d) \Rightarrow \alpha_{d'}(f \circ \mathrm{Id}_d) = \varphi(f)(\alpha_d(\mathrm{Id}_d))$$

Thus, $\alpha_{d'}(f) = \varphi(f)(\alpha_d(\mathrm{Id}_d))$, so α is fully defined by $\alpha_d(\mathrm{Id}_d)$. That any choice of $\alpha_d(\mathrm{Id}_d)$ will do is clear, as the above equation $\alpha_{d'}(f) = \varphi(f)(\alpha_d(\mathrm{Id}_d))$ is the sole restriction coming from the naturality condition. □

Corollary 7.3.1.1. *For any pair of objects $d, d' \in \mathrm{ob}(\mathcal{D})$, each natural transformation from $\hom(d, \cdot)$ to $\hom(d', \cdot)$ has the form $\hom(f, \cdot)$ for some unique $f : d' \to d$.*

Proof. By the Yoneda lemma we have the bijection

$$\gamma : \mathrm{Nat}(\hom(d, \cdot), \hom(d', \cdot)) \leftrightarrow \hom(d', d)$$

[6] I suspect because it's a result with more applications in algebraic geometry/topology than anything, but he also spends some time on representation functors so who knows really.

Pick any natural transformation α from $\hom(d,\cdot)$ to $\hom(d',\cdot)$, and define $f = \gamma(\alpha)$. Pick any $c \in \mathrm{ob}(\mathcal{D})$ and $g : d \to c$. Then

$$\alpha_c(g) = \hom(d', g)(\alpha_d(\mathrm{Id}_d)) = g \circ (\alpha_d(\mathrm{Id}_d)) = g \circ \gamma(\alpha) = g \circ f = \hom(f, c)(g)$$

Thus, $\alpha = \hom(f, \cdot)$. The uniqueness of f follows from γ being a bijection. □

This is the first really important observation from the Yoneda lemma: the set of natural transformations between two hom-functors $\hom(d,\cdot), \hom(d',\cdot)$ is in bijection with the set of arrows from d' to d. Moreover, we've got a simple formula for finding all said natural transformations! But we're not even close to done, because the lemma has far deeper consequences.

γ (the Yoneda map), it turns out, is not the main object of interest that arises from the Yoneda lemma. Define a functor $\mathcal{Y} : \mathcal{D}^* \to \mathrm{Func}(\mathcal{D}, \mathbf{Set})$ given by, for any $d, d' \in \mathrm{ob}(\mathcal{D})$ and $f : d' \to d \in \hom(\mathcal{D})$

$$\mathcal{Y}(d) = \hom(d, \cdot)$$
$$\mathcal{Y}(f^*) = \hom(f, \cdot)$$

This is called the *Yoneda functor*, and it is this functor which allows us to demonstrate the real idea behind the Yoneda lemma. Suppose we want to know if $d \cong d'$. By functoriality[7], this is possible only if $\mathcal{Y}(d) \cong \mathcal{Y}(d')$. Of course if $\mathcal{Y}(d) \cong \mathcal{Y}(d')$, then there exists a natural isomorphism α from $\mathcal{Y}(d)$ to $\mathcal{Y}(d')$. But by the Yoneda lemma' corollary $\mathcal{Y}$ is fully faithful, so there exists a unique $f : d' \to d$ and $g : d \to d'$ such that $\mathcal{Y}(f^*) = \alpha, \mathcal{Y}(g^*) = \alpha^{-1}$. It follows that $\mathcal{Y}(f \circ g) = \mathrm{Id} \Rightarrow f \circ g = \mathrm{Id}_d$, and hence that $d \cong d'$. What we have shown here is the following.

Corollary 7.3.1.2. *$d \cong d'$ if and only if $\hom(d,\cdot) \cong \hom(d',\cdot)$.*

This is where the true magic of the Yoneda lemma happens, it tells us that all the information about an object is encoded in the morphisms from that object. We therefore may, instead of studying objects, study only morphisms. Of course, it seems odd that only arrows away from the object would matter. This is easy to fix! None of our above proofs change if we replace covariant with contravariant functors. That is, the following results hold.

Theorem 7.3.2 (Yoneda Lemma II)**.** *Suppose $\mathcal{D}$ has small* hom*-sets, $\varphi : \mathcal{D} \to \mathbf{Set}$ is a contravariant functor, and $d \in \mathrm{ob}(\mathcal{D})$. Then there exists a bijection*

$$\gamma : \mathrm{Nat}(\hom(\cdot, d), \varphi) \cong \varphi(d)$$

which sends a natural transformation $\alpha \in \mathrm{Nat}(\hom(\cdot, d), \varphi)$ to $\alpha_d(\mathrm{Id}_d) \in \varphi(d)$.

Proof. This follows by noting that contravariant functors from $\mathcal{D}$ are covariant functors from $\mathcal{D}^*$, and that $\hom(\cdot, d)$ in $\mathcal{D}$ is the same as $\hom(d,\cdot)$ in $\mathcal{D}^*$. □

[7]The properties of being a functor.

Corollary 7.3.2.1. *For any pair of objects $d, d' \in \mathrm{ob}(\mathcal{D})$, each natural transformation from $\hom(\cdot, d)$ to $\hom(\cdot, d')$ has the form $\hom(\cdot, f)$ for some unique $f : d \to d'$.*

Corollary 7.3.2.2. *$d \cong d'$ if and only if $\hom(\cdot, d) \cong \hom(\cdot, d')$.*

Our next goal will be to show that the Yoneda map is natural in φ and d (note that this implies its contravariant version would be as well). To do this of course, we need to be a little more precise. First, note that $\varphi \in \mathrm{ob}(\mathrm{Func}(\mathcal{D}, \mathbf{Set}))$. We will consider two functors $N, E : \mathcal{K} = \mathrm{Func}(\mathcal{D}, \mathbf{Set}) \times \mathcal{D} \to \mathbf{Set}$, given by

$$\begin{aligned} N(\varphi, d) &= \mathrm{Nat}(\hom(d, \cdot), \varphi) & N(f, g)(\alpha)_a(h) &= f_a(\alpha_a(h \circ g)) \\ E(\varphi, d) &= \varphi(d) & E(f, g)(x) &= (\gamma_{\psi, d'} \circ N(f, g) \circ \gamma_{\varphi, d}^{-1})(x) \end{aligned}$$

where $(\varphi, d), (\psi, d') \in \mathrm{ob}(\mathcal{K})$, $(f, g) : (\varphi, d) \to (\psi, d')$, $\alpha \in \mathrm{Nat}(\hom(d, \cdot), \varphi)$, $a \in \mathrm{ob}(\mathcal{D})$, $h : d' \to a$, and $x \in \varphi(d)$. Armed with this, we make precise the naturality of γ.

Proposition 7.3.3. *γ is a natural isomorphism from N to E.*

Proof. It suffices to show that γ is a natural transformation from N to E. To that end, pick any $(\varphi, d), (\psi, d') \in \mathrm{ob}(\mathcal{K})$ and $(f, g) : (\varphi, d) \to (\psi, d')$. Then for any $\alpha \in N(\varphi, d)$, we get

$$\begin{aligned} (E(f, g) \circ \gamma_{\varphi, d})(\alpha) &= E(f, g)(\alpha_d(\mathrm{Id}_d)) = (\gamma_{\psi, d'} \circ N(f, g) \circ \gamma_{\varphi, d}^{-1} \circ \gamma_{\varphi, d})(\alpha) \\ &= (\gamma_{\psi, d'} \circ N(f, g))(\alpha) \end{aligned}$$

Thus, the following diagram commutes

$$\begin{array}{ccc} N(\varphi, d) & \xrightarrow{\gamma_{\varphi, d}} & E(\varphi, d) \\ {\scriptstyle N(f,g)}\downarrow & & \downarrow{\scriptstyle E(f,g)} \\ N(\psi, d') & \xrightarrow[\gamma_{\psi, d'}]{} & E(\psi, d') \end{array}$$

as required. □

Let's end off with something very silly : an incredibly overly complicated proof of Cayley's theorem (Theorem 2.2.5).

Proof. Suppose G is a group. We can represent G by a category $\mathcal{C}$ consisting of one object $*$ and morphisms which are all invertible. Note that in this setup, a natural transformation from $\hom(\cdot, *)$ to itself is equivalent to a group homomorphism, and $\hom(*, *)$ is just G. By Yoneda's lemma, we get a bijection

$$\gamma : \mathrm{Nat}(\hom(\cdot, *), \hom(\cdot, *)) \leftrightarrow \hom(*, *) = G$$

and $\mathrm{Nat}(\hom(\cdot, *), \hom(\cdot, *))$ is just a subset of S_G. We just then need to check that γ is a homomorphism. To that end, pick any $\alpha, \beta \in \mathrm{Nat}(\hom(\cdot, *), \hom(\cdot, *))$. By Yoneda's lemma, there exist unique $f, g \in G$ such that $\alpha = \hom(\cdot, f), \beta = \hom(\cdot, g)$. Thus, we get

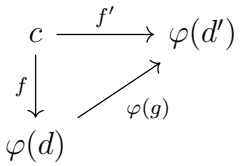
$$\begin{aligned} \gamma_*(\alpha \circ \beta) &= \gamma_*(\hom(\cdot, f) \circ \hom(\cdot, g)) = \gamma_*(\hom(f \circ g, \cdot)) = \mathrm{Id}_* \circ f \circ g = (\mathrm{Id}_* \circ f) \circ (\mathrm{Id}_* \circ g) \\ &= \gamma_*(\alpha) \circ \gamma_*(\beta) \end{aligned}$$

as required. □

7.4 Universals

We next cover the concept of universal arrows and objects, which will be of great use in the following section.

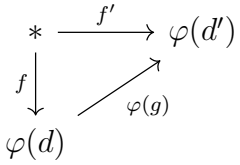
Definition 7.4.1. Suppose $\varphi : \mathcal{D} \to \mathcal{C}$ is a functor and $c \in \text{ob}(\mathcal{C})$. A universal arrow from c to φ is a pair (d, f) consisting of an object $d \in \text{ob}(\mathcal{D})$ and an arrow $f : c \to \varphi(d)$ such that for any other pair (d', f'), with $d' \in \text{ob}(\mathcal{D})$ and $f' : c \to \varphi(d')$, there exists a unique arrow $g : d \to d'$ making the following diagram commute.



When $\mathcal{C} = \mathbf{Set}$, we often instead refer to *universal elements*.

Definition 7.4.2. Suppose $\varphi : \mathcal{D} \to \mathbf{Set}$ is a functor. A universal element of φ is a pair (d, x) consisting of an object $d \in \text{ob}(\mathcal{D})$ and an element $x \in \varphi(d)$ such that for any other such pair (d', x'), there exists a unique arrow $f : d \to d'$ such that $\varphi(f)(x) = x'$.

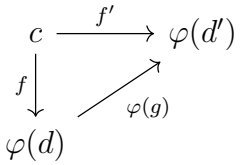
Let's take a moment to consider how these are equivalent. First, suppose that (d, x) is a universal element. Set $c = \{*\}$, the one-element set. Define the arrow $f : c \to \varphi(d)$ by $f : * \mapsto x$. Note that if (d', f') is any other pair, picking the function $f' : * \to \varphi(d')$ is equivalent to picking an element of d' which is in its image, call this element x'. Then by the definition of the universal element, there exists a unique $g : d \to d'$ such that $\varphi(g)(x) = x'$, that is a unique g such that the following diagram commutes



Hence, (d, f) is a universal arrow from $*$ to φ. Going the other way is a little more difficult.

Proposition 7.4.3. *Suppose $\mathcal{C}$ has small hom sets, that is for any $a, b \in \text{ob}(\mathcal{C})$, $\hom(a, b)$ is a set. Let $\varphi : \mathcal{D} \to \mathcal{C}$ be a functor, $c \in \text{ob}(\mathcal{C})$. Pick any $d \in \text{ob}(\mathcal{D})$ and $f : c \to \varphi(d)$. Define a functor $F : \mathcal{D} \to \mathbf{Set}$ in the following manner. For any $a, b \in \text{ob}(\mathcal{D})$ and arrow $g : a \to b$, $F(a) = \hom(c, \varphi(a))$ and $F(g)$ is the map from $\hom(c, \varphi(a))$ to $\hom(c, \varphi(b))$ given by composing elements of $\hom(c, \varphi(a))$ by $\varphi(g)$ on the left. Then (d, f) is a universal arrow from c to φ if and only if $(d, \varphi(f))$ is a universal element of F.*

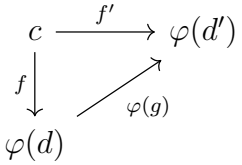
Proof. First, suppose that (d, f) is a universal arrow. Note that $f \in \hom(c, \varphi(d)) = F(d)$. Pick any $d' \in \text{ob}(\mathcal{D})$ and $f' \in \hom(c, \varphi(d')) = F(d')$. Then by the properties of universal arrows there exists a unique arrow $g : d \to d'$ making the following diagram commute



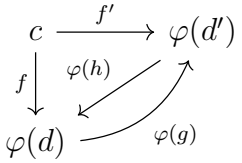
Hence, g is the unique arrow such that $\varphi(g) \circ f = f' \leftrightarrow F(g)(f) = f'$, making (d, f) the desired universal object of F. Now, suppose that (d, f) is a universal object of F. Pick any $d' \in \text{ob}(\mathcal{D})$ and $f' \in \hom(c, \varphi(d'))$. Then by the properties of the universal object, there exists a unique arrow $g : d \to d'$ such that $F(g)(f) = f' \leftrightarrow \varphi(g) \circ f = f'$, thus making (d, f) the desired universal arrow. □

Of course, universality sounds like it should imply some form of uniqueness. This is, in fact, the case.

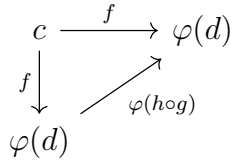
Theorem 7.4.4. *If $(d, f), (d', f')$ are a pair of universal arrows from c to φ, then there exists a unique isomorphism $g : d \cong d'$ making the following diagram commute*



Proof. Uniqueness follows from the uniqueness of g in the definition. By definition, there exists also a unique $h : d' \to d$ making the following diagram commute



That is, the following diagram commutes

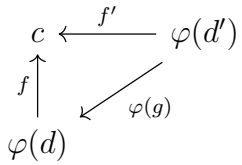


So by the uniqueness property of universal arrows, $h \circ g = \text{Id}_d$. The proof that $g \circ h = \text{Id}_{d'}$ is identical. □

Corollary 7.4.4.1. *Suppose $(d, x), (d', x')$ are a pair of universal elements for φ. Then there exists a unique isomorphism $f : d \to d'$ such that $\varphi(f)(x) = x'$.*

We can also take the dual of the axioms of a universal arrow from an object to a functor to get a universal arrow from a functor to an object.

Definition 7.4.5. Suppose $\varphi : \mathcal{D} \to \mathcal{C}$ is a functor and $c \in \text{ob}(\mathcal{C})$. A universal arrow from φ to c is a pair (d, f) consisting of an object $d \in \text{ob}(\mathcal{D})$ and an arrow $f : \varphi(d) \to c$ such that for any other pair (d', f'), with $d' \in \text{ob}(\mathcal{D})$ and $f' : \varphi(d') \to c$, there exists a unique arrow $g : d' \to d$ making the following diagram commute.

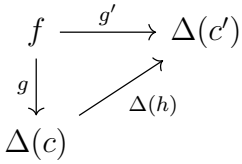


Again, we get that the object d is unique up to some notion of a unique isomorphism.

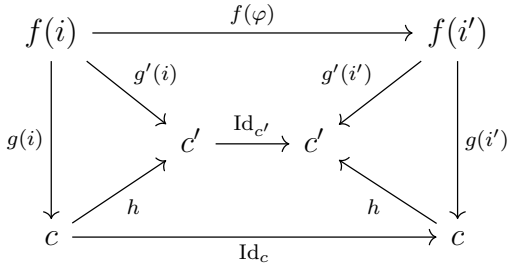
7.5 (Co)limits

Definition 7.5.1. Let $\mathcal{C}, \mathcal{I}$ be categories, and define the diagonal functor $\Delta : \mathcal{C} \to \text{Func}(\mathcal{I}, \mathcal{C})$ by sending each $c \in \text{ob}(\mathcal{C})$ to the constant functor taking the value c at each object of $\mathcal{I}$ and each arrow of $\mathcal{I}$ to Id_c. Arrows $\varphi : c \to c'$ are sent to the natural transformation taking each object of $\mathcal{I}$ to φ. Let $f : \mathcal{I} \to \mathcal{C}$ be an arbitrary functor. A colimit (direct limit) of f is a universal arrow (c, g) from f to Δ.

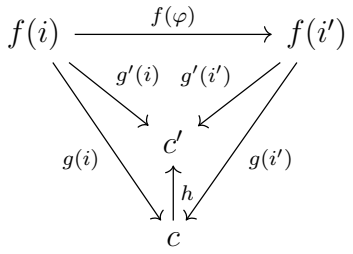
That's a rather dense definition, so let's unpack what it looks like in practice. If (c, g) is a colimit of f then g is a natural transformation from f to $\Delta(c)$, and for all $c' \in \text{ob}(\mathcal{C})$ and natural transformations g' from f to $\Delta(c')$ there exists a unique $h : c \to c'$ such that the following diagram commutes



That is, at any particular $i, i' \in \text{ob}(\mathcal{I})$ or arrow $\varphi : i \to i'$, we get the commutative diagram

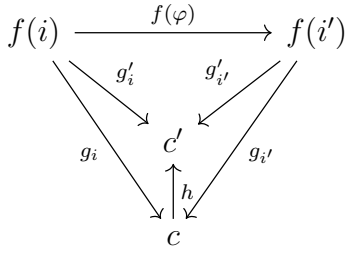


Or, more concisely



This allows us to re-phrase the definition in the following manner.

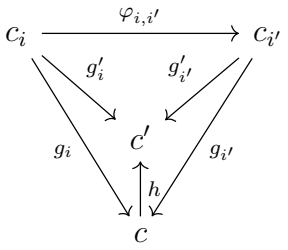
Definition 7.5.2. Let $f : \mathcal{I} \to \mathcal{C}$ be an arbitrary functor. A colimit (direct limit) of f is an object $c \in \text{ob}(\mathcal{C})$ and collection of arrows $\{g_i : f(i) \to c\}_{i \in \text{ob}(\mathcal{I})}$ satisfying, for all arrows $\varphi : i \to i'$, $g_i = g_{i'} \circ f(\varphi)$, such that for any other object $c' \in \text{ob}(\mathcal{C})$ and collection of arrows $\{g'_i : f(i) \to c'\}_{i \in \text{ob}(\mathcal{I})}$ satisfying the same conditions there exists a unique arrow $h : c \to c'$ making the following diagrams commute, for all arrows $\varphi : i \to i'$



This hopefully makes it a bit more clear the sense in which we call this a "limit". The object c is, of course, unique up to unique isomorphism if it exists (in the sense of the previous section), and we therefore denote it (since we only care about things up to isomorphism) by $\varinjlim f$. The functor g is called the *colimiting cone*.

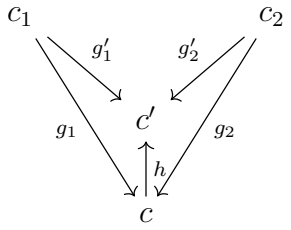
We'll go now over some standard examples. Perhaps the most common (and in fact a way I've seen the colimit defined) is to take $\text{ob}(\mathcal{I})$ to be any partially ordered set, with a unique arrow $i \to i'$ if and only if $i \leq i'$. In this case, denoting $\text{ob}(\mathcal{I}) = I$, our definition becomes the following.

Example 7.5.1. Let $\{c_i\}_{i \in I}$ be a collection of objects in $\mathcal{C}$. For each $i \leq i'$, choose an arrow $\varphi_{i,i'} : c_i \to c_{i'}$, in a manner such that if $i \leq i' \leq i''$, then $\varphi_{i,i''} = \varphi_{i',i''} \circ \varphi_{i,i'}$. A colimit of these collections is an object $c \in \text{ob}(\mathcal{C})$ along with a collection of arrows $\{g_i : c_i \to c\}_{i \in I}$ satisfying $g_i = g_{i'} \circ \varphi_{i,i'}$ such that for any other such object and collection $c', \{g'_i\}$ satisfying this condition, there exists a unique arrow $h : c \to c'$ such that the following diagrams commute



Another common example is to take $\mathrm{ob}(\mathcal{I})$ to be a set with two elements and only identity arrows. In this case, we get

Example 7.5.2. Let $\{c_1, c_2\}$ be a pair of objects in $\mathcal{C}$. The colimit of these objects is an object $c \in \mathrm{ob}(\mathcal{C})$ along with a pair of arrows $g_i : c_i \to c$ such that for any other such object and pair c', g_i', there exists a unique arrow $h : c \to c'$ such that the following diagrams commute

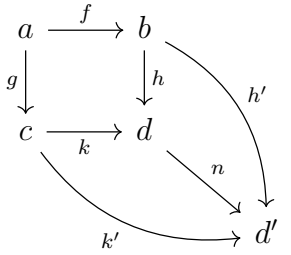


In this case, we call c the *coproduct* or *direct sum* of c_1, c_2, and denote it $c_1 \coprod c_2$ or $c_1 \oplus c_2$. We can, of course, extend this idea to take the coproduct of a collection of elements.

Note. Taking the coproduct of two modules over R is the same as taking their direct sum.

If we instead take $\mathrm{ob}(\mathcal{I}) = \{a, b, c\}$, with identity arrows and $\omega : a \to b, \gamma : a \to c$, we get the following.

Example 7.5.3. Chose three objects $a, b, c \in \mathrm{ob}(\mathcal{C})$ and two arrows $f : a \to b, g : a \to c$. A colimit of this diagram is an object $d \in \mathrm{ob}(\mathcal{C})$ and pair of arrows $h : b \to d, k : c \to d$ such that for any other object $d' \in \mathrm{ob}(\mathcal{C})$ and pair of arrows $h' : b \to d', k' : c \to d'$ there exists a unique arrow $n : d \to d'$ making the following diagrams commute



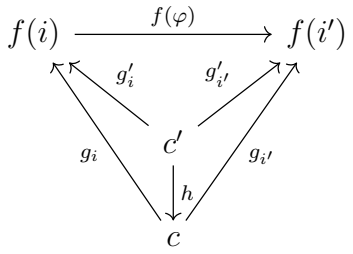
We call d the *pushout* of the diagram, and denote if $b \coprod_{(f,g)} c$.

We can also dualize these definitions to get limits.

Definition 7.5.3. Let $f : \mathcal{I} \to \mathcal{C}$ be an arbitrary functor. A limit (indirect limit) of f is a universal arrow (c, g) from Δ to f.

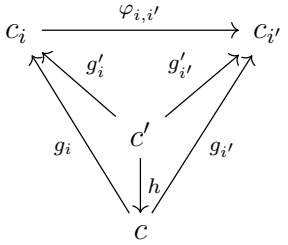
Again, we can expand this out to the following equivalent definition.

Definition 7.5.4. Let $f : \mathcal{I} \to \mathcal{C}$ be an arbitrary functor. A limit of f is an object $c \in \mathrm{ob}(\mathcal{C})$ and collection of arrows $\{g_i : c \to f(i)\}_{i \in \mathrm{ob}(\mathcal{I})}$ satisfying, for each arrow $\varphi : i \to i'$, $g_{i'} = f(\varphi) \circ g_i$, such that for any other object $c' \in \mathrm{ob}(\mathcal{C})$ and collection of arrows $\{g_i' : c' \to f(i)\}_{i \in \mathrm{ob}(\mathcal{I})}$ satisfying the same conditions there exists a unique arrow $h : c' \to c$ making the following diagrams commute



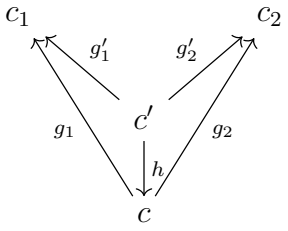
and c is unique up to some notion of unique isomorphism, so we denote it $\varprojlim f$. The map g is called the limiting cone. All the examples from before also carry over pretty much unchanged. The first becomes

Example 7.5.4. Let $\{c_i\}_{i\in I}$ be a collection of objects in $\mathcal{C}$. For each $i \leq i'$, choose an arrow $\varphi_{i,i'} : c_i \to c_{i'}$, in a manner such that if $i \leq i' \leq i''$, then $\varphi_{i,i''} = \varphi_{i',i''} \circ \varphi_{i,i'}$. A limit of these collections is an object $c \in \text{ob}(\mathcal{C})$ along with a collection of arrows $\{g_i : c \to c_i\}_{i\in I}$ satisfying $g_{i'} = \varphi_{i',i} \circ g_i$ such that for any other such object and collection $c', \{g'_i\}$, satisfying the same conditions there exists a unique arrow $h : c' \to c$ such that the following diagrams commute



The second becomes

Example 7.5.5. Let $\{c_1, c_2\}$ be a pair of objects in $\mathcal{C}$. The limit of these objects is an object $c \in \text{ob}(\mathcal{C})$ along with a pair of arrows $g_i : c \to c_i$ such that for any other such object and pair c', g'_i, there exists a unique arrow $h : c' \to c$ such that the following diagrams commute

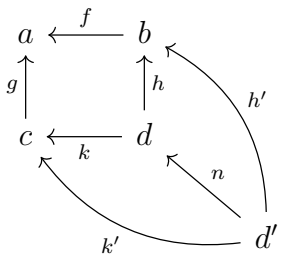


In this case, we call c the *product* or *direct product* of c_1, c_2, and denote it $c_1 \prod c_2$. We can, of course, extend this idea to take the product of a collection of elements.

Note. Taking the product of two modules over R is the same as taking their direct product.

The third changes a little more.

Example 7.5.6. Chose three objects $a, b, c \in \text{ob}(\mathcal{C})$ and two arrows $f : b \to a, g : c \to a$. A limit of this diagram is an object $d \in \text{ob}(\mathcal{C})$ and pair of arrows $h : d \to b, k : d \to c$ such that for any other object $d' \in \text{ob}(\mathcal{C})$ and pair of arrows $h' : d' \to b, k' : d' \to c$ there exists a unique arrow $n : d' \to d$ making the following diagrams commute



We call d the *pullback* of the diagram, and denote if $b \prod_{(f,g)} c$.

7.6 Representations

This section was rather difficult to write, and as a result I ended up pulling from more sources. In particular, ideas for proofs and some definitions in this section were pulled from [ET20], [nLa24a], [nLa24b], [nLa24c], [nLa24d], and [nLa24e].

The Yoneda lemma, as stated so far, is useful, but its interesting consequences apply only to functors of the form $\hom(d, \cdot)$ or $\hom(\cdot, d)$. Our goal in this section will be to show that, in fact, almost any functor to **Set** looks like one of these functors. The first step of our setup is to give a special name these functors.

Definition 7.6.1. Let $\mathcal{D}$ be a category with small hom-sets. A representation of a functor $\varphi : \mathcal{D} \to \mathbf{Set}$ is a pair (d, τ), where $d \in \mathrm{ob}(\mathcal{D})$ and $\tau \in \mathrm{Nat}(\hom(d, \cdot), \varphi)$. A functor is called representable if it has a representation, and the d is called the representing object.

The representation of a contravariant functor is defined similarly, just replacing $\hom(d, \cdot)$ with $\hom(\cdot, d)$.

The second step of our setup is a little more complicated. You may remember the Yoneda functor from section 7.3. We wish to study the dual of this functor

$$\mathcal{Y}^* : \mathcal{D}^{**} \to \mathrm{Func}(\mathcal{D}, \mathbf{Set})^*$$

I first claim that this is actually (via composition with some isomorphisms) a functor from $\mathcal{D} \to \mathrm{Func}(\mathcal{D}^*, \mathbf{Set})$. The first part is fairly obvious, as we already know that $\mathcal{D}^{**} \cong \mathcal{D}$. The second is not an isomorphism per-say, but it is an isomorphism if you restrict $\mathrm{Func}(\mathcal{D}, \mathbf{Set})$ to functors of the form $\hom(d, \cdot)$ (which is the range of $\mathcal{Y}$ anyway). Our functor for the second part is then going to take $\hom(d, \cdot)$ to $\hom(\cdot, d)$. For natural transformations $\tau : \hom(d, \cdot) \dot{\to} \hom(d', \cdot)$, it takes this to a natural transformation ℓ by, for any $f : c \to d$ in $\mathcal{D}^*$, $\ell_c(f) = \tau_c(f_*)^*$. We check that this pair of maps Γ is, in fact, a split monomorphism.

That the object function is injective is clear. For the arrow function, suppose $\ell^1 = \ell^2$. Then for any $f : c \to d$, we get $\tau^1_c(f_*)^* = \tau^2_c(f_*)^* \Rightarrow \tau^1_c(f_*) = \tau^2_c(f_*)$, and hence $\tau^1 = \tau^2$, as required. Finally, we check that Γ is a functor. We just need to check identity and being well-behaved under composition. First, suppose that $d = d'$ and τ is the identity natural transformation for $\hom(d, \cdot)$. Then for any $f : c \to d$, we get $\ell_c(f) = \tau_c(f_*)^* = f^*_* = f$, as

required. Now, suppose that $\tau : \hom(d, \cdot) \dot{\to} \hom(d', \cdot), \alpha : \hom(d', \cdot) \dot{\to} \hom(d'', \cdot)$, and pick any $f : c \to d$. Then we get

$$\begin{aligned}\Gamma(\alpha \circ \tau)_c(f) &= (\alpha \circ \tau)_c(f_*)^* = (\alpha_c \circ \tau_c)(f_*)^* = \alpha_c(\tau_c(f_*))^* = \alpha_c((\tau_c(f_*)^*)_*)^* \\ &= \Gamma(\alpha)_c(\tau_c(f_*)^*) = (\Gamma(\alpha)_c \circ \Gamma(\tau)_c)(f)\end{aligned}$$

as required.

Let's sum that all up. In the end, we instead say that out "new" $\mathcal{Y}^*$ is a functor from $\mathcal{D} \to \mathrm{Func}(\mathcal{D}^*, \mathbf{Set})$, which sends an object d to $\hom(\cdot, d)$ and an arrow $f : d \to d'$ to $\Gamma(\hom(f_*, \cdot)^*)$ which acts by, for any $g : c \to d$,

$$\mathcal{Y}^*(\hom(f_*, \cdot)^*)_c(g) = (\hom(f_*, \cdot)^*_c(g_*))^* = (g_* \circ f_*)^* = f \circ g = \hom(\cdot, f)(g)$$

Hence, our map is $d \to \hom(\cdot, d)$ and $f \to \hom(\cdot, f)$. We call this new $\mathcal{Y}^*$ by the name Ω. Again, the Yoneda lemma implies that Ω is fully faithful. Often, we call $\mathcal{Y}$ the contravariant Yoneda embedding and Ω the covariant Yoneda embedding, this naming convention presumably arising from their respective domains. The point of constructing this Ω was exactly that, to allow us to say things directly about $\mathcal{D}$ rather than $\mathcal{D}^*$ using the Yoneda lemma.

We only need one last pair of definitions before we move on.

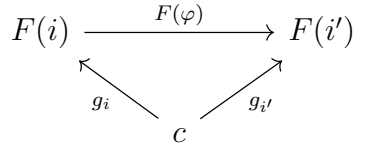
Definition 7.6.2. A category $\mathcal{C}$ is (co)complete[8] if every functor from a small cateogry into $\mathcal{C}$ has a (co)limit.

Definition 7.6.3. A functor $f : \mathcal{C} \to \mathcal{D}$ is (co)continuous if, given any functor $g : \mathcal{I} \to \mathcal{C}$ from a small category $\mathcal{I}$, then $\varprojlim g$ ($\varinjlim g$) existing implies that $\varprojlim(f \circ g)$ ($\varinjlim(f \circ g)$) exists and $\varprojlim(f \circ g) = f(\varprojlim g)$ ($\varinjlim(f \circ g) = f(\varinjlim g)$), with a similar preservation occurring for the corresponding maps.

Alright, let's finally start proving things.

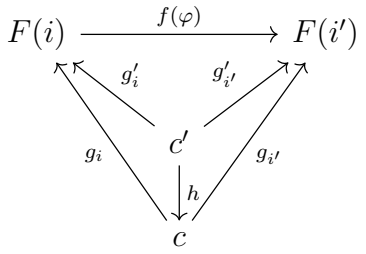
Proposition 7.6.4. Set *is (co)complete.*

Proof. Let $\mathcal{I}$ be any small category, and $F : \mathcal{I} \to \mathbf{Set}$ a functor. For completeness, start by defining c to be the set of all cartesian products of the form $\prod_{i \in \mathcal{I}} x_i$, where $x_i \in F(i)$. Note that since $\mathcal{I}$ is small, c is well-defined. In particular, we'll refine c to keep only the products such that for any arrow $\varphi : i \to i'$, $F(\varphi)(x_i) = x_{i'}$. We define the $g_i : c \to F(i)$ to be the projection maps. First, we show the commutativity of all triangles of the form



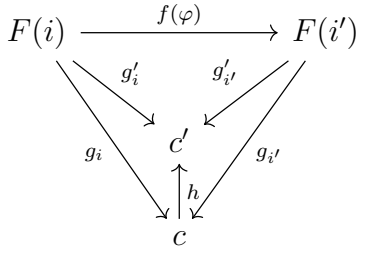
for $\varphi : i \to i'$. Indeed, suppose $(x_i)_{i \in \mathrm{ob}(\mathcal{I})} \in c$. Then by construction, $(\varphi \circ g_i)(x_i)_{i \in \mathrm{ob}(\mathcal{I})} = \varphi(x_i) = x_{i'} = g_{i'}(x_i)_{i \in \mathrm{ob}(\mathcal{I})}$, as required. Now, suppose that $(c', \{g_i'\})$ is any other such pair. Suppose there existed a map $h : c' \to c$ such that the diagram

[8] You will occationally see this referred to as small (co)complete instead.



commutes. Then for any $y \in c'$, we'd need for $(g_i \circ h)(y) = g'_i(y) \Rightarrow h(y) = (g'_i(y))_{i \in \text{ob}(\mathcal{C})}$. We just need to check, then, that h is well-defined like this to show that $c = \varprojlim F$. But this follows immediately by noting that $\varphi \circ g'_i = g'_{i'}$.

Now, for co-completeness. Define $C = \bigcup_{i \in \text{ob}(\mathcal{I})} F(i) \times \{i\}$, and restrict this C to only include an element $x \in F(i)$ if, for any given $i' \in \text{ob}(\mathcal{I})$ and arrows $\varphi, \psi : i \to i'$, $F(\varphi)(x) = F(\psi)(x)$. Again, since F is small, these constructions are all well-defined sets. We'll say that $(x, i) \sim_0 (x', i')$ in C if there exists some $\varphi : i \to i'$ such that $F(\varphi)(x) = x'$ or if there exists some $\varphi : i' \to i$ such that $F(\varphi)(x') = x$. We then iteratively define $\sim_k$, for each $k \in \mathbb{N}$, by saying $(x, i) \sim_k (x', i')$ if there exists some (x'', i'') such that $(x, i) \sim_{k-1} (x'', i'') \sim_{k-1} (x', i')$. Finally, we'll say that $(x, i) \sim (x', i')$ if there exists some $k \in \mathbb{N}$ such that $(x, i) \sim_k (s', i')$. $\sim$ is an equivalence relation on C, and we'll define $c = C/\sim$, with projection map q. Define each g_i by the inclusion $F(i) \hookrightarrow F(i) \times \{i\}$ followed by applying q. We'll show that these satisfy the required properties. First, suppose that $\varphi : i \to i'$ and $x \in F(i)$. Then $(x, i) \sim_0 (x', i')$, so by construction $g_i(x) = (g_{i'} \circ \varphi)(x)$ as required. Now, suppose that $(c', \{g'_i\})$ is another such pair. Suppose there existed a map $h : c \to c'$ such that the diagram



commutes. Pick any $(x, i) \in C$. Then we need for $g'_i(x) = (h \circ g_i)(x) \Rightarrow h([(x, i)]) = g'_i(x)$. Thus, we're done as long as $(x, i) \sim (x', i')$ implies that $g'_i(x) = g'_{i'}(x')$. This follows from two observations.

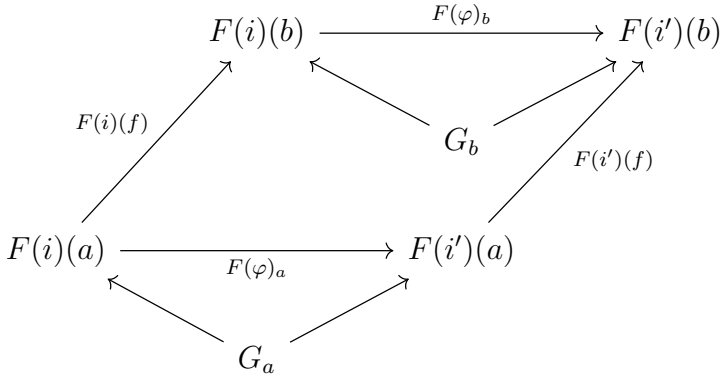
1. $\sim$ is the weakest equivalence such that $(x, i) \sim_0 (x', i') \Rightarrow (x, i) \sim (x', i')$.

2. Consider the projection map $r : C \to c'$ given by all the g'_i. r defines an equivalence relation on C, call this R. If $(x, i) \sim_0 (x', i')$, then (without loss of generality) there exists some $\varphi : i \to i'$ such that $F(\varphi)(x) = x'$. Thus, $g'_i(x) = (g'_{i'} \circ F(\varphi))(x) = g'_{i'}(x')$, so R is an equivalence relation such that $(x, i) \sim_0 (x', i') \Rightarrow (x, i) R (x', i')$.

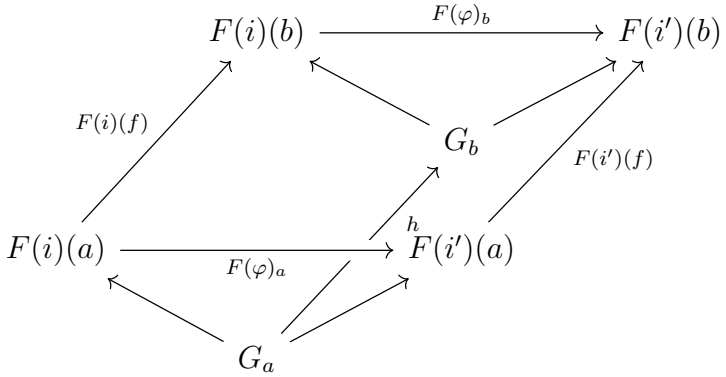
□

Proposition 7.6.5. *If $\mathcal{C}$ is a category, then* $\text{Func}(\mathcal{C}, \mathbf{Set})$ *is (co)complete.*

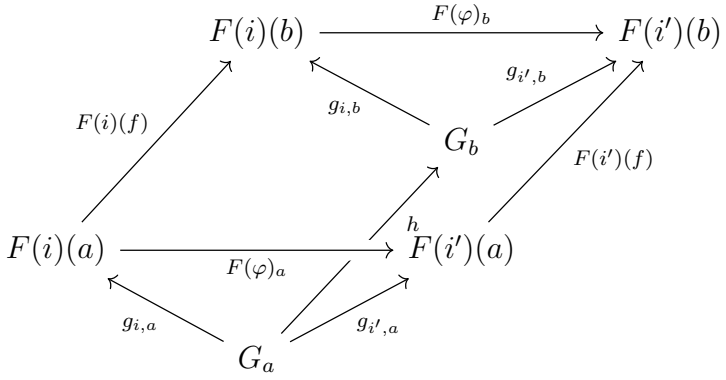
Proof. We'll do the proof only for completeness, co-completeness is a similar proof. Suppose $\mathcal{I}$ is small, and $F : \mathcal{I} \to \text{Func}(\mathcal{C}, \mathbf{Set})$ a functor. First we note that, by the previous proposition, $F(\cdot)(c)$ has a limit for any $c \in \text{ob}(\mathcal{C})$. We'll define then, for each $c \in \text{ob}(\mathcal{C})$, that $G_c = \varprojlim F(\cdot)(c)$. Then for any $a, b \in \text{ob}(\mathcal{C})$, $f : a \to b$, and $\varphi : i \to i'$, we get a commutative diagram



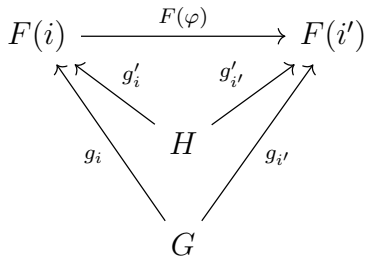
where the unlabelled arrows are those arising from the limiting cones. Thus, by the universal property of the limit, there exists a unique arrow $h : G_a \to G_b$ such that the diagram



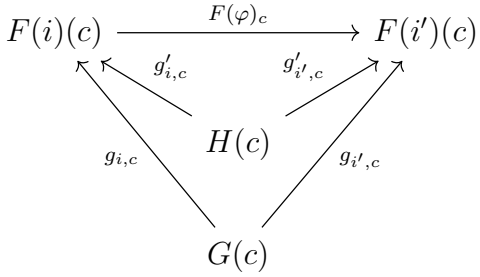
always commutes. Thus, we can define a functor $G : \mathcal{C} \to \mathbf{Set}$ by the rules $G(c) = G_c, G(f) = h$. That this is a functor is guaranteed by the uniqueness of each h. We wish to show that $G = \varprojlim F$. To do this, we give some names to the unlabelled arrows in the above diagram



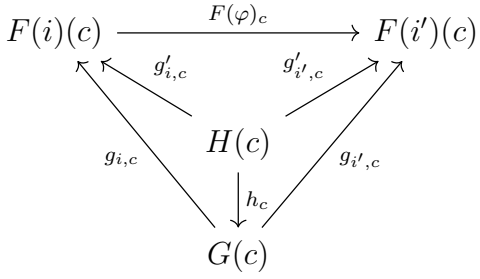
We'll then define natural transformations $g_i : G \dot{\to} F(i)$ by $(g_i)_a = g_{i,a}$. All that remains then is to check that G with these natural transformations satisfies the required universal property. To that end, pick any other functor $H : \mathcal{C} \to \mathbf{Set}$ and natural transformations g'_i satisfying the requisite commutative diagram. Then we get a commutative diagram



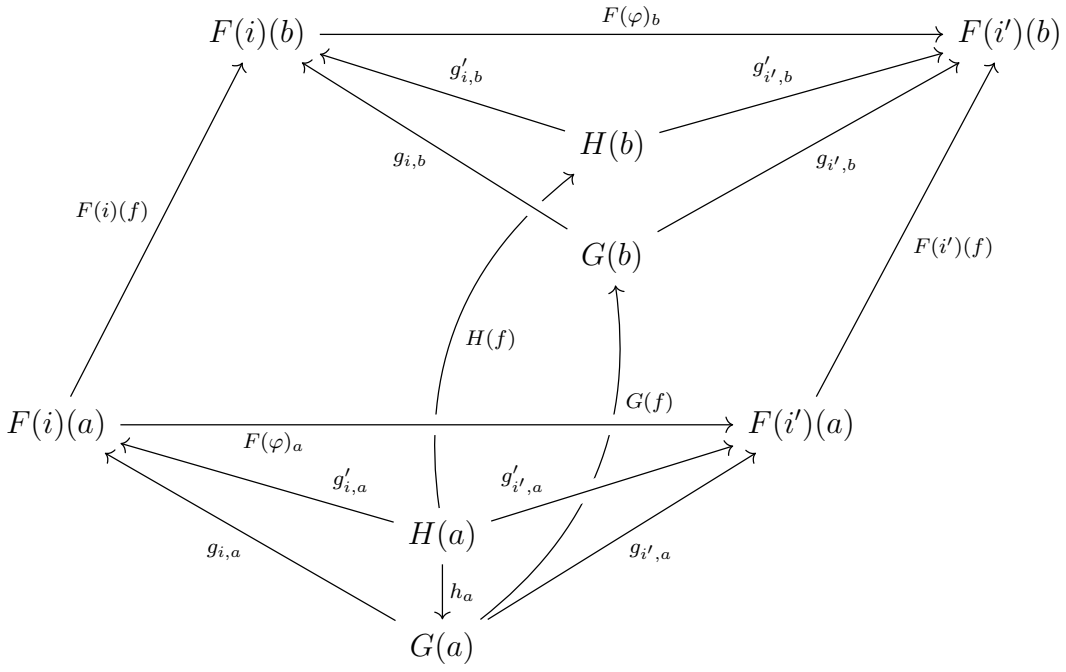
Evaluating at any particular $c \in \mathrm{ob}(\mathcal{C})$, this becomes



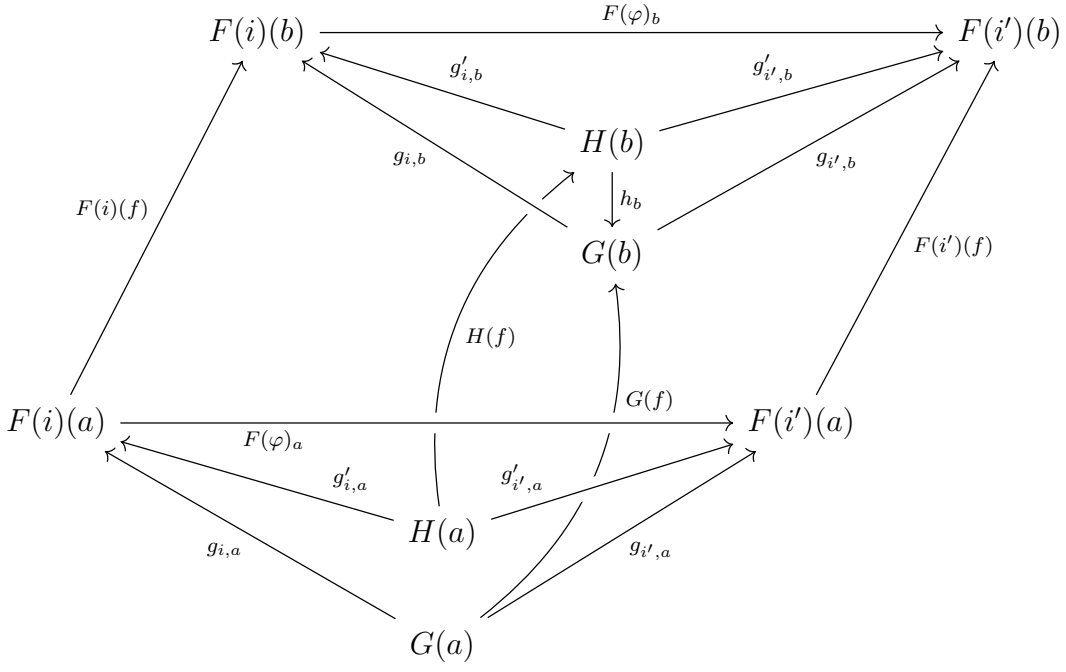
So by the universal property of the limit, and since $G(c) = \varprojlim F(\cdot)(c)$, there exists a unique $h_c : H(c) \to G(c)$ such that



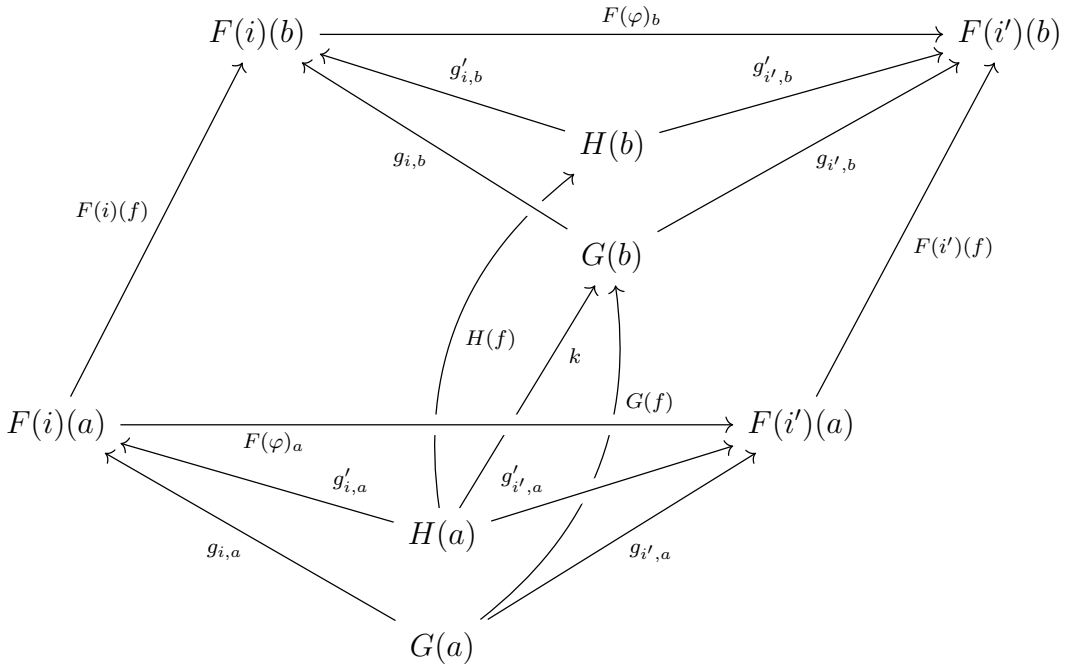
commutes. This uniquely defines our desired natural transformation $h : c \mapsto h_c$, we need now just check that this is indeed a natural transformation. Indeed, this follows by noting that for any $f : a \to b$, the diagrams



and



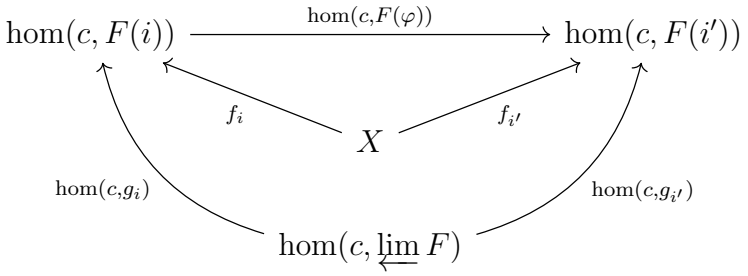
commute, and by the universal property of the limit there exists a unique arrow $k : H(a) \to G(b)$ such that



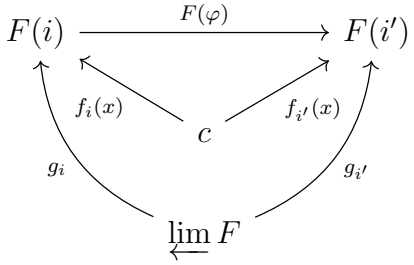
commutes. □

Proposition 7.6.6. *If $\mathcal{C}$ has small hom-sets, then each* $\hom(c,\cdot)$ $(\hom(\cdot,c))$ *is (co)continuous.*

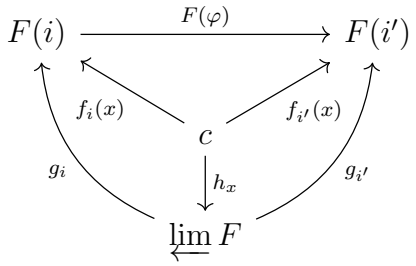
Proof. Again, and for the rest of this section, we prove only the first statement as the other is dual. Fix some $c \in \mathrm{ob}(\mathcal{C})$ and small category $\mathcal{I}$. Suppose $F : \mathcal{I} \to \mathcal{C}$ is a functor with a limit. We'll show that $\varprojlim \hom(c, F(\cdot)) = \hom(c, \varprojlim F)$, with the expected maps. To that end, pick any set X and maps $\{f_i\}$ such that for any $\varphi : i \to i'$, the diagram



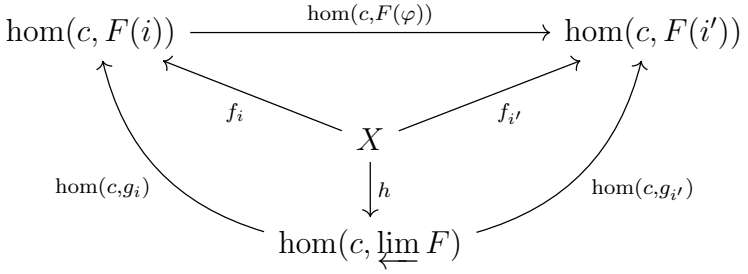
commutes. Each $x \in X$ associates under f_i to some unique arrow $f_i(x) : c \to F(i)$, satisfying $(F(\varphi) \circ f_i)(x) = f_{i'}(x)$. That is, the diagram



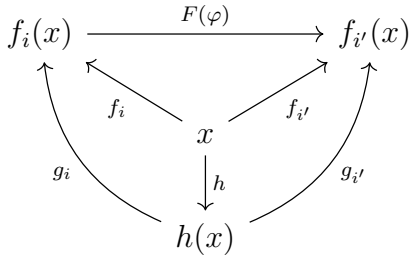
commutes, and hence there exists a unique $h_x : c \to \varprojlim F$ such that



commutes. Note that $h_x \in \hom(c, \varprojlim F)$, so we've defined a function $h : X \to \hom(c, \varprojlim F)$ given by $h(x) = h_x$. Our goal will be to show that



commutes, and h is the unique function for which this holds. Indeed, at any particlar $x \in X$ this diagram becomes



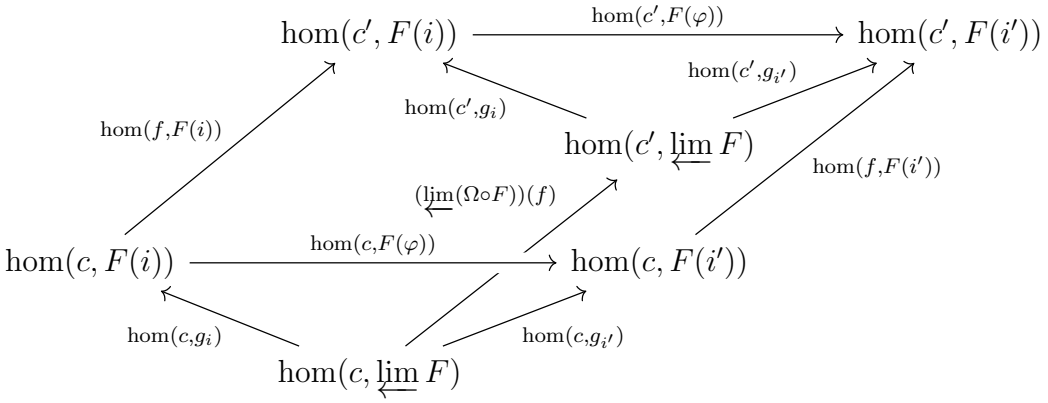
which commutes by definition. This also makes uniqueness clear. □

Proposition 7.6.7. *The (contravariant) Yoneda embedding is (co)continuous.*

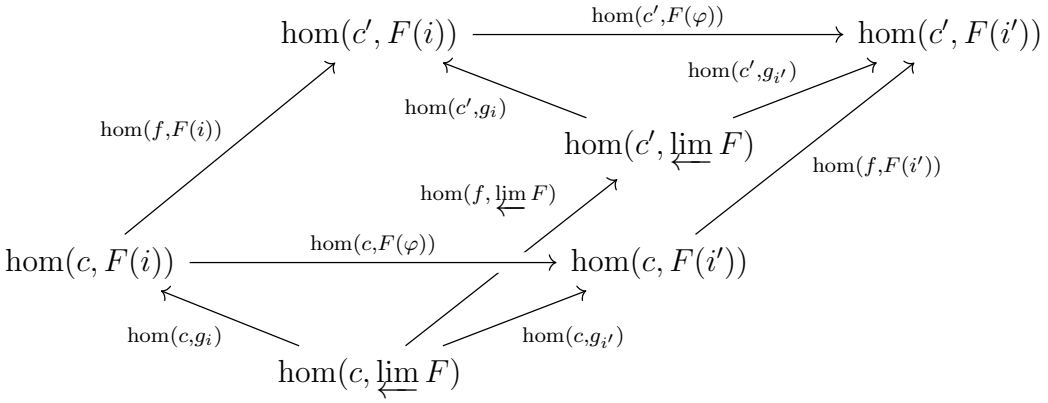
Proof. Suppose $\mathcal{I}$ is small, $\mathcal{C}$ has small hom-sets, and $F : \mathcal{I} \to \mathcal{C}$ is a functor with a limit. Since functor categories are complete, $\varprojlim(\Omega \circ F)$ exists. Furthermore, we know by previous results that at any $c \in \mathrm{ob}(\mathcal{C})$

$$(\varprojlim(\Omega \circ F))(c) = \varprojlim((\Omega \circ F)(\cdot)(c)) = \varprojlim(\hom(c, F)) = \hom(c, \varprojlim F) = \Omega(\varprojlim F)(c)$$

Also, by the previous proposition and the construction of the limits in proposition 7.6.5, our corresponding maps are the desired $\hom(\cdot, g_i)$, where g_i are the original maps from F. Suppose $f : c \to c'$ is an arrow in $\mathcal{C}^*$. Then $(\varprojlim(\Omega \circ F))(f)$ is the unique arrow making the diagram



However, the diagram[9]



commutes, as for any $k \in \hom(c, \varprojlim F)$ we get $g_i \circ k \circ f_*$ via both paths through the left side (and similarly for the right side) of the diagram. □

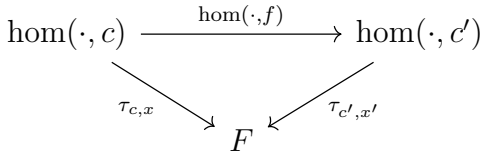
To summarize, we know now that every category $\mathcal{C}$ with small hom-sets is embedded continuously in a complete category via the Yoneda embedding. What we really want, then, is for every element in $\mathrm{Func}(\mathcal{C}^*, \mathbf{Set})$ to be the limit of something composed with the Yoneda embedding, that is we want everything to just be a limit of functors of the form $\hom(\cdot, c)$, as we understand these functors very well.

Theorem 7.6.8 (Density)**.** *Let $\mathcal{C}$ be small. Then any contravariant (covariant) functor $F : \mathcal{C} \to \mathbf{Set}$ is a colimit of representable contravariant (covariant) functors.*

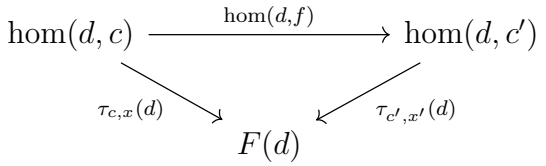
Proof. We prove only the contravariant case, the covariant case is similar. Define $\mathcal{I}$ to be the category having as objects pairs (x, c) with $c \in \mathrm{ob}(\mathcal{C})$ and $x \in F(c)$ and having as arrows $(x, c) \to (x', c')$ maps $g : c \to c'$ in $\mathcal{C}$ such that $F(g)(x') = x$. We can define a covariant functor $G : \mathcal{I} \to \mathrm{Func}(\mathcal{C}^*, \mathbf{Set})$ by $G(x, c) = \hom(\cdot, c)$ and $G(f) = \hom(\cdot, f)$, that is by composing the forgetful functor and Yoneda embedding. We claim that $F = \varinjlim G$. Indeed, we get by the Yoneda lemma a bijection $\gamma^{-1} : F(c) \to \mathrm{Nat}(\hom(\cdot, c), F)$ for each $c \in \mathrm{ob}(\mathcal{C})$.

[9] All of these diagrams are only really well-defined up to applying some unique isomorphism to $\hom(c, \varprojlim F)$. But of course we only care about objects up to unique isomorphisms anyhow, so that's fine!

Thus, we can get for each $x \in F(c)$ a unique natural transformation $\tau_{c,x} : \hom(\cdot, c) \dot{\to} F$. The diagram



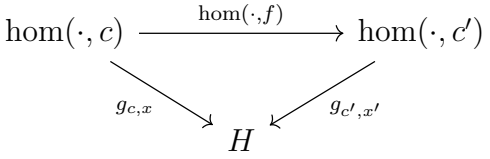
commutes, as for any $d \in \text{ob}(\mathcal{C})$ this becomes



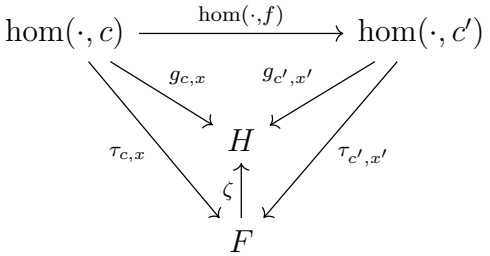
Picking some $g \in \hom(d, c)$, one can calculate that $\tau_{c,x}(d)(g) = F(g)(x)$ and

$$(\tau_{c',x'}(d) \circ \hom(d, f))(g) = \tau_{c',x'}(d)(f \circ g) = F(f \circ g)(x') = (F(g) \circ F(f))(x') = F(g)(x)$$

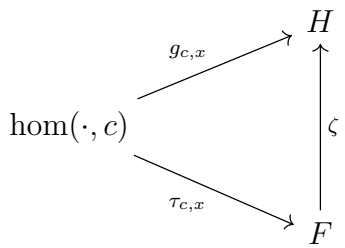
as required. We just need to show now that this commutative diagram satisfies the desired universal property. To that end, pick any other $H \in \text{Func}(\mathcal{C}^*, \mathbf{Set})$ and set of maps $g_{c,x} : \hom(\cdot, c) \dot{\to} G$ such that



commutes. By the Yoneda lemma, each $g_{c,x}$ arises from the bijection $\gamma^{-1} : \text{Nat}(\hom(\cdot, c), H) \leftrightarrow H(c)$. In particular, we'll say that each $g_{c,x}$ arises from some $z_{c,x} \in H(c)$. Now, suppose we had a natural transformation $\zeta : F \dot{\to} H$ which made the diagram



commute. In particular, by our assumptions, we need only check that



commutes. Evaluating at c and $\mathrm{Id}_c \in \hom(c,c)$, this implies

$$z_{c,x} = g_{c,x}(c)(\mathrm{Id}_c) = (\zeta(c) \circ \tau_{c,x}(c))(\mathrm{Id}_c) = \zeta(c)(x)$$

Thus, if such a natural transformation exists it must be given by $\zeta(c)(x) = z_{c,x}$. We check that this is, in fact, a natural transformation. To that end, fix some $a, b \in \mathrm{ob}(\mathcal{C})$ and $g \in \hom(a,b)$. Consider the diagram

$$\begin{array}{ccc} F(b) & \xrightarrow{\zeta(b)} & H(b) \\ {\scriptstyle F(g)}\downarrow & & \downarrow{\scriptstyle H(g)} \\ F(a) & \xrightarrow[\zeta(a)]{} & H(a) \end{array}$$

Chasing any $x \in F(b)$ around the diagram, we get the expressions

$$H(g)(\zeta(b)(x)) = H(g)(z_{b,x}), \zeta(a)(F(g)(x)) = z_{a,F(g)(x)}$$

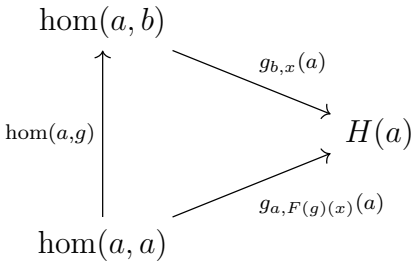
So we need to show that these two expressions are equal. But of course the diagrams

$$\begin{array}{ccc} \hom(b,b) & \xrightarrow{g_{b,x}(b)} & H(b) \\ {\scriptstyle \hom(g,b)}\downarrow & & \downarrow{\scriptstyle H(g)} \\ \hom(a,b) & \xrightarrow[g_{b,x}(a)]{} & H(a) \end{array}$$

commute, and hence

$$g_{b,x}(a)(g) = (g_{b,x}(a) \circ \hom(g,b))(\mathrm{Id}_b) = (H(g) \circ g_{b,x}(b))(\mathrm{Id}_b) = H(g)(z_{b,x})$$

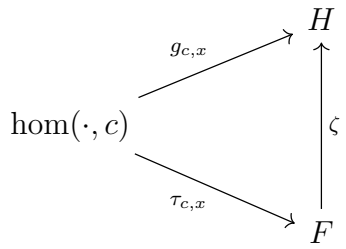
Furthermore, the diagram



commutes by assumption, so chasing around Id_a we get

$$z_{a,F(g)(x)} = g_{a,F(g)(x)}(a)(\mathrm{Id}_a) = (g_{b,x} \circ \hom(a,g))(\mathrm{Id}_a) = g_{b,x}(a)(g)$$

Thus, we get $z_{a,F(g)(x)} = g_{b,x}(a)(g) = H(g)(z_{b,x}) = H(g)(\zeta(b)(x))$, making ζ a natural transformation. Finally, we check that



commutes. Fix any $b \in \mathrm{ob}(\mathcal{C})$, and $g \in \hom(b, c)$. Then we get

$$g_{c,x}(b)(g) = z_{b,F(g)(x)} = \zeta(b)(F(g)(x)) = (\zeta(b) \circ \tau_{c,x}(b))(g)$$

as required. □

We can sum up the above section with a dogma : **Every functor to Set is uniquely specified by representable functors, which arise from the (co)continuous Yoneda embedding, and two objects in a category are isomorphic if and only if their corresponding representable functors are isomorphic**.

Or, put more succinctly.

All the information about small categories are contained in their representable functors to Set.

We need not look at objects or arrows at all really, only functors.

7.7 Some Final Remarks

I cannot emphasize enough how brief of an overview of category theory this was. I leave off here in an attempt not to get too diverted from our goal of studying algebra, as categories often crop up more in situations arising from algebraic topology or geometry. Indeed, we will see them again in detail in chapters 9 and 10, as we start to cover homology, cohomology, Abelian categories, monoidal categories, and the like.

There exists also generalizations of categories, leading to n-categories and ∞-categories. The interested reader can start there study of these in [Lan10], but I believe there are also books dedicated exclusively to that subject which are likely better. On that note, I have yet to find a good introductory book for category theory. It's really one of those subjects that you just need to get your hands dirty with to fully understand what's going on. There are plenty out there though, so feel free to try and find one. It will, at the very least, be better than [Lan10].

Finally, I would be remiss if I didn't mention the existence of homotopy type theory [Uni13]. Is it useful? Not really. But it does seem very cool, and I need to justify my purchase of that book somehow.

Part IV

Advanced Algebra

Chapter 8

Field Extensions and Galois Theory

8.1 Algebraic Extensions

Much of this section is based off of notes from lectures by Dr. Sujatha Ramdorai, and similar sections in [Lan05].

Let's start with the most basic definition in field theory.

Definition 8.1.1. Let F, K be fields. We say that K is a field extension of F, denoted K/F if there exists an embedding (i.e. an injective homomorphism) $F \hookrightarrow K$.

We take the convention of identifying a field with its embedding in an extension K, as this is a much easier way to work with things. The first tool we'd like, in field theory, is a way to measure the "size" of an extension. To do this, we note that K is a vector space over F, and define

Definition 8.1.2. Let K/F be a field extension. The degree of the extension, denoted $[K : F]$, is $\dim_F(K)$. K/F is called a finite extension if $[K : F] < \infty$.

Proposition 8.1.3. *If $E/K/F$ is a tower of field extensions (i.e. E/K is a field extension and K/F is a field extension), then $[E : F] = [E : K][K : F]$.*

Proof. Let $\{x_i\}_{i\in I}$ be a basis for K over F and $\{y_j\}_{j\in J}$ a basis for E over K. It suffices to show that $\{x_i y_j\}_{i\in I, j\in J}$ is a basis for E over F. First, we show that it is spanning. Pick any $z \in E$. Then there exist $\alpha_j \in K$ such that $z = \sum_{j\in J} \alpha_j y_j$. Furthermore, there exist $\beta_{j,i} \in F$ such that $\alpha_j = \sum_{i\in I} \beta_{j,i} x_i$. Thus, we get

$$z = \sum_{i\in I, j\in J} \beta_{j,i} x_i y_j$$

as required. Next, we show linear independence. Suppose there exist $\alpha_{i,j} \in F$ such that

$$0 = \sum_{i\in I, j\in J} \alpha_{i,j} x_i y_j$$

Then we get

$$0 = \sum_{j \in J} \Big(\sum_{i \in I} \alpha_{i,j} x_i \Big) y_j$$

Each $\sum_{i \in I} \alpha_{i,j} x_i \in K$, so by the linear independence of the y_j we get that $\sum_{i \in I} \alpha_{i,j} x_i = 0$ for each $j \in J$. But the x_i are also linearly independent, and hence $\alpha_{i,j} = 0$ for all $i \in I, j \in J$. □

Note. We drop the underline notation for vectors here, as we want to emphasize that these are really field elements.

Corollary 8.1.3.1. *If $K/F, E/K$ are finite, then so is E/F.*

Now that we've got the basics out of the way, we can get to the namesake of this section.

Definition 8.1.4. Let K/F be a field extension. We say that $a \in K$ is algebraic over F if there exists a polynomial $p \in F[x]$ such that $p(a) = 0$. We say that K/F is an algebraic extension if every $a \in K$ is algebraic over F.

We'll next define a couple of related notions, which we will show in the end are all related.

Definition 8.1.5. Let K/F be a field extension, $a \in K$ algebraic over F. A minimal polynomial of a over F is a monic polynomial in $F[x]$ of minimal degree of which a is a root.

Proposition 8.1.6. *Minimal polynomials are unique and irreducible.*

Proof. Suppose $f \in F[x]$ satisfied $f = gh$, where $g, h \in F[x]$, and $f(a) = 0$. Then, without loss of generality, $g(a) = 0$, so either f is not irreducible or $h \in F$. In the first case we get $\deg(g) < \deg(f)$, and hence f cannot be a minimal polynomial of a. Thus, minimal polynomials must be irreducible. For uniqueness, suppose that $f, g \in F[x]$ were two minimal polynomials of a. Then they are both irreducible, and both have the same degree $n \in \mathbb{N}$. Since F is a field, $(f, g) = (h)$, where $h \in F[x]$ is the GCD of f and g. If $hk = f$, where $k \in F[x]$, then since f is irreducible either $h \in F$, in which case $(h) = F[x]$, or h is a unit multiple of f. The first case would imply that every polynomial in $F[x]$ has a as a root, which is impossible. Hence, h must be a unit multiple of both f and g, making f a unit multiple of g. Since f, g are both monic, we conclude that $f = g$. □

Since minimal polynomials are uniquely defined when they exist, we denote the minimal polynomial of a over F by $\min(a, F)$. Note that $a \in F$ if and only if $\min(a, F) = x - a$.

Definition 8.1.7. Let K/F be a field extension, and $a \in K$. We denote by $F(a)$ the smallest field contained in K containing F and a.

Note. It is not too hard to see that $F(a) \cong \mathrm{FF}(F[a])$.

Theorem 8.1.8. *Let K/F be a field extension, and $a \in K$. Then the following are equivalent.*

1. *a is algebraic over F.*
2. *$F(a) = F[a]$.*

3. $[F(a) : F]$ *is finite.*

Proof. First, suppose that a is algebraic over F. That $F[a] \subset F(a)$ is clear, so we just need to show that $F[a]$ is a field to prove the second statement. To that end, let $p = \min(a, F)$, and write $p = \sum_{i=0}^{n} \alpha_i x^i$. Since $p(a) = 0$, we conclude that any $z \in F[a]$ can be written in the form $z = \sum_{i=0}^{r} \beta_i a^i$ for some $\beta_i \in F$, where $r < n$, as $a^n = \sum_{i=0}^{n-1} \alpha_i x^i$. Assume $z \notin F$, as in that case finding an inverse for z is either trivial or $z = 0$. Since p is irreducible, no factor of $z(x) = \sum_{i=0}^{r} \beta_i x^i$ divides $p(x)$, and vice-versa. Thus, $\text{GCD}(p(x), z(x)) = 1$, so there exists some $h, k \in F[x]$ such that $p(x)h(x)+z(x)k(x) = 1 \Rightarrow p(a)h(a)+z(a)k(a) = 1 \Rightarrow z(a)k(a) = 1$, giving us an inverse for z. Thus, $F[a]$ is a field as claimed. The third statement follows from the second by noting that since p is of minimal degree, $\{1, a, \dots, a^{n-1}\}$ form a basis for $F[a]$. That the third statement implies the first is immediate, as otherwise $\{1, a, a^2, \dots\}$ would be linearly independent. □

Note. In particular, $[F(a) : F] = \deg(\min(a, F))$ when this is well-defined.

Corollary 8.1.8.1. *Let $E/K/F$ be a tower of algebraic extensions. Then E/F is algebraic.*

Proof. Pick any $a \in E$. Since a is algebraic over K, $p = \min(a, K)$ is well-defined. Write $p = \sum_{i=0}^{n} \alpha_i x^i$, where $\alpha_i \in K$. We get from this a tower of field extensions $F(\alpha_1, \dots, \alpha_n, a)/F(\alpha_1, \dots, \alpha_n)/F$, each of which is finite by Theorem 8.1.8. Since clearly $F(a) \subset F(\alpha_1, \dots, \alpha_n, a)$, it follows that $F(a)/F$ is a finite extension, and hence a is algebraic over F. □

Note. If K can be written in the form $F(a_1, \dots, a_n)$ for some $a_i \in K$, then we call K a finitely generated over F. It is clear that all finitely generated extensions generated by algebraic elements are both algebraic and finite. However, algebraic extensions need not be finite, or finitely generated.

There's one more way of combining fields we need to cover in this section.

Definition 8.1.9. Let F, K be subfields of a field E. The composition of F, K, denoted FK, is the smallest subfield of E containing F and K.

Note. If K, E are both overfields (i.e. fields containing) some F, then KE is the union of all fields of the form $F(\alpha_1, \dots, \alpha_n)$, where $\alpha_i \in K \cup E$.

Note. $[KE : F] = [K : F][E : F]$ does not, in general, hold. It does happen to hold if $K \cap E = F$.

Proposition 8.1.10. *If $K/F, E/F$ are algebraic and K, E are subfields of some common field, then KE/F is algebraic.*

Proof. Pick any $a \in KF$. By the above observation, there exist some $\alpha_1, \dots, \alpha_n \in K \cup E$ such that $a \in F(\alpha_1, \dots, \alpha_n)$. But since K, E are both algebraic extensions of F, each α_i is algebraic over F, and hence $[F(\alpha_1, \dots, \alpha_n) : F] < \infty$, making this extension algebraic over F. □

Proposition 8.1.11. *If K/F is a field extension, E/F an algebraic field extension, and E, K are subfields of a common field, then KE/K is algebraic.*

Proof. Pick $a \in KE$. We know there exist elements $\alpha_i \in K$ and $\beta_j \in E$ such that $a \in F(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m)$. Hence, $a \in K(\beta_1, \ldots, \beta_m)$. Since each β_i is algebraic over F, they are certainly algebraic over K, and hence $[K(\beta_1, \ldots, \beta_m) : K] < \infty$. □

The above three properties are very nice ones for field extensions to have, so we end off by giving this collection of properties a name.

Definition 8.1.12. Let $\mathcal{C}$ be a class of field extensions. We call $\mathcal{C}$ distinguished if

1. $K/F, E/K \in \mathcal{C} \iff E/F \in \mathcal{C}$.
2. $K/F, E/F \in \mathcal{C}$ and K, E both being subfields of a common field implies that $KE/F \in \mathcal{C}$.
3. $E/F \in \mathcal{C}$, K/F a field extension, and E, K being subfields of a common field implies that $KE/K \in \mathcal{C}$.

It is of course obvious from this definition that the class of algebraic (and finite) extensions is distinguished.

8.2 Splitting Fields and Algebraic Closures

This section, like the previous, is based off of notes from lectures by Dr. Sujatha Ramdorai and similar sections in [Lan05].

It should come as no surprise at this point in our studies that there is a deep connection between the roots of polynomials and algebraic field extensions, and we dedicate this section to studying this relationship. Let's start with the basics.

Definition 8.2.1. Let K/F be a field extension, and $p \in F[x]$. We say that p splits completely over K if p is a product of linear factors in $K[x]$.

Proposition 8.2.2. *Every $p \in F[x]$ splits completely over some field extension.*

Proof. It suffices to show that for any irreducible polynomial $p \in F[x]$ of degree at least two, we can find a field extension where p has a root. By the proof of Theorem 3.9.8, $F[x]/(p)$ is a field. But of course evaluating in $F[x]/(p)$, we get $p([x]) = 0$. □

The above proof is a neat little sleight of hand, and also a good reminder that $F[x]/(p)$ is a field if and only if f is irreducible. It's at this point that we start to use some category theory[1].

Definition 8.2.3. Let F be a field. The category of all field extensions of F, denoted $\mathbf{Extn}(F)$, is the category whose objects are fields extending F and morphisms field homomorphisms between extensions fixing F. We call these morphisms F-homomorphisms, and denote them in the form $\mathrm{Hom}_F(K, E)$.

[1] We don't have to, but the above chapter needs a purpose!

Note. One needs to be careful about what exactly they mean by "fixing" F. What this is really saying is that the embedding of F in K is mapped to the embedding of F in E, and the composite map $F \hookrightarrow K \to E \to F$ is the identity, where that last arrow is the inverse of the embedding $F \hookrightarrow E$.

Definition 8.2.4. Let F be a field. We say that F is algebraically closed if every polynomial in F of degree at least one splits completely over F.

Before proceeding, we take a quick detour.

Definition 8.2.5. Let R be a commutative ring. An ideal $I \subsetneq R$ is called maximal if there exists no ideal $J \subsetneq R$ such that $I \subsetneq J$.

Lemma 8.2.6. *Let R be a commutative ring and $I \subsetneq R$ an ideal. Then R/I is a field if and only if I is maximal.*

Proof. This follows immediately by noting that

1. Ideals in R containing I are in bijection with ideals in R/I.
2. R/I is a field if and only if its only two ideals are (0) and R/I.

□

Lemma 8.2.7. *Let R be a commutative ring and $I \subsetneq R$ an ideal. I is contained in a maximal ideal of R.*

Proof. Let Σ be the set of ideals in R containing I except R, with partial order imposed by inclusion. It is non-empty since $I \in \Sigma$. Let $\Omega \subset \Sigma$ be any totally ordered subset. It's clear that $I = \bigcup_{J \in \Omega} J$ is an ideal in R, in particular an upper bound of Ω. We wish to show that $I \in \Sigma$. If this were not the case, then $I = R \Rightarrow 1 \in I$. But this would imply that $1 \in J$ for some $J \in \Omega$, and hence that J is R, which is impossible. Hence, $I \in \Sigma$. The result then follows by Zorn's lemma. □

Note. This also implies, by taking $I = (0)$, that every commutative ring has a maximal ideal.

Alright, now back to field theory.

Proposition 8.2.8. *Every field F has an algebraically closed field extension.*

Proof. For each $p \in F[x]$ of degree at least one, we write a letter x_p. Let $S = \{x_p\}_{p \in F[x]}$[2], and consider the ring $F[S]$. Specifically, we look at the ideal $I \subset F[S]$ generated by $\{p(x_p)\}_{p \in F[x]}$. We first claim that $I \neq F[S]$. Indeed, suppose it were. Then there would exist finitely many distinct polynomials $p_i \in F[x], g_i \in F[S]$ such that $\sum_{i=1}^{n} g_i p_i(x_{p_i}) = 1$. By the above proposition, we can find a field extension K/F such that each p_i has some root $\alpha_i \in K$. Applying the evaluation homomorphism to $F[S]$ given by evaluating each x_{p_i} to α_i and all other variables to zero, we'd conclude that $0 = 1$, which is impossible. Hence, $I \neq F[S]$, and is therefore contained in some maximal ideal $M \subset F[S]$. $K_1 = F[S]/M$ is therefore a field in

[2]Abusing notation a bit here, technically it's not indexed over all of $F[x]$

which every polynomial $p \in F[x]$ of degree at least one has a root, in particular $[x_p]$. Applying this process iteratively, we get an infinite tower of field extensions $F \subset K_1 \subset K_2 \subset \cdots$, such that each polynomial in $K_i[x]$ of degree at least one has a root in K_{i+1}. Let $K = \bigcup_{i=1}^{n} K_i$. It is not too difficult to check that K is a field, and one in which F is embedded and every polynomial of degree at least one has a root. Since every polynomial in $K[x]$ of degree at least one has a root in K, it follows that every polynomial in $K[x]$ splits completely over K. □

Note. Algebraically closed fields have no proper algebraic extensions.

Corollary 8.2.8.1. *Every field F has an algebraic algebraically closed field extension.*

Proof. Let K be an algebraically closed extension of F, and let $\{K_i\}_{i\in I}$ be all the sub-fields of K which are algebraic extensions of F. Since the composite of two algebraic extensions is algebraic, $K' = \bigcup_{i\in I} K_i$ is an algebraic extension of F, and is a sub-field of K. Pick any $p \in K'[x]$. Every root a of p is in K, and a is by definition algebraic over K', so by the transitivity of algebraic extensions a is algebraic over F and hence $a \in K'$. Thus, K' is algebraically closed. □

We call such algebraic algebraically closed field extensions *algebraic closures* of F, and the above corollary guarantees their existence. We'll work now towards proving their uniqueness.

Lemma 8.2.9. *Let $\sigma : F \hookrightarrow L$ be a field embedding, where L is algebraically closed, and pick any root a of a polynomial in F such that $\alpha \notin F$. Then there exists an extension τ of σ to $F(a)$, i.e. a field embedding $\tau : F(a) \hookrightarrow L$ such that $\tau|_F = \sigma$. In particular, the number of such extensions is equal to the number of distinct roots of $\sigma(\min(a, F))$ in L.*

Proof. Let $p = \min(a, F)$. We know that $F(a) \cong F[x]/(p)$. Note that the field embeddings of $F[x]/(p)$ in L must all arise from field homomorphisms from $F[x]$ to L which are zero on (p). In particular, to extend σ, these homomorphisms must embed F in $\sigma(F)$. Every such homomorphism is therefore just an evaluation homomorphism, and is entirely specified by the image of x. In order for the homomorphism to be zero on (p), it must map x to a root of $\sigma(p)$ in L. □

Theorem 8.2.10. *Let $\sigma : F \hookrightarrow L$ be a field embedding, where L is algebraically closed, and let E/F be an algebraic field extension. Then there exists an extension τ of σ to E, i.e. a field embedding $\tau : E \hookrightarrow L$ such that $\tau|_F = \sigma$.*

Proof. Let S be the set of pairs (K, τ) where $E/K/F$ is a tower of algebraic field extensions and τ is an extension of σ to K. We can impose a partial order on this set by saying that $(K, \tau) \leq (K', \tau')$ if $K \subset K'$ and $\tau'|_K = \tau$. Note that $(F, \sigma) \in S$, so S is non-empty. Let $\Sigma \subset S$ be any totally ordered subset of S, say $\Sigma = \{(K_i, \tau_i)\}_{i\in I}$. Set $K = \bigcup_{i\in I} K_i$, and define $\tau : K \hookrightarrow L$ by saying $\tau|_{K_i} = \tau_i$. Note that K is a field, and τ is a well-defined field embedding extending σ, so $(K, \tau) \in S$. Furthermore, (K, τ) is clearly an upper bound for Σ. Thus, by Zorn's lemma there exists a maximal element $(K, \tau) \in S$. We just need to show now that $K = E$. If this were not the case, then we could find some $a \in E\backslash K$. By the previous lemma, τ extends to $K(a)$, and hence so does σ. But (K, τ) is maximal, so we conclude that $K(a) = K \Rightarrow a \in K$. □

Corollary 8.2.10.1. *Any two algebraically closed, algebraic extensions of a field F are F-isomorphic.*

Proof. Let $\sigma : F \hookrightarrow L$ be the "standard" embedding of F in an algebraically closed, algebraic extension of F. Let K be some other algebraically closed, algebraic extension of F. Then σ has a unique extension to K, which since K is algebraically closed must be bijective. □

Since we only care about field extensions up to F-isomorphism, we can therefore refer to such fields as the *algebraic closure* of F, which we denote $\overline{F}$.

Definition 8.2.11. A splitting field for a non-unit polynomial $p \in F[x]$ is an algebraic field extension of F over which p splits completely, and which can be embedded into any other such algebraic field extension.

Proposition 8.2.12. *Every polynomial has a splitting field, and that splitting field is unique up to F-isomorphism.*

Proof. We start with existence. That there exists an algebraic field extension in which p splits completely is clear. Now, suppose that K/F is any field extension over which p splits completely. Let $a_1, \ldots, a_n$ be the distinct roots of p in $\overline{F}$. By Theorem 8.2.10, there exists an extension of the standard embedding of F into $\overline{F}$ to K. By the proof of lemma 8.2.9, this extension must map the roots of p in K to roots of p in $\overline{F}$. Thus, we can invert to get an embedding of $F(a_1, \ldots, a_n)$ into K. This gives existence, uniqueness follows by noting that field homomorphisms can be viewed as linear maps over F. □

Again, this allows us to talk of the *splitting field* of a polynomial, which we denote F_p. We can extend this to a family of polynomials in the obvious way, and note that the same theorem applies as the splitting field of a family of polynomials $\{p_i\}_{i \in I}$ is just $\prod_{i \in I} F_{p_i}$.

8.3 Separable Extensions

This section is based off of notes from lectures by Dr. Sujatha Ramdorai, andsimilar sections in [Lan05],[Jac09].

Definition 8.3.1. A polynomial $p \in F[x]$ of degree at least one is separable if p has no repeated roots. If K/F is a field extension, we say that $a \in K$ is separable over F if it is algebraic over F and $\min(a, F)$ is separable. K/F is separable if every $a \in K$ is separable over F.

While this is the intuitive definition of separable extensions, it's actually not the quickest way to prove the things we want to about separable extensions. Therefore, we instead focus on the *separable degree* of an extension. For this, we will need another result on algebraic closures.

Theorem 8.3.2. *Let E/F be an algebraic extension, and $\sigma : F \hookrightarrow L$ be an embedding of F in the algebraic closure L of $\sigma(F)$. Let $\tau : F \hookrightarrow L'$ be an embedding of F in the algebraic closure L' of $\tau(F)$. Set S_σ to be the number of extensions of σ to E, and define S_τ similarly. Then $|S_\sigma| = |S_\tau|$.*

Proof. Consider the isomorphism $\sigma \circ \tau^{-1}$. By Theorem 8.2.10, this extends to an isomorphism $\gamma : L' \to L$. Now, pick any extension ω of τ to E. Then $\gamma \circ \omega$ is an extension of σ to E, and since γ is bijective distinct extensions of τ lead to distinct extensions of σ under this mapping. Furthermore, composing by γ^{-1}, we can see that any such extension of σ arises in this manner. □

Because of this, $|S_\sigma|$ depends only on the field extension E/F. We can then define the separable degree of the extension, denoted $[E : F]_s$, to be this cardinality. Let's get some basic results on separable degree.

Proposition 8.3.3. *Let $E/K/F$ be a tower of algebraic extensions. Then*

$$[E : F]_s = [E : K]_s[K : F]_s$$

and if $[E : F]$ is finite then $[E : F]_s \leq [E : F]$.

Proof. Starting with the first statement, let L be an algebraic closure of $\sigma(F)$, where $\sigma : F \hookrightarrow L$ is an embedding. The first statement follows by noting that each extension of σ to E can be written as an extension of σ to K, then extended to E, and that the number of such second extensions is independent of the extension of σ chosen (as it is always an embedding). For the second, we write can find some $\alpha_i \in E$ such that

$$F \subset F(\alpha_1) \subset \cdots \subset F(\alpha_1, \ldots, \alpha_n) = E$$

Is a tower of field extensions. By lemma 8.2.9, each of these field extensions satisfies $[F(\alpha_1, \ldots, \alpha_m) : F(\alpha_1, \ldots, \alpha_{m-1})]_s \leq [F(\alpha_1, \ldots, \alpha_m) : F(\alpha_1, \ldots, \alpha_{m-1})]$. The result then follows by multiplicativity of separable and usual degree. □

It turns out that there's really no difference between separable degree and separable extensions. To do this we'll need to be a little careful, and start with finite extensions.

Theorem 8.3.4. *Let K/F be a finite algebraic field extension. Then*

1. *K/F is separable if and only if $[K : F]_s = [K : F]$.*
2. *$a \in K$ is separable over F if and only if $F(a)/F$ is separable.*

Proof. Note that by the same argument used to prove that $[K : F]_s \leq [K : F]$, we need only show that the first statement holds for $K = F(a)$, for some $a \in K$. Suppose $[F(a) : F]_s = [F(a) : F]$. By lemma 8.2.9, $[F(a) : F]_s$ is equal to the number of distinct roots of $\min(a, F)$, and we know that $[F(a) : F] = \deg(\min(a, F))$. Thus, $\min(a, F)$ must be separable, and hence a is separable over F. That a is separable over F implies that $[F(a) : F]_s = [F(a) : F]$ is also clear from this. Thus, we just need to show that $a \in K$ being separable over F implies that $F(a)/F$ is separable. To that end, pick any $b \in K$. Then we get a tower of extensions $F(a)/F(b)/F$, which satisfy

$$[F(a) : F] = [F(a) : F(b)][F(b) : F], [F(a) : F]_s = [F(a) : F(b)]_s[F(b) : F]_s$$

If a is separable over F then it certainly is over $F(b)$ as well, so it follows from this that $[F(b) : F] = [F(b) : F]_s$, and hence b is separable over F. □

To extend this equivalence to infinite extensions, we can use the following result.

Proposition 8.3.5. *Let K/F be an algebraic field extension. Then it is separable if and only if for every finite sub-extension L/F, where $L \subset K$, L/F is separable.*

Proof. This follows immediately by noting that any $a \in K$ is in some finite sub-extension, namely $F(a)/F$, and that K is the union of all finite sub-extensions. □

We'll next prove that separable extensions are a distinguished class, but for this we'll first need a quick lemma.

Lemma 8.3.6. *Suppose K/F is a field extension, and $a \in K$ is algebraic over F. Then $\min(a, F)$ divides every polynomial in $F[x]$ of which a is a root.*

Proof. Suppose this were not the case. Set $p = \min(a, F)$, and let $g \in F[x]$ be a polynomial of which a is a root such that $p \nmid g$. Then clearly $\deg(\mathrm{GCD}(p, g)) < \deg(p)$. But there exists $k, h \in F[x]$ such that $pk + gh = \mathrm{GCD}(p, g)$, and hence $\mathrm{GCD}(p, g)$ has a as a root. This contradicts the minimality of $\deg(p)$. □

Theorem 8.3.7. *Separable extensions of a field F are a distinguished class of extensions.*

Proof. First, suppose that $E/K/F$ is a tower of separable field extensions. Pick any $a \in E$, and let $\min(a, K) = \sum_{i=0}^{n} \alpha_i x^i$. Then $F(a, \alpha_1, \ldots, \alpha_n)/F(\alpha_1, \ldots, \alpha_n)/F$ is a tower of finite algebraic extensions, each of which is separable. That $F(a, \alpha_1, \ldots, \alpha_n)/F$ is separable, and hence a is separable over F, follows by the Theorem 8.3.4 and proposition 8.3.3. Thus, E/F is separable. Now, suppose that E/F is separable, and pick any $a \in E$. By lemma 8.3.6, a is separable over K, and hence E/K is separable. K/F being separable is immediate.

Next, suppose that K, E are subfields of a common field, K/F is separable and E/F is a field extension. By lemma 8.3.6, every element of K is separable over E, and hence for any $a \in K$ we get that $E(a)/E$ is separable. Repeating this argument, we conclude that by proposition 8.3.5 that EK/E is separable. If E/F is also separable, then by the first part of the proof EK/F is separable. □

We now move on to studying when such separable extensions exist. Let's begin with the notion of a separable closure.

Definition 8.3.8. Let F be a field, and $F \subset K \subset \overline{F}$ a field extension. We say that K is separably closed if every separable extension of F can be F-embedded in K.

Proposition 8.3.9. *Every field F has a separably closed extension, which is unique up to F-isomorphism.*

Proof. For existence, we simply note that by Theorem 8.3.7 the composition of all separable extensions of F is separable. The uniqueness is then clear from the definition. □

We call this "unique" field the separable closure of F, denoted F^{sep}. We'll come back to these later.

Definition 8.3.10. A field F is called perfect if every algebraic extension of F is separable.

It turns out that most fields are perfect! In order to show this, we'll need to take a bit of a detour.

Definition 8.3.11. Let R be a ring, and $p(x) = \sum_{i=0}^{n} a_i x^i$ a polynomial in $R[x]$. The derivative of the polynomial is

$$p'(x) = \sum_{i=1}^{n} i a_i x^{i-1}$$

where we define multiplication by integers using the standard ring inclusion $\mathbb{Z} \hookrightarrow R$ given by mapping 1 in $\mathbb{Z}$ to 1 in R.

It is not too hard to verify that R is a linear map from $R[x]$ to itself, just as we'd expect, and satisfies the usual product rule from calculus. It also has some more properties that we'd expect.

Proposition 8.3.12. *Suppose F is algebraically closed. Then $p(x) \in F[x]$ of degree at least 1 is separable if and only if $GCD(p, p') = 1$.*

Proof. The above statement is equivalent to saying that p is separable if and only if p, p' share no roots. Indeed, since F is algebraically closed we can factor p to get

$$p(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n)$$

where $\alpha_i \in F$ are the (possibly non-distinct) roots of p. This, in turn, gives

$$p'(x) = \sum_{i=1}^{n} \prod_{j \neq i} (x - \alpha_j)$$

If α_i is not a repeated root, then

$$p'(\alpha_i) = \prod_{j \neq i} (\alpha_i - \alpha_j) \neq 0$$

Otherwise, there is a factor $(x - \alpha_i)$ in each term of the sum, and hence α_i is a root of p'. The result follows from this. □

Lemma 8.3.13. *Suppose F is a field, K/F a field extension, and $p, q \in F[x]$. Then $GCD(p, q)$ is the same in $F[x]$ and $K[x]$.*

Proof. Suppose $(GCD)(p, q) = r$ in $F[x]$. Let $k, h \in F[x]$ be the polynomials such that $p = kr, q = hr$. It suffices to show that k, h are coprime in $K[x]$. Hence, it suffices to show that the leamma holds in the case where $\mathrm{GCD}(p, q) = 1$. Assume this. Then there exist $h, k \in F[x]$ such that $ph + qk = 1$. Thus, $(p, q) = (1)$ in $K[x]$ as well, completing the proof. □

Corollary 8.3.13.1. *$p \in F[x]$ is separable if and only if $GCD(p, p') = 1$.*

Definition 8.3.14. Let F be a field. The character of a field, denoted $\text{char}(F)$, is the smallest $n \in \mathbb{N}$ such that $\sum_{i=1}^{n} 1 = 0$ in F. If no such F exists, we write $\text{char}(F) = 0$.

Proposition 8.3.15. *If* $\text{char}(F) \neq 0$*, then* $\text{char}(F)$ *is prime.*

Proof. Let $\varphi : \mathbb{Z} \hookrightarrow F$ be the standard ring inclusion given by mapping 1 in $\mathbb{Z}$ to 1 in F. Note that $n = \text{char}(F)$ if and only if $\ker(\varphi) = (n)$. Also, one can note that $\varphi(\mathbb{Z})$ is a subfield of F. Since $\varphi(\mathbb{Z}) \cong \mathbb{Z}/(n)$, (n) must be maximal and hence n must be prime. □

Proposition 8.3.16. *If* $\text{char}(F) = 0$*, then* F *is perfect.*

Proof. It suffices to show that every irreducible polynomial $f \in F[x]$ is separable. Indeed, write $f(x) = \sum_{i=0}^{n} a_i x^i$. Then $\deg(f') < \deg(f)$, $f' \neq 0$, and hence since f is irreducible it must be coprime to f'. The result then follows from lemma 8.3.13. □

Lemma 8.3.17. *If* $\text{char}(F) = p$ *and* $a \in F$*, then* $x^p - a$ *is either irreducible or has one root of multiplicity* p *in* F*.*

Proof. Let $f(x) = x^p - a$, and let $b \in \overline{F}$ be any root of f. Suppose that f is reducible. Since $b^p = a$, $(x - b)^p = x^p - a$. Thus, the only possible factors of f are powers of $(x - b)$. In particular, we can assume without loss of generality that $\exists 0 < k < p$ such that $b^k \in F$. Furthermore, since p is prime, there exists $q, r \in \mathbb{Z}$ such that $qk + pr = 1$. Thus, since $(b^k)^a (b^p)^r \in F$, we conclude that $b \in F$, and hence $f(x) = (x - b)^p$ in $F[x]$. □

Proposition 8.3.18. *If* $\text{char}(F) = p$*, then* F *is perfect if and only if* $F^p = F$*.*

Proof. If F is perfect, then every irreducible polynomial in F must be separable. Hence, by the above lemma every polynomial of the form $x^p - a$ cannot be irreducible. Thus, there exists for every $a \in F$ some $b \in F$ such that $b^p = a$, and therefore $F^p = F$. Now, suppose that $F^p = F$, and pick any irreducible $f \in F[x]$. Suppose f is not separable. Then $\text{GCD}(f, f') \neq 1$, and hence by the proof of proposition 8.3.16 every non-zero term in f must be of the form $a_{kp} x^{kp}$, where $k \in \mathbb{N}$. Write

$$f(x) = \sum_{k=0}^{n} a_{kp} x^{kp}$$

Suppose for each a_{kp} there exists a $b_k \in F$ such that $b_k^p = a_{kp}$. Then we get

$$f(x) = \sum_{k=0}^{n} b_k^p x^{kp} = \Big(\sum_{k=0}^{n} b_k x^k \Big)^p$$

which again is impossible. Thus, there exists some $a_{kp} \in F$ which is not in F^p, a contradiction, and hence f must be separable. □

Corollary 8.3.18.1. *If* $\text{char}(F) = p$*, then for every irreducible non-trivial* $f \in F[x]$ *there exists an irreducible, separable* $g \in F[x]$ *and some* $k \in \mathbb{N} \cup \{0\}$ *such that* $f(x) = g(x^{p^e})$*.*

Proof. If f is separable, then we're done. Otherwise, by the proof of the previous proposition, we know that f can be written in the form

$$f(x) = \sum_{k=0}^{n} a_k x^{kp}$$

Set $g(x) = \sum_{k=0}^{n} a_k x^k$. Then $f(x) = g(x^p)$. Repeating this process, we get the desired result. □

Note. Not every field is perfect!

Definition 8.3.19. Let K/F be a field extension. The extension is called simple if there exists some $a \in K$ such that $K = F(a)$. In this case, we call a a primitive element of K over F.

Theorem 8.3.20. *A finite field extension K/F is simple if and only if there exist a finite number of subfields of K containing F.*

Proof. First, suppose that F is finite. Then by the Theorem 3.10.2 $K^\times$ is cyclic, and from which the desired result immediately follows. Thus, we assume that F is infinite.

Suppose that K has only a finite number of subfields containing F. We proceed by induction to prove the claim that, for any $a_1, \ldots, a_n \in K$, there exist $c_2, \ldots, c_n \in F$ such that

$$F(a_1, \ldots, a_n) = F(a_1 + c_2 a_2 + \cdots + c_n a_n)$$

The base case of $n = 1$ is clear. Now, pick any $a_1, \ldots, a_n \in K$, and assume that the result holds for $n - 1$. Then there exist $c_2, \ldots, c_{n-1} \in F$ such that

$$F(a_1, \ldots, a_{n-1}) = F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1})$$

Furthermore, the set $\{F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} + c_n a_n)\}_{nc_n \in F}$ must be finite. Thus, there exist distinct $c_n, c_n' \in F$ such that

$$F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} + c_n a_n) = F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} + c_n' a_n)$$

In particular, this implies that

$$a_n(c_n - c_n') \in F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} + c_n a_n)$$

and hence

$$a_n, a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} \in F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} + c_n a_n)$$

Thus, we conclude that

$$F(a_1, \ldots, a_n) = F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1}, a_n) = F(a_1 + c_2 a_2 + \cdots + c_{n-1} a_{n-1} + c_n a_n)$$

as required. Since K/F is finite, it follows from this that K/F is simple.

Now, suppose K/F is simple, say $K = F(a)$. Let $p = \min(a, F)$, and let E be an extension of F contained in K. Let $q = \min(a, E)$. Then by lemma 8.3.6, $q \mid p$, so since $E[x]$ is a UFD there exists a unique $h \in E[x]$ such that $p = qh$. Any polynomial dividing p must be equal to a product of expressions of the form (x, α), where α is a root of p, and there are only finitely many such polynomials. Hence, we get a finite number of options for q. But $K = E(a)$, so $E \cong K[x]/(q)$, and hence there are only finitely many such E. □

Theorem 8.3.21 (Primitive Element)**.** *Every finite separable extension of a field F is simple.*

Proof. We can note from the above proof that every finite extension of a finite field is simple, and hence assume that F is infinite. Using the same argument as in the infinite case above, it suffices to show that if K/F is separable and finite and $a, b \in K$, then $F(a, b)$ is simple. If $F(a, b) = F$ this is obvious, so assume this is not the case. Let $\sigma_1, \ldots, \sigma_n$ be the F-embeddings of $F(a, b)$ in F^{sep}. Note that by separability and lemma 8.2.9, $n = [F(a, b) : F]$. Define a polynomial $p \in F[x]$ by

$$p(x) = \prod_{i \neq j} (\sigma_i(a) + \sigma_i(b)x - \sigma_j(a) - \sigma_j(b)x)$$

Note that since $n \geq 2$, $p \neq 0$, and hence there exists by Theorem 3.10.1 some $c \in F$ such that $p(c) \neq 0$. That is, each $\sigma_i(a + cb)$ must be distinct, so by lemma 8.2.9 $[F(a + cb) : F] = n$. Thus, since $a + cb \in F(a, b)$, $F(a, b) = F(a + cb)$. □

8.4 Normal Extensions

This section is based off of notes from lectures by Dr. Sujatha Ramdorai, along with a similar section from [Lan05].

Definition 8.4.1. An algebraic field extension K/F is called normal if every irreducible polynomial of $F[x]$ with a root K has all of its roots in K.

Normality turns out not to be transitive (and hence Normal extensions do not form a distinguished class), but we do get a slightly weaker condition.

Proposition 8.4.2. *Let $L/K/F$ be a tower of algebraic extensions. If L/F is normal, then L/K is normal.*

Despite this, it turns out that Normal extensions are intimately connected with splitting fields.

Theorem 8.4.3. *Let K/F be an algebraic field extension, where $K \subset \overline{F}$. Then the following are equivalent.*

1. *K/F is normal.*
2. *K is the splitting field of a family of polynomials in $F[x]$.*
3. *Every embedding $K \hookrightarrow \overline{F}$ induces an automorphism of K.*

Proof. First, suppose that (iii) holds. Pick any $a \in K$, and let $p = \min(a, K)$. If $b \in \overline{F}$ is a root of p, then there exists an F-isomorphism $\tau : F(a) \to F(b)$ such that $\tau(a) = b$ (as p is irreducible). By Theorem 8.2.10, this can be extended to an F-embedding $\tau : K \hookrightarrow \overline{F}$, which by (iii) must be an isomorphism on K. Hence, every root of p must be in K, and so K/F is normal. Now, suppose that K/F is normal. Then K is clearly the splitting field of all irreducible polynomials in $F[x]$ with a root in K, so (ii) holds. Finally, suppose that K is the splitting field of $\{f_i\}_{i\in I} \subset F[x]$. Let $\tau : K \hookrightarrow \overline{F}$ be an embedding. Then $\tau(K)$ is another splitting field of $\{f_i\}_{i\in I}$ so by the infinite version of proposition 8.2.12 $\tau : K \to \tau(K)$ is an F-isomorphism. Since $K, \tau(K) \subset \overline{F}$ and thus contain the same roots of polynomials in our family, it follows that τ is an automorphism. □

Using this, we can show that normal extensions aren't that poorly behaved.

Proposition 8.4.4. *Let $K/F, E/F$ be a pair of normal extensions, where $K, E \subset L$ are contained in some common field. Then KE/F and $(K \cap E)/F$ are both normal.*

Proof. By Theorem 8.4.3, each of K, E is the splitting field of a family of polynomials in $F[x]$. KE is just the splitting field of the union of those two families, and hence is normal over F. Now suppose that $f \in F[x]$ has a root in $K \cap E$. Then it has a root in K, so all of its roots are in K, and the same with E. Hence, all of its roots are in $K \cap E$. □

There's one last result I'd like to mention here which can be useful.

Proposition 8.4.5. *Every algebraic extension is contained in a normal extension.*

Proof. Suppose K/F is algebraic. Then the desired normal extension is just the splitting field of $\{\min(a, F)\}_{a\in K}$. □

8.5 Purely Inseparable Extensions

This section is based off of lectures by Dr. Sujatha Ramdorai and a similar section in [Lan05].

We have already spent substantial time characterizing separable extensions. In this section, we try to expand this to include all algebraic extensions, and will discover that algebraic extensions are formed by a separable and *purely inseparable* extension. Since every field of characteristic zero is perfect, we assume for this section that our fields are not perfect and are of characteristic p.

Definition 8.5.1. Let K/F be an algebraic extension. We call $a \in K$ purely inseparable over F if there exists some $n \in \mathbb{N}$ such that $a^{p^n} \in F$. K/F is purely inseparable if every $a \in K$ is purely inseparable over F.

Theorem 8.5.2. *Let F be a field, and $f \in F[x]$ an irreducible polynomial. Then all the roots of f have the same multiplicity, which is p^k for some $k \in \mathbb{N} \cup \{0\}$ if $\mathrm{char}(F) = p$.*

Proof. By corollary 8.3.18.1 there exists some separable irreducible $g \in F[x]$ and $k \in \mathbb{N}\cup\{0\}$ such that $f(x) = g(x^{p^k})$. Thus, it suffices to consider polynomials of the form $x^{p^k} - a$, where $a \in \overline{F}$. But in $\overline{F}$, $x^{p^k} - a = (x - a^{1/p^k})^{p^k}$, so our single root a^{1/p^k} has multiplicity p^k. □

Corollary 8.5.2.1. *Let K/F be an algebraic extension, and pick any $a \in K$. Let $f = \min(a, F)$, and let $g \in F[x]$ be the separable irreducible polynomial such that $f(x) = g(x^{p^k})$, for some $k \in \mathbb{N} \cup \{0\}$. Then $[F(a) : F]_s = \deg(g)$, $[F(a) : F] = p^k[F(a) : F]_s$, and a^{p^k} is separable over F.*

Proof. If $a \in F$ this is trivial, so assume that $a \notin F$. By lemma 8.2.9, $[F(a) : F]_s$ is equal to the number of distinct roots of f, which is precisely $\deg(g)$. Thus, since $[F(a) : F] = \deg(f)$, we get the first two results. The third is just a restatement of Theorem 8.5.2. □

The above result also shows that $[F(a) : F]_s \mid [F(a) : F]$, and hence by the multiplicity of separable and total degree we get $[K : F]_s \mid [K : F]$ for any finite extension K/F. We call $[K : F]/[K : F]_s = [K : F]_i$ the *inseparable degree*, which it follows is also multiplicative. It is also clear that K/F is separable if and only if $[K : F]_i = 1$.

Theorem 8.5.3. *Let K/F be an algebraic field extension. Then the following are equivalent.*

1. *$[K : F]_s = 1$.*
2. *K/F is purely inseparable.*
3. *Every $a \in K$ has a minimal polynomial of the form $x^{p^k} - b$ in $F[x]$.*
4. *K is generated by a set of purely inseparable generators over F.*

Proof. First, suppose that (i) holds. Then in particular we must get $[F(a) : F]_s = 1$ for any $a \in K$, and hence the minimal polynomial of a over F is of the form $x^{p^k} - b$, where $b \in F$. Thus, $a^{p^k} \in F$ and K/F is purely inseparable. Now, suppose that K/F is purely inseparable. Then for any $a \in K$, there exists some $b \in F$ and minimal $k \in \mathbb{N} \cup \{0\}$ such that $p(x) = x^{p^k} - b$ has a as a root. Hence, $\min(a, F) \mid p$. But by the proof of Theorem 8.5.2, p has only one root, and hence by corollary 8.5.2.1 we get $\deg(\min(a, F)) = p^k$. Thus, $\min(a, F) = x^{p^k} - b$, proving (iii). Suppose that (iii) holds. Then clearly every $a \in K$ is purely inseparable over F, and hence any set of generators for K will be, so (iv) holds. Finally, suppose that (iv) holds. Let $\{a_i\}_{i \in I}$ be a set of purely inseparable generators for K over F. Any extension of an embedding of F in $\overline{F}$ to K is fully specified by where it sends each a_i. But each a_i must be sent to another root of $\min(a_i, F)$, and by a previous part of this proof each of these has one root, so there is exactly one such extension and $[K : F]_s = 1$. □

Proposition 8.5.4. *The class of purely inseparable extensions of a field F is distinguished.*

Proof. The condition on towers follows immediately from the multiplicativity of inseparable/separable degrees and (i) in Theorem 8.5.3. Now, suppose that K/F is an inseparable extension and E/F a field extension. By Theorem 8.5.3, there exist a set of purely inseparable generators $\{a_i\}_{i \in I}$ for K over F. Since $KE = E(\{a_i\}_{i \in I})$, it suffices to show that each a_i is purely inseparable over E. By Theorem 8.5.3, $\min(a_i, F) = x^{p^k} - b$ for some $b \in F$. Then since $\min(a_i, E) \mid \min(a_i, F)$, $\min(a_i, E)$ has exactly one root. By Theorem 8.5.2 that root must have multiplicity $p^{k'}$ for some $k' \in \mathbb{N} \cup \{0\}$, and hence $\min(a_i, E)$ is of the desired form. □

Theorem 8.5.5. *Let K/F be an algebraic field extension. Let L be the composite of all subfields of K which are separable extensions of F. Then L/F is separable, and K/L is purely inseparable.*

Proof. That L/F is separable is immediate from separable extensions being a distinguished class. Now, suppose that $a \in K$. By assumption, a cannot be separable over L. Let g be the separable, irreducible polynomial in $L[x]$ such that $\min(a, L)(x) = g(x^{p^k})$. Then a^{p^k} is a root of g in K, and hence since g is separable $a^{p^k} \in L$. Thus, K/L is purely inseparable. □

Corollary 8.5.5.1. *An algebraic extension K/F is separable and purely inseparable if and only if $K = F$.*

Corollary 8.5.5.2. *Let K, L be finite extensions of F, with K/F purely inseparable, L/F separable, and L, K subfields of a common field. Then*

$$[KL : K] = [L : F] = [KL : F]_s$$
$$[KL : L] = [K : F] = [KL : F]_i$$

Proof. KL/K is separable, and $K : F$ is purely inseparable, so $[KL : F]_s = [L : F] = [KL : K]$. Furthermore, L/F is separable and KL/L is purely inseparable, so $[KL : L] = [K : F] = [KL : F]_i$. □

It is worth asking whether our decomposition can be done in the other way, that is whether we can decompose an algebraic field extension into a purely inseparable followed by separable extension. The answer is, in general, no, but there is a special case where we can do this which can be found in [Lan05].

8.6 Finite Fields*

This section closely follows the same section in [Lan05], which I find to be an excellent, systematic, and thorough piece of exposition.

The point of this section is to demonstrate that we can precisely describe, up to isomorphism, all finite fields. This isn't explicitly needed in any of the later parts of this book, but it's the sort of information that tends to crop up and be useful when you least expect it to.
As was observed in section 8.3, every field of characteristic p can have the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ embedded into it, and hence every finite field of characteristic p is an extension of $\mathbb{F}_p$. In fact, it turns out that these extensions are generated by a very select set of polynomials.

Theorem 8.6.1. *Let p be a prime number, and $n \geq 0$. Then the subset of $\overline{\mathbb{F}_p}$ consisting of roots of the polynomial $x^{p^n} - x$ is a field, in particular the splitting field of $x^{p^n} - x$, and is of size p^n. Furthermore, any finite field of character p is isomorphic to some such splitting field.*

Proof. We start by showing that the subset $S \subset \overline{\mathbb{F}_p}$ consisting of the roots of the polynomial $x^{p^n} - x$ is a field. That $0, 1 \in S$ is clear. Now, suppose that $a, b \in S$. Then $(ab)^{p^n} = a^{p^n} b^{p^n} =$

ab, so $ab \in S$. Furthermore, $(a+b)^{p^n} = a^{p^n} + b^{p^n} = a + b$, so $a + b \in S$. We finish off by noting that if $a \neq 0$, $(a^{-1})^{p^n} = (a^{p^n})^{-1} = a^{-1}$ and $(-a)^{p^n} = (-1)^{p^n} a^{p^n} = -a$, so multiplicative and additive inverses are present. Thus, S is a field. Furthermore, if $f(x) = x^{p^n} - x$ then $f'(x) = -1$, and hence f is separable, from which it follows that $|S| = p^n$ as claimed. Finally, suppose that K is any other field of character p. Since K is a vector space over $\mathbb{F}_p$, $|K| = p^n$ for some $n \in \mathbb{N} \cup \{0\}$. Furthermore, $a^{p^n - 1} = 1$ for any non-zero $a \in K$. Thus, K is isomorphic to the splitting field of $x^{p^n} - x$. □

We denote this unique (up to isomorphism) field of order p^n by $\mathbb{F}_{p^n}$. We also get the fairly immediate consequences

Corollary 8.6.1.1. *Any degree m extension of $\mathbb{F}_{p^n}$ is isomorphic to $\mathbb{F}_{p^{nm}}$. Furthermore, every finite extension of a finite field is normal.*

The last thing of interest here is that finite fields have a very simple structure on their group of automorphisms. Indeed, let K be a finite field of characteristic p. We define the *Frobenius Automorphism* on K by $\varphi(a) = a^p$. Note that this is bijective since K is finite and φ is an injective vector space homomorphism, and hence φ is indeed a field automorphism. There's actually an immediate consequence of this observation as well.

Proposition 8.6.2. *Every finite field is perfect.*

However, this isn't the point at the moment. We're more interested in the following result.

Theorem 8.6.3. *Let F be a finite field of characteristic P. Then the group of F-automorphisms of the degree-n extension K of F is cyclic of order n, and generated by φ^n.*

Proof. Pick any $\tau \in \mathrm{Aut}_F(K)$. Since F is the splitting field of $x^{p^m} - x$, where $|F| = p^m$, τ must fix every root of $x^{p^m} - x$. Since $K^\times$ is cyclic, τ is entirely defined by where it sends a generator of $K^\times$. That is, if $a \in K$ is our generator, we get $\tau(a) = a^r$ for some $r \in \mathbb{N} \cup \{0\}$. Thus, $\tau(b) = b^r$ for any $b \in K$. But of course $F^\times$ is cyclic of order $p^m - 1$, so since τ fixes F we require $r = km$ for some $k \in \mathbb{N}$. Furthermore, since $K^\times$ is cyclic of order $p^{nm} - 1$ we get that each p^{km} produces a different automorphism for $1 \leq k \leq n$, and $\tau = \mathrm{Id}_K$ for $k = n$, from which the result follows. □

8.7 Galois Extensions

This section is based off of lectures by Dr. Sujatha Ramdorai and a similar section in [Lan05].

Let's start with our two fundamental definitions.

Definition 8.7.1. A field extension K/F is Galois if it is algebraic, separable, and normal.

Note. Since normal extensions are not a distinguished class, neither are Galois extensions.

Definition 8.7.2. The Galois group of a field extension K/F is the group

$$\mathrm{Gal}(K/F) = \mathrm{Aut}_F(K)$$

Note. If K/F is algebraic, then $\mathrm{Gal}(K/F)$ can also be viewed as all extensions of the standard embedding of F into $\overline{F}$ to K, that is all F-embeddings of K in $\overline{K}$.

Galois theory is all about relating the properties of field extensions to those of their Galois groups. Doing so requires building up a lot of machinery, which we begin to do now.

Definition 8.7.3. Let F be a field and G a subgroup of $\mathrm{Aut}(F)$. The subset of F fixed by G, which we denote F^G, is called the fixed field of G.

It is not too difficult to show that the fixed field is, in fact, a field.

Proposition 8.7.4. *Let K/F be a Galois extension, and set $G = \mathrm{Gal}(K/F)$. Then $F = K^G$. Furthermore, the map $L \to \mathrm{Gal}(K/L)$ from intermediate field extensions $F \subset L \subset K$ to subgroups of G is injective.*

Proof. Suppose that $a \in K^G$. Since $F \subset F(a) \subset K$ and K/F is separable, $F(a)/F$ is separable. That is, $[F(a) : F]_s = [F(a) : F]$. But $F(a)$ cannot have a non-trivial F-automorphism, as this could then be extended to an F-automorphism of K not fixing a, and hence $[F(a) : F]_s = 1 \Rightarrow a \in F$. That $F \subset K^G$ is clear, so $F = K^G$.

Now, suppose that $F \subset L \subset K$ is an intermediate extension. Then K/F being normal, separable, and algebraic implies that K/L is as well, and hence K/L is Galois. Thus, if $H = \mathrm{Gal}(K/L) \subset G$, then $K^H = L$. Injectivity follows from this. □

Note. We say that H belongs to L if $H = \mathrm{Gal}(K/L)$, and that L is associated to H.

Corollary 8.7.4.1. *Let K/F be Galois, with $\mathrm{Gal}(K/F) = G$, and let L, E be two intermediate field extensions with subgroups H, R belonging to them respectively. Then*

1. *$H \cap R$ belongs to LE.*
2. *The fixed field of the smallest subgroup of G containing H, R is $L \cap E$.*
3. *$L \subset E$ if and only if $R \subset H$.*

Proof. The first item follows by noting that K/LE is Galois, if an automorphism of K fixes L and E then it fixes LE, and if an automorphism of K fixes LE then it fixes L and E. The second comes from noting that any element of a subgroup generated by H, R must fix $L \cap E$, and vice-versa. The third is also immediate. □

Proposition 8.7.5 (Artin)**.** *Let K be a field and G a finite group of automorphisms of K. Then K is a finite Galois extension of K^G and $|G| = [K : K^G]$.*

Proof. Pick any $a \in K$. Let $\sigma_1, \ldots, \sigma_n \in G$ be a maximal set such that $\sigma_1(a), \ldots, \sigma_n(a)$ are distinct. Pick any other $\tau \in G$. Then since τ is injective, $(\tau \circ \sigma_1)(a), \ldots, (\tau \circ \sigma_n)(a)$ must be a permutation of $\sigma_1(a), \ldots, \sigma_n(a)$, as otherwise $\sigma_1, \ldots, \sigma_n$ would not be minimal. Define a polynomial

$$p(x) = \sum_{i=1}^{n} (x - \sigma_i(a))$$

Then p is fixed under G, and hence $p \in K^G[x]$. Furthermore, p is separable, and has a as a root as otherwise $\mathrm{Id}_K(a)$ would be distinct contradicting the minimality of the $\sigma_i(a)$. Since a was arbitrary, it follows that K/K^G is Galois with Galois group G. As K/K^G is separable, G being finite further implies that K/K^G is finite. K/K^G is therefore simple, and hence by the Dedikind's lemma $[K : K^G] = |G|$. □

Corollary 8.7.5.1. *Let K/F be a finite Galois extension. Then each subgroup of $\mathrm{Gal}(K/F)$ is associated to an intermediate extension of K/F.*

Proposition 8.7.6. *Let K/F be a normal extension, and set $G = \mathrm{Gal}(K/F)$. Let L be an intermediate field extension, and set $H = \mathrm{Gal}(K/L)$. Then*

1. *$H \trianglelefteq G$ if and only if L/F is normal.*
2. *If L/F is normal, then $\mathrm{Gal}(L/F) \cong G/H$.*

Proof. 1. First, suppose that L/F is normal. Then L is the splitting field of a family of polynomials in $F[x]$, and any F-automorphism of L can be specified entirely by permuting the roots of these polynomials. Pick any $\sigma \in H, \tau \in G$. Since τ fixes F, it fixes any polynomial in $F[x]$. Thus, τ can only act on L by permuting roots in that same family of polynomials, making τ restrict to an F-automorphism of L. It follows that $\tau \circ \sigma \circ \tau^{-1} \in H$, and hence $H \trianglelefteq G$. Now, suppose that $H \trianglelefteq G$. Let $\mathcal{F} = \{\min(a, F)\}_{a \in L}$. Our goal will be to show that L is the splitting field of $\mathcal{F}$, and hence that L/F is normal. It is clear that any field over which $\mathcal{F}$ splits must contain L, so we need only show that $\mathcal{F}$ splits over L. Suppose this were not the case, that is there existed some $a \in L$ such that $\min(a, F)$ had some root $b \in K$ which was not in L. We could then find an F-automorphism $\sigma : K \to K$ such that $\sigma(a) = b$. Note that $\min(a, F)$ must have some other root $c \in K \backslash \{b\}$ which is not in L, as otherwise we'd conclude that $b \in K$. We can find some $\tau \in H$ such that $\tau(b) = c$. Then we'd get

$$(\sigma^{-1} \circ \tau \circ \sigma)(a) = \sigma^{-1}(c) \neq a \Rightarrow \sigma^{-1} \circ \tau \circ \sigma \notin H \Rightarrow H \not\trianglelefteq G$$

as claimed.

2. Now, suppose that L/F is normal. Define a homomorphism $q : G \to \mathrm{Gal}(L/F)$ by restriction, that is $q(\sigma) = \sigma|_L$. That it has the claimed codomain follows from L/F being normal, as we saw in the previous part this implies that every element of G restricts to an F-automorphism of L. That $\ker(q) = H$ is clear, as is that q is surjective since K/L is normal. The desired result follows from this.

□

We can combine all these observations into one large, fundamental result.

Theorem 8.7.7 (Fundamental Theorem of Finite Galois Theory)**.** *Let K/F be a finite Galois extension, and set $G = \mathrm{Gal}(K/F)$. Then*

1. *$|G| = [K : F]$.*

2. *There is a bijective correspondence between subgroups $H \subset G$ and intermediate field extensions L, given by $L = K^H$.*

3. *L/F is normal if and only if $H \trianglelefteq G$, and in this case* $\mathrm{Gal}(L/F) \cong G/H$.

Essentially, **all the information about Galois extensions is contained in their Galois groups**.

Note. This theorem does not hold for infinite Galois extensions, but there is an analogue of it [Lan05].

Corollary 8.7.7.1. *Let K/F be a Galois extension. We say that the extension is Abelian (cyclic) if* $\mathrm{Gal}(K/F)$ *is Abelian (cyclic). If K/F is Abelian (cyclic), then every intermediate field extension is also Galois and Abelian (cyclic) over F.*

8.8 Properties of Galois Extensions

This section is based off of lectures by Dr. Sujatha Ramdorai and a similar section in [Lan05]. In it, we go over a collection of useful properties and facts about Galois extensions.

Lemma 8.8.1 (Dedikind)**.** *Let F be a field and $\sigma_1, \dots \sigma_n \in \mathrm{Aut}(F)$ distinct. Then these automorphisms are linearly independent over F.*

Proof. Suppose this was false. Then, without loss of generality, there would be some minimal m and coefficients $\alpha_i \in K$ such that $\sum_{i=1}^m \alpha_i \sigma_i = 0$. Clearly we need for $m \geq 2$, and so since these automorphisms are distinct there exists some non-zero $u \in K$ such that $\sigma_1(u) \neq \sigma_m(u)$. Note that any non-zero element of K can be written in the form cu for some non-zero $c \in K$, and

$$\sum_{i=1}^m \alpha_i \sigma_i(cu) = 0, \sum_{i=1}^m \alpha_i \sigma_1(u)\sigma_i(c) = 0$$

Thus,

$$\sum_{i=2}^m \alpha_i(\sigma_1(u) - \sigma_i(u))\sigma_i(c) = 0$$

Since $\sigma_1(u) - \sigma_m(u) \neq 0$, this would make m non-minimal, a contradiction. □

Theorem 8.8.2. *Let K/F be a finite field extension. Then* $|\mathrm{Gal}(K/F)| \leq [K : F]$.

Proof. Let $n = [K : F]$, and suppose there existed distinct $\sigma_1, \dots, \sigma_{n+1} \in \mathrm{Gal}(K/F)$. By Dedikind's lemma these would be algebraically independent. Furthermore, since K is an n-dimensional F-vector space, we find an F-basis $\{k_1, \dots, k_n\}$ for K. Consider the set of linear equations in $n+1$ unknowns $x_1, \dots, x_{n+1}$ given by

$$\begin{cases} x_1\sigma_1(k_1) + \cdots + x_{n+1}\sigma_{n+1}(k_1) = 0 \\ \vdots \\ x_1\sigma_1(k_n) + \cdots + x_{n+1}\sigma_{n+1}(k_n) = 0 \end{cases}$$

A solution to this would imply that the σ_i are linearly dependent. But of course this is a system of n linear equations with $n+1$ unknowns, so a solution exists, contradicting our assumption that the σ_i were distinct. □

The other results I will simply state, as I don't find their proofs particularly interesting or enlightening. If you would like to see the proofs, see section 6.1 in [Lan05].

Proposition 8.8.3. *Let K/F be Galois, and L/F some other field extension such that K, L are contained in some common field. Then $KL/L, K/(K \cap L)$ are Galois, and $\mathrm{Gal}(KL/L) \cong \mathrm{Gal}(K/(K \cap L))$, with the isomorphism being the expected restriction map. Furthermore, if K/F is finite, then $[KL : L] \mid [K : F]$.*

Proposition 8.8.4. *Let $K_1, \ldots, K_n$ be Galois extensions of F. Then $K_1 \cdots K_n/F$ is Galois. Furthermore, if $K_j \cap \prod_{i \neq j} K_i = F$ for any j and $G_i = \mathrm{Gal}(K_i/F)$, then*

$$\mathrm{Gal}(K_1 \cdots K_n/F) \cong G_1 \times \cdots \times G_n$$

Finally, if K/F is a finite Galois extension such that $\mathrm{Gal}(K/F) = G_1 \times \cdots \times G_n$, and we set K_i to be the fixed field of $\prod_{j \neq i} G_j$, then the K_i meet the above conditions and $\prod_{i=1}^{n} K_i = K$.

Note. The second part of this proposition gives us a way of decomposing finite Galois extensions into simpler ones. In particular, any finite Abelian Galois extension can be decomposed into the composite of cyclic Galois extensions.

Proposition 8.8.5. *Let K, L be field extensions of F contained in a common field, and E an intermediate extension of K/F. Then*

1. *If K, L are Abelian over F, then KL/F is Abelian.*
2. *If K/F is Abelian, then KL/L is Abelian.*
3. *If K/F is Abelian, then $K/E, E/F$ are Abelian.*

Note that this last proposition implies that the composition of all Abelian extensions of F is Abelian over F, and hence that there exists an *Abelian closure* of F. We denote this by F^{ab}.

8.9 Norm and Trace

This section is a re-organization of a similar one in [Lan05].

We begin with the basic definitions.

Definition 8.9.1. Let K/F be a finite field extension, and for any $a \in K$ let $T_a : K \to K$ be the F-linear map $x \mapsto ax$. We define the norm and trace of a over F to be $N_F^K(a) = \det(T_a), \mathrm{Tr}_F^K(a) = \mathrm{Tr}(T_a)$ respectively.

These quantities are of interest precisely because they turn out to relate to F-automorphisms of K.

Proposition 8.9.2. *Let K/F be a finite field extension, with $[K:F]_s = n, [K:F]_i = m$. Let $\sigma_1, \ldots, \sigma_n : L \hookrightarrow \overline{F}$ be the distinct F-embeddings of K in $\overline{F}$. Then for any $a \in K$*

$$N_F^K(a) = \Big(\prod_{i=1}^{n} \sigma_i(a)\Big)^m \qquad \mathrm{Tr}_F^K(a) = m\sum_{i=1}^{n} \sigma_i(a)$$

Proof. First, we justify there being exactly n distinct F-embeddings of K in $\overline{F}$[3]. By Theorem 8.5.5, there exists some intermediate field $F \subset L \subset K$ such that K/L is purely inseparable and L/F is separable. Then every extension of Id_F to K arises from exactly one extension of Id_F to L, of which there are n. By Theorem 8.3.21, L/F is simple, and hence there exists some $b \in L$ such that $L = F(b)$. Since any extension of Id_F to L is entirely defined by its action on b, it follows that all the roots of $\min(b, F)$ are in L and that each σ_i is just mapping b to a distinct root of $\min(b, F)$. We can extend σ_i from L to K by just having it act as the identity on each new element we add on, and each σ_i can only be extended from L to K in one manner. Since every F-embedding of K in $\overline{F}$ must arise in this manner, there are n distinct such F-embeddings.

Now, we consider the map T_a. Let $\min(a, F) = x^r + \sum_{j=1}^{r} \alpha_j x^{r-j}$. We know that $\{1, a, \ldots, a^{r-1}\}$ is an F-basis for $F(a)$. Let $\{v_j\}_{j=1}^{\ell}$ be a basis for K over $F(a)$, where we note that $\ell = nm/r$. Then by the proof of proposition 8.1.3,

$$\{a^i v_j \mid 0 \le i \le r-1, 1 \le j \le \ell\}$$

is a basis for K over F. In this basis, the matrix for T_a would be block-diagonal, with blocks of the form

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -\alpha_r \\ 1 & 0 & 0 & \cdots & 0 & -\alpha_{r-1} \\ 0 & 1 & 0 & \cdots & 0 & -\alpha_{r-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\alpha_1 \end{pmatrix}$$

where we've ordered here by i then j, each in ascending order[4]. The determinant of each of these blocks is $(-1)^r\alpha_r$, and there are ℓ blocks, so $\det(T) = (-1)^{nm}\alpha_r^{\ell}$. Similarly, the trace of each block is $-\alpha_1$, so $\mathrm{Tr}(T) = -\ell\alpha_1$. We split now into two cases.

Case 1: Suppose that K/F is separable. As noted above, each $\sigma_i : F(a) \hookrightarrow \overline{F}$ is just mapping a to one of the r distinct root of $\min(a, F)$. Furthermore, by the separability of $K/F(a)$, each of the said extensions then has ℓ extensions to an F-embedding of K. Thus, if $a = a_1, a_2, \ldots, a_r$ are the roots of $\min(a, F)$, then

$$\prod_{i=1}^{n} \sigma_i(a) = \Big(\prod_{i=1}^{r} a_i\Big)^{\ell} = ((-1)^r\alpha_r)^{\ell} = (-1)^n\alpha_r^{\ell} = N_F^K(a)$$

$$\sum_{i=1}^{n} \sigma_i(a) = \ell\sum_{i=1}^{r} a_i = -\ell\alpha_1 = \mathrm{Tr}_F^K(a)$$

[3]Perhaps unnecessary, but I like how it makes this proof stand alone.

[4]You may recognize this as the rational canonical form of T_a. That is, this also implies that the invariant factors of T_a are just copies of the minimal polynomial of a over F

as claimed.

Case 2: Suppose that K/F is not separable. Let $p = \text{char}(F)$, where p is prime and non-zero. By corollary 8.3.18.1, there exists some separable irreducible $g \in F[x]$ such that $\min(a, F)(x) = g(x^{p^z})$, for some $z \geq 0$. Again, each $\sigma_i : K \hookrightarrow \overline{F}$ can be thought of as arising from some extension of Id_F to $F(a)$. As noted above, each of these extensions is just mapping a to one of the $q = \deg(g) = r/(p^z) = [F(a) : F]_s$ distinct root of $\min(a, F)$. Each of the said extensions then has $[K : F(a)]_s$ extensions to an F-embedding of K. Thus, if $a = a_1, a_2, \ldots, a_q$ are the roots of $\min(a, F)$, then

$$\begin{aligned}\Big(\prod_{i=1}^{n} \sigma_i(a)\Big)^m &= \Big(\prod_{i=1}^{q} a_i\Big)^{m[K:F(a)]_s} = \Big(\prod_{i=1}^{q} a_i\Big)^{nm/q} = \Big(\prod_{i=1}^{q} a_i^{p^z}\Big)^{nm/r} = \Big((-1)^q \alpha_r\Big)^{\ell} \\ &= (-1)^{\ell q} \alpha_r^{\ell} = (-1)^{nm} \alpha_r^{\ell}\end{aligned}$$

as claimed. The last step works because raising (-1) to the exponent any non-zero power of p gives (-1) in F. Finally, we get

$$m \sum_{i=1}^{n} \sigma_i(a) = m[K : F(a)]_s \sum_{i=1}^{q} a_i = \frac{m[K : F(a)]_s}{p^z} \sum_{i=1}^{q} p^z a_i = -[K : F(a)]\alpha_1 = -\ell\alpha_1$$

as claimed. □

Corollary 8.9.2.1. *Let K/F be a finite field extension. For any $a \in K$, let $\min(a, F) = x^r + \sum_{i=1}^{r} \alpha_i x^{r-i}$. Then*

$$N_F^K(a) = (-1)^{[K:F]} \alpha_r^{[K:F]/r} \qquad \text{Tr}_F^K(a) = -\frac{[K : F]\alpha_1}{r}$$

Furthermore, if K/F is separable and $\sigma_1, \ldots, \sigma_n : L \hookrightarrow \overline{F}$ are the distinct F-embeddings of K in $\overline{F}$ then

$$N_F^K(a) = \prod_{i=1}^{n} \sigma_i(a) \qquad \text{Tr}_F^K(a) = \sum_{i=1}^{n} \sigma_i(a)$$

and if K/F is not separable then $\text{Tr}_F^K(a) = 0$. Finally, N_F^K is a multiplicative homomorphism and Tr_F^K an additive homomorphism.

Let's go over another basic property of these operations.

Proposition 8.9.3 (Transitivity)**.** *Let $K/L/F$ be a tower of finite field extensions. Then*

$$N_F^K = N_F^L \circ N_L^K \qquad \text{Tr}_F^K = \text{Tr}_F^L \circ \text{Tr}_L^K$$

Proof. This follows by noting that every F-embedding of K is an extension of a unique F-embedding of L. □

The trace function also has a strong connection to dual spaces.

Theorem 8.9.4. *Let K/F be a finite separable field extension. Then the map $f : (x, y) \mapsto \mathrm{Tr}_F^K(xy)$ is bilinear, and the map $x \mapsto f(x, \cdot)$ is an isomorphism between K and its dual space over F.*

Proof. That f is bilinear follows immediately from the linearity of Tr_F^K. For the second part, it's clear that $g : x \mapsto f(x, \cdot)$ is an F-homomorphism from K to $\overleftarrow{K}$. Suppose $g(x) = 0$. Then for if $x \neq 0$, we'd conclude that $\mathrm{Tr}_F^K(xx^{-1}) = 0$. Since $\mathrm{Tr}_F^K(1) = [K : F] \neq 0$, $x = 0$. Thus, g is injective. Since $\dim_F(K) < \infty$, this is sufficient for surjectivity. □

Corollary 8.9.4.1. *Let K/F be a finite separable field extension, and $x_1, \ldots, x_n$ a basis for K over F. Then there exists another basis $y_1, \ldots, y_n$ of K over F such that $\mathrm{Tr}_F^K(x_i y_j) = \delta_{ij}$.*

Proof. Since g is an isomorphism, there exists a basis of K over F dual to $g(x_1), \ldots, g(x_n)$. This is the desired basis. □

Theorem 8.9.5. *Let K/F be a finite separable field extension. Let $a \in K$ be an element such that $F(a) = K$, and let $f = \min(a, F)$. Define*

$$\frac{f(x)}{(x-a)} = b_0 + b_1 x + \cdots + b_{n-1}x^{n-1}$$

where $b_i \in K$. Then the dual basis of $1, a, \ldots, a^{n-1}$ for K over F is $\frac{g(b_0)}{f'(a)}, \ldots, \frac{g(b_{n-1})}{f'(a)}$, where g is the function from Theorem 8.9.4.

Proof. Let $a = a_1, a_2, \ldots, a_n$ be the distinct roots of f. Then

$$\sum_{i=1}^{n} \frac{f(x)}{(x-a_i)} \frac{a_i^r}{f'(a_i)} = x^r$$

for any $0 \leq r \leq n-1$. Indeed, plugging in any a_j to the left-hand side we get

$$\sum_{i=1}^{n} \frac{f(a_j)}{(a_j-a_i)} \frac{a_i^r}{f'(a_i)}$$

so the difference between the two sides is a polynomial of degree at most $n-1$ with $a_1, \ldots, a_n$ as roots, which implies that it must be zero.

We can extend Tr_F^K linearly to a map $K[x] \to F[x]$ by having it act on each coefficient separately. Using this new definition and the above equation, we'd get

$$\mathrm{Tr}_F^K\Big(\sum_{i=1}^{n} \frac{f(x)}{(x-a_i)} \frac{a_i^r}{f'(a_i)}\Big) = nx^r$$

Since $K = F(a)$, any F-embedding of K into $\overline{F}$ is just a choice of which root of f to map a to. Let $\sigma_i : K \hookrightarrow \overline{F}$ be the embedding defined by $\sigma_i(a) = a_i$, and in general write $\sigma_i(a_j) = a_{ij}$, where each a_{ij} is a root of f. By proposition 8.9.2, we know that

$$\mathrm{Tr}_F^K\Big(\frac{f(x)}{(x-a_i)} \frac{a_i^r}{f'(a_i)}\Big) = \sum_{j=1}^{n} \sigma_j\Big(\frac{f(x)}{(x-a_i)} \frac{a_i^r}{f'(a_i)}\Big) = \sum_{j=1}^{n} \frac{f(x)}{(x-a_{ij})} \frac{a_{ij}^r}{f'(a_{ij})} = \sum_{j=1}^{n} \frac{f(x)}{(x-a_j)} \frac{a_j^r}{f'(a_j)}$$

which is completely independent of i. Thus, we get

$$\mathrm{Tr}_F^K\left(\frac{f(x)}{(x-a)}\frac{a^r}{f'(a)}\right) = x^r$$

Or, expanding this out

$$\mathrm{Tr}_F^K\left(\sum_{i=0}^{n-1}\frac{a^r}{f'(a)}b_i x^i\right) = x^r$$

Comparing terms, we conclude that

$$\mathrm{Tr}_F^K\left(\frac{b_i}{f'(a)}a^r\right) = \delta_{ir}$$

which is the desired result. □

We'll end off with a pair of related theorems, proved by Hilbert, which will be useful in the next section.

Theorem 8.9.6 (Hilbert's Theorem 90)**.** *Let K/F be a finite cyclic Galois field extension of degree n with $G = \mathrm{Gal}(K/F) = \langle\sigma\rangle$. Pick any $a \in K$. Then $N_F^K(a) = 1$ if and only if there exists some non-zero $b \in K$ such that $a = \frac{b}{\sigma(b)}$.*

Proof. Suppose that $N_F^K(a) = 1$. Clearly, $a \neq 0$. By lemma 8.8.1, we know

$$\mathrm{Id}_K + a\sigma + a\sigma(a)\sigma^2 + \cdots + a\sigma(a)\sigma^2(a)\cdots\sigma^{n-2}(a)\sigma^{n-1} \neq 0$$

Thus, there exists some $b \in K$ such that

$$\Big(\mathrm{Id}_K + a\sigma + a\sigma(a)\sigma^2 + \cdots + a\sigma(a)\sigma^2(a)\cdots\sigma^{n-2}(a)\sigma^{n-1}\Big)(b) = c \neq 0$$

Then

$$\begin{aligned} a\sigma(c) &= a\sigma\Big(\Big(\mathrm{Id}_K + a\sigma + a\sigma(a)\sigma^2 + \cdots + a\sigma(a)\sigma^2(a)\cdots\sigma^{n-2}(a)\sigma^{n-1}\Big)(b)\Big) \\ &= a\Big(\sigma + \sigma(a)\sigma^2 + \sigma(a)\sigma^2(a)\sigma^3 + \cdots + \sigma(a)\sigma^2(a)\sigma^3(a)\cdots\sigma^{n-1}(a)\sigma^n\Big)(b) \\ &= \Big(a\sigma + a\sigma(a)\sigma^2 + a\sigma(a)\sigma^2(a)\sigma^3 + \cdots + aN_F^K(a)\,\mathrm{Id}_K\Big)(b) = c \end{aligned}$$

Thus, $a = \frac{c}{\sigma(c)}$ as desired. Now, suppose that $a = \frac{b}{\sigma(b)}$ for some $b \neq 0$. Then since $N_F^K(\sigma(b)) = N_F^K(b)$, we get $N_F^K(a) = 1$. □

There's another form of this which sometimes gets brought up.

Theorem 8.9.7 (Hilbert's Theorem 90 Additive)**.** *Let K/F be a finite cyclic Galois field extension of degree n with $G = \mathrm{Gal}(K/F) = \langle\sigma\rangle$. Pick any $a \in K$. Then $\mathrm{Tr}_F^K(a) = 0$ if and only if there exists some non-zero $b \in K$ such that $a = b - \sigma(b)$.*

Proof. Suppose that $\mathrm{Tr}_F^K(a) = 0$, and set $b \in K$ such that $\mathrm{Tr}_F^K(b) = 1$[5]. Let

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

Then

$$\begin{aligned} a + \sigma(c) &= a + \sigma(a)\sigma^2(b) + (\sigma(a) + \sigma^2(a))\sigma^3(b) + \cdots + (\sigma(a) + \cdots + \sigma^{n-1}(a))\sigma^n(b) \\ &= a\sigma(b) + (a + \sigma(a))\sigma^2(b) + \cdots + (a + \sigma(a) + \cdots + \sigma^{n-2}(a))\sigma^{n-1}(b) \\ &\quad - a(\sigma(b) + \sigma^2(b) + \cdots + \sigma^{n-1}(b)) + (\sigma(a) + \cdots + \sigma^{n-1}(a))b + a \\ &= c - a(\mathrm{Tr}_F^K(b) - b) + (\mathrm{Tr}_F^K(a) - a)b + a = c - a(1 - b) - ab + a = c \end{aligned}$$

so $a = c - \sigma(c)$. If there exists some $b \in K$ such that $a = b - \sigma(b)$, then the result follows from $\mathrm{Tr}_F^K(\sigma(b)) = \mathrm{Tr}_F^K(b)$. □

8.10 Cyclotomic and Cyclic Extensions

This section is based off a similar one in [Lan05], along with lectures by Dr. Sujatha Ramdorai. In it, we study the seemingly simple equation $x^n - 1$.

Definition 8.10.1. Let F be a field. An nth root of unity over F is a solution to the polynomial $x^n - 1$ in $\overline{F}$.

Note. 1 is always an nth root of unity.

Sometimes, 1 is our only nth root of unity.

Proposition 8.10.2. *Let* $\mathrm{char}(F) = p \neq 0$*, and suppose that* $n = p^k$ *for some* $k \geq 0$*. Then the only* n*th root of unity is* 1*.*

Proof. In $\overline{F}[x]$, $x^{p^k} - 1 = (x - 1)^{p^k}$. □

Theorem 8.10.3. *Suppose* $\mathrm{char}(F) = 0$ *or* $GCD(\mathrm{char}(F), n) = 1$*. Then* $x^n - 1$ *is separable and the* n*th roots of unity over* F *form a cyclic multiplicative group.*

Proof. The separability follows from $(x^n - 1)' = nx^{n-1} \neq 0$, since $p \nmid n$ or $\mathrm{char}(F) = 0$. Thus, $x^n - 1$ has n distinct roots, call them $1, \alpha_1, \ldots, \alpha_{n-1}$. These form a multiplicative group, as any finite multiplicative subgroup of a field is cyclic. □

We denote this group of nth roots of unity by μ_n, and call a generator of μ_n a *primitive* nth root of unity. By convention, we'll use the symbol ζ_n to denote a primitive nth root of unity.

Proposition 8.10.4. *Let* n, m *be coprime. Then* $\mu_{nm} \cong \mu_n \times \mu_m$*.*

Proof. This follows from the observations $\mu_n \cap \mu_m = \{1\}$ and $|\mu_n \mu_m| = mn$. □

We now take a quick detour to talk about some basic rings.

[5] Such an element must exist by corollary 8.9.4.1

Definition 8.10.5. The Euler totient function $\varphi : \mathbb{N} \to \mathbb{N}$ is given by

$$\varphi(n) = \{m \in \mathbb{N} \mid m \leq n, \text{GCD}(n, m) = 1\}$$

Proposition 8.10.6. *Let $n \in \mathbb{Z}$. Then $(\mathbb{Z}/n\mathbb{Z})^\times$ is a multiplicative group of order $\varphi(n)$.*

Proof. First, note that $a \in (\mathbb{Z}/n\mathbb{Z})^*$ if and only if it is coprime to n. Furthermore, if a is coprime to n, then there exist $b, c \in \mathbb{Z}$ such that $ab + nc = 1$, so $ab \equiv 1 \bmod n$. Without loss of generality, we may assume that $0 < b < n$. But this also implies that b is coprime to n. □

Theorem 8.10.7. *Suppose ζ_n is a primitive nth root of unity over F, where $\text{char}(F) \nmid n$. Then there exists an embedding $\text{Gal}(F(\zeta_n)/F) \hookrightarrow (\mathbb{Z}/n\mathbb{Z})^\times$. In particular, $F(\zeta_n)/F$ is Abelian and Galois.*

Proof. This follows by noting that every $\sigma \in \text{Gal}(F(\zeta_n)/F)$ acts by $\sigma(\zeta_n) = \zeta_n^k$ for some k, ζ_n^k must be primitive for this to be an automorphism, and ζ_n^k is primitive only if $\text{GCD}(k, n) = 1$. □

Corollary 8.10.7.1. $[F(\zeta_n) : F] \mid \varphi(n)$.

In general, the above embedding is not an isomorphism. However, there is a rather important case where it is.

Theorem 8.10.8. *Let ζ_n be a primitive nth root of unity over $\mathbb{Q}$. Then $\text{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \cong (\mathbb{Z}/n\mathbb{Z})^\times$.*

Proof. It suffices to show that the embedding from the previous theorem is surjective. That is, it suffices to show that for any number k less than and coprime to n, ζ_n^k is a primitive nth root of unity. In particular, we only need to demonstrate the truth of this for prime k. To that end, let $f = \min(\zeta_n, \mathbb{Q})$, and let $h \in \mathbb{Q}[x]$ be the polynomial such that $x^n - 1 = f(x)h(x)$. h, f both have leading coefficients 1, and hence are primitive and in $\mathbb{Z}[x]$ by lemma 3.9.12 and Gauss' lemma[6]. It will suffice to show that, for any prime p coprime to and less than n, ζ_n^p is a root of f. Suppose this were not the case. Then ζ_n^p would be a root of h. Thus, $f(x) \mid h(x^p)$, that is there exists some $g \in \mathbb{Z}[x]$ such that

$$h(x^p) = f(x)g(x)$$

In mod-p arithmetic, this equality becomes

$$h(x)^p \equiv f(x)g(x) \bmod p$$

In particular, f, h are not coprime when regarded as being in $\mathbb{Z}/p\mathbb{Z}[x]$. Since $x^n - 1 = f(x)h(x)$ still holds in $\mathbb{Z}/p\mathbb{Z}[x]$, this would imply that $x^n - 1$ is not separable in $\mathbb{Z}/p\mathbb{Z}[x]$, which is impossible. □

[6] This is not as obvious as I'm making it seem, but proving it is a good review of these concepts. It helps to look at Theorem 3.9.4.

Corollary 8.10.8.1. *Suppose n, m are coprime. Then $\mathbb{Q}(\zeta_n) \cap \mathbb{Q}(\zeta_m) = \mathbb{Q}$.*

Proof. Since n, m are coprime, $\zeta_n \zeta_m = \zeta_{nm}$. By Theorem 8.10.8, $\mathrm{Gal}(\mathbb{Q}(\zeta_{mn})/\mathbb{Q}) \cong \mathbb{Z}/(nm)\mathbb{Z}$ $\cong \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$, where the last isomorphism comes from n, m being coprime. The result then follows from the fixed fields of $\mathbb{Z}/n\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$ being precisely $\mathbb{Q}(\zeta_n)$ and $\mathbb{Q}(\zeta_m)$, along with proposition 8.8.4. □

Let's look a bit more at factoring the equation $x^n - 1$. We know that in an algebraic closure

$$x^n - 1 = \prod_{\zeta \in \overline{F}, \zeta^n = 1} (x - \zeta)$$

In particular, we can group these ζ by the smallest m for which $\zeta^m = 1$. We define the $\Phi_m(x)$, the mth Cyclotomic polynomial, to be the product of all the factors in $x^n - 1$ where m is the smallest number such that $\zeta^m = 1$. It is immediate from this that

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

One may also note that the roots of ζ_d are precisely the primitive dth roots of unity over F. We can rearrange the above equation as well, to get

$$\Phi_n(x) = \frac{x^n - 1}{\prod_{d|n, d<n} \Phi_d(x)}$$

This gives us a recursive formula for calculating Φ_n, as $\Phi_1(x) = x - 1$ is known. In particular, by Theorem 3.9.4, this formula implies that over $\mathbb{Q}$, $\Phi_n \in \mathbb{Z}[x]$ and is always monic. Since there are precisely $\varphi(n)$ primitive nth roots of unity over any field F, we conclude that $\deg(\Phi_n) = \varphi(n)$ and by Theorem 8.10.8 that over $\mathbb{Q}$

$$\Phi_n(x) = \min(\zeta_n, \mathbb{Q})$$

This also provides a slick proof of the identity

$$n = \sum_{d|n} \varphi(d)$$

There are many other neat identities involving these polynomials, the interested reader is directed to section 6.3 of [Lan05].

Let's use these concepts, along with our results on norm and trace, to talk about cyclic extensions.

Theorem 8.10.9. *Let F be a field, and $n \in \mathbb{N}$. Suppose that $\mathrm{char}(F)$ is zero or coprime to n, and that there is a primitive nth root of unity in F. Then the following hold.*

1. *If K/F is a cyclic Galois extension of degree n, then there exists some $a \in K, b \in F$ such that $K = F(a)$ and a is a root of $x^n - b$.*

2. *If $x^n - b \in F[x]$, and a is a root of $x^n - b$, then $F(a)/F$ is a Cyclic Galois extension of degree $d \mid n$.*

Proof. 1. Let $\mathrm{Gal}(K/F) = \langle\sigma\rangle$, and let $\zeta_n \in F$ be a primitive nth root of unity. By corollary 8.9.2.1, $N_F^K(\zeta_n) = \zeta_n^n = 1$. Thus, by Theorem 8.9.6, there exists some non-zero $a \in K$ such that $\zeta_n = \frac{a}{\sigma(a)}$. This implies $a^n = \sigma(a^n)$, and hence that a^n is fixed by $\langle\sigma\rangle$, meaning $a^n = b \in F$. $x^n - b$ therefore has distinct roots $a, \zeta_n a, \zeta_n^2 a, \ldots, \zeta_n^{n-1} a$, making it separable and making $F(a)$ the splitting field of $x^n - b$. Thus, $K/F(a)/F$ is a tower of Galois extensions. But clearly $|\mathrm{Gal}(F(a)/F)| = n$, as any choice of $\zeta_n^k a$ for a gives an automorphism of $F(a)$, and so $[F(a) : F] = n \Rightarrow K = F(a)$.

2. Now, suppose that $x^n - b \in F[x]$, and a is a root of $x^n - b$. Then there exists some minimal $d \mid n$ such that $a^d \in F$. We simply apply the above observations to $x^d - a^d$ to get the second result, as $\zeta_n^{n/d}$ is a primitive dth root of unity in F.

□

Theorem 8.10.10 (Artin-Schreier)**.** *1. Suppose K/F is a cyclic Galois extension of degree $p = \mathrm{char}(F)$. Then $K = F(a)$, where a is a root of some $x^p - x - b \in F[x]$.*

2. Suppose $b \in F$. Then $x^p - x - b$ either splits completely or is irreducible over F, and its splitting field is a cyclic Galois extension of degree p if it is irreducible.

Proof. 1. Let $\mathrm{Gal}(K/F) = \langle\sigma\rangle$. Since $\mathrm{Tr}_F^K(-1) = -p = 0$, there exists some $a \in K$ such that $-1 = a - \sigma(a)$. Thus, we get

$$\sigma(a^p - a) = \sigma(a)^p - \sigma(a) = (a-1)^p - (a-1) = a^p - a$$

so $a^p - a = b \in F$, and a is a root of $f(x) = x^p - x - b \in F[x]$. It suffices now to show that f is irreducible over F. Note that $\sigma(a) \neq a \Rightarrow a \notin F$, and $a, a+1, \ldots, a+p-1$ are all distinct roots of f in $F[x]$. $F(a)/F$ is therefore a non-trivial cyclic Galois extension, whose Galois group has an order dividing p. But $\mathbb{Z}/p\mathbb{Z}$ is the only non-trivial cyclic group of order dividing p, so $F(a) = K$.

2. Suppose $b \in F \backslash F^p$, and set $f(x) = x^p - x - b$. Let a be a root of f. Then $a, a+1, \ldots, a+p-1$ are the distinct roots of f, making f separable and $F(a)/F$ Galois (and $F(a)$ the splitting field of f). It's clear that if any root of f is in F, then all roots of f are. Thus, We just need to show that f is irreducible otherwise. Indeed, suppose $f(x) = g_1(x) \cdots g_n(x)$ is a decomposition of f into irreducible polynomials in $F[x]$. Then since all the roots of f differ only by addition of constants in F, each g_i must be the same degree d. We conclude that $dn = p \Rightarrow d = p$ or $n = p$. The second of these cases is impossible, as we've assumed that some root of f is not in F, so $d = p$ and hence f is irreducible.

□

We end off my mentioning some interesting results whose proofs are beyond the scope of this text, or in the case of the Inverse Galois Problem simply too much of a digression to be worth pursuing.

Theorem 8.10.11 (Kronecker-Weber)**.** *We call an extension of $\mathbb{Q}$ Cyclotomic if it can be obtained by adjoining roots of unity to $\mathbb{Q}$. With that definition, any Abelian extension of $\mathbb{Q}$ is a sub-extension of a Cyclotomic extension of $\mathbb{Q}$. [Gha99]*

Conjecture 8.10.12 (Inverse Galois Problem)**.** *Every finite Abelian group occurs as the Galois group of some Galois extension of $\mathbb{Q}$. [Gha99]*

8.11 Solvable and Radical Extensions

This section is based on a similar one from [Lan05]. We start with any finite field extension K/F. Denote by K^{norm} the splitting field of the family of polynomials $\{\min(a, F)\}_{a\in K}$, this is the smallest normal extension of F containing K, and is often called the *normal closure* of K over F.

Definition 8.11.1. A finite field extension K/F is solvable if $\mathrm{Gal}(K^{norm}/F)$ is solvable.

We can also define K_{sep} to be the composition of all separable sub-extensions of a finite field extension K/F. By Theorem 8.5.5, K/K_{sep} is purely inseparable and K_{sep}/K is separable. Thus, since purely inseparable extensions have trivial Galois groups (indeed, all the minimal polynomials have a single root) $\mathrm{Gal}(K/F) = \mathrm{Gal}(K_{sep}/F)$. Furthermore, $(K_{sep})^{norm}/F$ is separable, as all the relevant minimal polynomials are separable. Thus, for the purposes of the following theorem, we may assume that all our field extensions are separable, as taking the separable subfield plays well with field composition and taking normal closures.

Theorem 8.11.2. *The class of solvable field extensions is distinguished.*

Proof. Let K/F be solvable, and $F \subset L \subset K$ any subfield of K containing F. Then $K^{norm}/L^{norm}/F$ is a tower of Galois field extensions, and hence $\mathrm{Gal}(L^{norm}/F) \cong \frac{\mathrm{Gal}(K^{norm}/F)}{\mathrm{Gal}(K^{norm}/L^{norm})}$. Since $\mathrm{Gal}(K^{norm}/F)$ is solvable, it follows by lemma 2.9.5 that $\mathrm{Gal}(K^{norm}/L^{norm}), \mathrm{Gal}(L^{norm}/F)$ are solvable, and hence $K/L, L/F$ are solvable. Now, suppose that $L/F, K/L$ are solvable. Then we again get that

$$\mathrm{Gal}(L^{norm}/F) \cong \frac{\mathrm{Gal}(K^{norm}/F)}{\mathrm{Gal}(K^{norm}/L^{norm})}$$

making $\mathrm{Gal}(K^{norm}/F)$ solvable by the same lemma, and hence K/F solvable.

Now, suppose that L/F is any algebraic extension, and K/F is solvable. It suffices to show that LK/L is solvable. Indeed, $(KL)^{norm}/L^{norm}$ is Galois by proposition 8.8.3, and from this same proposition we get $\mathrm{Gal}((KL)^{norm}/L^{norm}) \cong \mathrm{Gal}(K^{norm}/(K^{norm} \cap L^{norm})) \hookrightarrow \mathrm{Gal}(K^{norm}/F)$, thus making $\mathrm{Gal}((KL)^{norm}/L^{norm})$ and hence KL/L solvable. □

In practice, we prefer to work with a seemingly simpler class of solvable extensions. Note at this point we drop the assumption that K/F is separable.

Definition 8.11.3. Let K/F be a finite extension, and $p = \text{char}(F)$. We call K/F solvable by radicals if there exists a finite field extension L/F containing K and a tower decomposition of this extension

$$F = E_0 \subset E_1 \subset \cdots \subset E_m = L$$

such that each E_{i+1} is formed by adjoining to E_i one of

1. A root of unity.
2. A root of $x^n - a \in E_i[x]$, where $\text{GCD}(n, p) = 1$, or $p = 0$, or $n = p$.
3. A root of $x^p - x - a \in E_i[x]$, if $p \neq 0$.

The trick is, this is actually an equivalent definition to our definition of solvability!

Theorem 8.11.4. *Let K/F be a finite extension. Then K/F is solvable if and only if it is solvable by radicals.*

Proof. First, suppose that K/F is solvable, and set $G = \text{Gal}(K^{norm}/F)$. We start with the case where K is separable over F. It is important to note that K^{norm}/F is finite, as K is finitely generated over F and hence we need only consider finitely many polynomials when constructing K^{norm}. Thus, by Theorem 2.9.6, we can find a cyclic composition series $G = G_0 \supset G_1 \supset \cdots \supset G_m = \{1\}$ of G. Define $E_i = (K^{norm})^{G_i} \cap K$. Then $E_i^{norm} = (K^{norm})^{G_i}$, and

$$K^{norm} = E_m^{norm}/E_{m-1}^{norm}/\cdots/E_0^{norm} = F$$

is a tower of field extensions with cyclic Galois groups. Let n be the product of all primes dividing $|G|$ not equal to $\text{char}(F)$, and let ζ_n be a primitive nth root of unity over F. Then the above composition series has Galois groups which contain those of the composition series

$$K^{norm}(\zeta_n) = E_m^{norm}(\zeta_n)/E_{m-1}^{norm}(\zeta_n)/\cdots/E_0^{norm}(\zeta_n) = F(\zeta_n)$$

Since $F(\zeta_n)/F$ is certainly finite, separable, and solvable by radicals, we may reduce to the case where K^{norm}/F is cyclic and F contains an nth root of unity ζ_n. Let H be a subgroup of G such that $|H| = p$, and let $E = (K^{norm})^H$. Then K^{norm}/E is a cyclic Galois extension of order p, and by Theorem 8.10.10 there exists some $a \in K^{norm}$ with minimal polynomial $x^p - x - b \in E[x]$ over F such that $K^{norm} = E(a)$. Doing this repeatedly, we may assume that $p \nmid |G|$, and hence $n = |G|$. Finally, in this case, we conclude by Theorem 8.10.9 that there exists some $a \in K^{norm}$ such that $K^{Gal} = F(a)$ and a is a root of $x^n - b \in F[x]$. This completes the proof in the case K/F being separable.

If K/F is not separable, we can construct a tower of extensions $K^{norm}/K_{sep}^{norm}/F$. By corollary 2.9.14.1, K_{sep}^{norm}/F is solvable by radicals. Thus, we need only then show that K^{norm}/K_{sep}^{norm} is solvable by radicals. To that end, pick any $a \in K^{norm}$. Since K^{norm}/K_{sep}^{norm} is purely inseparable, there exists some $b \in F$ such that a is a root of $x^{p^k} - b$, where $k \in \mathbb{N}$. Thus, K^{norm}/K_{sep}^{norm} is solvable by radicals, as claimed.

Now, suppose that K/F is solvable by radicals. Let L be the finite field extension of K such that L/F has the desired tower decomposition

$$F = E_0 \subset E_1 \subset \cdots \subset E_m = L$$

Since $K \subset L$, $K^{norm} \subset L^{norm}$. Thus, since subgroups of solvable groups are solvable, it suffices to prove that L^{norm}/F is solvable. Note that L^{norm}/L is radical, as we are simply adjoining the other roots of the equations used in the tower $E_m/E_{m-1}\cdots/E_1/E_0$. Hence, we may assume that $L = L^{norm}$. Let n be the product of all the primes dividing $[L : F]$ except $p = \text{char}(F)$. Then $F(\zeta_n)/F$ is certainly a radical extension, and it's pretty clear that $E_{i+1}(\zeta_n)/E_i(\zeta_n)$ is still a radical extension as well, and $L \subset L(\zeta_n)$, we may assume that $\zeta_n \in F$. Then by Theorem 8.10.9, Theorem 8.10.10, and purely inseparable extensions having a trivial Galois group, each $\text{Gal}(E_{i+1}/E_i)$ is cyclic, and in fact each E_{i+1}/E_i is normal. We claim that this implies $\text{Gal}(E_i/F) \trianglelefteq \text{Gal}(E_{i+1}/F)$, and

$$\frac{\text{Gal}(E_{i+1}/F)}{\text{Gal}(E_i/F)} \cong \text{Gal}(E_{i+1}/E_i)$$

giving us the desired decomposition of $\text{Gal}(L/F)$ and making it solvable. Indeed, suppose $\sigma \in \text{Gal}(E_{i+1}/F)$. σ is entirely defined by how it acts on the series of elements $a_1, \ldots, a_{i+1}$ such that $E_j(a_{j+1}) = E_{j+1}$. Since at each stage our extension is normal, all the other roots of the minimal polynomial of a_j are in E_j. Thus, $\sigma(a_j) \in E_j$, and hence σ restricts to an F-automorphism of E_j. This would, in turn, imply that any such σ can be decomposed into an F-automorphism of E_i and an E_i automorphism of E_{i+1}, which gives the desired result. □

8.12 Solving Polynomials

This section is based on a multitude of different sources, the foundations of which are notes by Dr. Sujatha Ramdorai and [Lan05].

We've spent all this time building up Galois extensions, solvable extensions, and the like. Now, we can finally talk about how they can be used to solve for (or determine when you could solve for) the roots of polynomials. Let's start with a basic definition.

Definition 8.12.1. Let $f \in F[x]$. We call f solvable if its splitting field is a solvable extension of F.

How does this relate to solving polynomials? Well, consider the quadratic equation

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

which gives the roots of a quadratic polynomial over any field except extensions of $\mathbb{F}_2$. It is immediate from this equation that every quadratic polynomial over F is solvable, as we just need to adjoin a root of $x^2 - (b^2 - 4ac)$ to get the splitting field. This is what we really mean

by "solving by radicals", we can solve the polynomials using only the operations of division, multiplication, addition, and taking roots[7].

Let's specialize to polynomials over $\mathbb{Q}$ in particular. For quadratic polynomials we may assume, without loss of generality, that $f(x) = x^2 + bx + c$. Then our roots are given by

$$-b \pm \sqrt{b^2 - 4c}$$

Furthermore, we can quickly see that the Galois group of the splitting field of f, which we denote $\mathrm{Gal}(f)$, is S_2 if $b^2 - 4ac \notin \mathbb{Q}^2$ and $\{1\}$ otherwise. We wish to try and do this for higher-degree polynomials. Again, we start by assuming our equation is of the form

$$f(x) = x^3 + bx^2 + cx + d$$

Setting $y = x + b/3$, we get

$$f(x) = y^3 + py + q$$

where

$$p = \frac{3c - b^2}{3} \qquad\qquad q = \frac{2b^3 - 9cb + 27d}{27}$$

Thus, we can assume our polynomial is in the form (called a depressed cubic)

$$f(x) = x^3 + px + q$$

It turns out that we can get a cubic formula for this equation, which is

$$\sqrt[3]{\frac{27q \pm \sqrt{(27q)^2 + 4 \cdot 27p^3}}{2}}$$

minus some special cases which we won't get into here (it is still always solvable though). The derivation of this is long and messy, and of no interest to us. What is of interest is attempting to generalize the discriminant to this new cubic case.

Definition 8.12.2. Let $f \in F[x]$ be a polynomial with leading coefficient 1 and degree $n \geq 2$. We define the discriminant of the polynomial, denoted $\mathrm{Disc}(f)$, to be

$$\mathrm{Disc}(f) = (-1)^{\frac{n(n-1)}{2}} \prod_{i \neq j} (\alpha_i - \alpha_j)$$

where $\alpha_i \in \overline{F}$ are the roots of f.

Note. $\mathrm{Disc}(f)$ is fixed by any permutation of the roots of f, and hence is in F.

[7] Of course, we're also allowed roots of $x^p - x - a$ for $p \neq 0$, but our focus in on $\mathbb{Q}$, so we don't care too much about that

Let's see what happens, for example, when $n = 2$. In this case, our formula becomes

$$\text{Disc}(f) = (\alpha_1 - \alpha_2)^2 = 4(b^2 - 4ac)$$

just a trivial perfect square factor away from our normal definition! For the cubic case, we get

$$\text{Disc}(f) = (\alpha_1 - \alpha_2)^2(\alpha_1 - \alpha_3)^2(\alpha_2 - \alpha_3)^2$$

Set $\delta(f) = \sqrt{\text{Disc}(f)}$. Then for any $\sigma \in \text{Gal}(f)$, $\sigma(\delta(f)) = \pm\delta(f)$. We can view $\text{Gal}(f)$ as a subset of S_3, with σ being a permutation of the roots. In this view, $\sigma(\delta(f)) = \text{sgn}(f)\delta(f)$.

Now, let's step back a bit. It's clear that f is irreducible over $\mathbb{Q}$ if and only if it has no roots in $\mathbb{Q}$. Assuming f is irreducible, let $\alpha \in \overline{Q}$ be a root. Then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$. If f splits completely over $\mathbb{Q}(\alpha)$, then $\text{Gal}(f) = A_3$ (i.e. we can only do 3-cycle permutations of the roots). Otherwise, $|\text{Gal}(f)| = 6 = |S_3|$, so $\text{Gal}(f) = S_3$.

Theorem 8.12.3. *Suppose f is cubic and irreducible over $\mathbb{Q}$. Then $\text{Gal}(f) = A_3$ if $\text{Disc}(f) \in \mathbb{Q}^2$, and $\text{Gal}(f) = S_3$ otherwise.*

Proof. Suppose that $\text{Disc}(f) \in \mathbb{Q}^2$. Then $\delta \in \mathbb{Q}$, so every $\sigma \in \text{Gal}(f)$ must fix δ. It follows that $\text{Gal}(f)$ contains only even permutation, and hence $\text{Gal}(f) = A_3$. Otherwise, $\delta \notin \mathbb{Q}$ and hence is not fixed by every $\sigma \in \text{Gal}(f)$. In particular, this implies that an odd permuation exists in $\text{Gal}(f) \Rightarrow \text{Gal}(f) \neq A_3$, so $\text{Gal}(f) = S_3$. □

One can generally solve the quartic as well, but doing so is a hideous exercise. There is also a generalization of the above theorem to the quartic case. I'll state it here, but not prove it.

Theorem 8.12.4. *Any monic quartic $f \in \mathbb{Q}[x]$ can be put in the form $f(x) = x^4+px^2+qx+r$. Suppose this f is irreducible over $\mathbb{Q}$. Define the resolvant cubic of f, $g(z)$, to be $g(z) = z^3 - pz^2 - 4rz + 4pr - q^2$. Then $\text{Disc}(f) = \text{Disc}(g)$, and*

1. *$\text{Gal}(f) \cong V_4$, the Klein 4-group, if and only if $\text{Disc}(f) \in \mathbb{Q}^2$ and $g(z)$ splits over $\mathbb{Q}$.*
2. *$\text{Gal}(f) \cong A_4$ if and only if $\text{Disc}(f) \in \mathbb{Q}^2$ and $g(z)$ has no roots in $\mathbb{Q}$.*
3. *$\text{Gal}(f) \cong S_4$ if and only if $\text{Disc}(f) \notin \mathbb{Q}^2$ and $g(z)$ has no roots in $\mathbb{Q}$.*
4. *$\text{Gal}(f) \cong \mathbb{Z}/4\mathbb{Z}$ if and only if $\text{Disc}(f) \notin \mathbb{Q}^2$ and $g(z)$ has exactly one root r' in $\mathbb{Q}$, and the polynomials $x^2 + r', x^2 + (r' - p)x + r$ both split over $\mathbb{Q}(\sqrt{\text{Disc}(f)})$.*
5. *$\text{Gal}(f) \cong D_4$ if and only if $\text{Disc}(f) \notin \mathbb{Q}^2$ and $g(z)$ has exactly one root r' in $\mathbb{Q}$, and (4) does not hold.*

Note. The above theorems will hold for any separable irreducible polynomial over a field of characteristic zero or a sufficiently high characteristic.

Of course, the theorems I've given above are rather useless to you if you can't calculate the discriminant, which as of now would require the roots anyway. The trick here is that there's another way to compute the discriminant without knowing the roots! In order to do this, we'll need to take a quick detour to talk about the *resultant*.

Definition 8.12.5. Let $f(x) = \sum_{i=0}^{n} a_i x^{n-i}, g(x) = \sum_{i=0}^{m} b_i x^{m-i} \in F[x]$, where $a_0, b_0 \neq 0$ and $n, m \geq 1$. We define the resultant of these polynomials $R(f, g)$ to be the determinant of the matrix

$$\begin{pmatrix} a_0 & a_1 & \cdots & a_n & \\ & a_0 & \cdots & a_{n-1} & a_n \\ & & \ddots & \ddots & \ddots \\ b_0 & b_1 & \cdots & b_m & \\ & b_0 & \cdots & b_{m-1} & b_m \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

where blank spaces are zero, the first pattern continues for m rows, and the second for n rows.

Note. The above equation is not supposed to imply that $m = n$.

The following remarkable result is true, but the proof is long enough that I will not cover it here. The interested reader could find it in [Lan05], section 4.8.

Theorem 8.12.6. *Suppose $\alpha_1, \ldots, \alpha_n$ are the roots of f and $\beta_1, \ldots, \beta_m$ those of g in $\overline{F}$. Then*

$$R(f, g) = a_0^m b_0^n \prod_{i=1}^{n} \prod_{j=1}^{m} (\alpha_i - \beta_j)$$

A corollary of this is what allows us to calculate the discriminant using the resultant, and hence knowing only the coefficients of the polynomial. Again, we state it without proof.

Corollary 8.12.6.1. $\operatorname{Disc}(f) = (-1)^{\frac{n(n-1)}{2}} a_0 \operatorname{Res}(f, f')$.

Let's go back now to the connection between solvable extensions and finding the roots of polynomials. Suppose there existed a general formula for finding the roots of a degree n polynomial, which used only multiplicative, division, addition, and taking roots. Then each degree n polynomial would necessarily be solvable by radicals, and hence solvable. That is,

Theorem 8.12.7. *There exists a general formula for finding the roots of a degree n polynomial in $F[x]$ which uses only multiplicative, division, addition, and taking roots only if each polynomial of degree at most n is solvable.*

We can use this to show that there is no extension of the quadratic formula to the quintic and beyond.

Lemma 8.12.8. *S_n is not solvable for $n \geq 5$.*

Proof. Suppose $H, N \subset S_n$ are two subgroups such that $N \trianglelefteq H$, H contains every 3-cycle, and H/N is Abelian. Pick any distinct $i, j, k, r, s \in \{1, \ldots, n\}$. Set $\sigma = (ijk), \tau = (krs)$. Then

$$\sigma\tau\sigma^{-1}\tau^{-1} = (ijk)(krs)(ikj)(ksr) = (irk)$$

Since H/N is Abelian, $[H, H] \subset N$. Thus, by the arbitrary choice of i, r, k, every 3-cycle is also in N. Now, suppose that S_n had an Abelian tower of subgroups

$$\{1\} = G_n \trianglelefteq G_{n-1} \trianglelefteq \cdots \trianglelefteq G_0 = S_n$$

Then applying our above observation repeatedly, we'd conclude that every 3-cycle is in G_n, which is impossible. □

Lemma 8.12.9. *If p is prime, then the only elements of order p in S_p are p-cycles.*

Proof. This is a simple consequence of writing permutations as the product of disjoint cycles, and noting that the order of an n-cycle is n. □

Lemma 8.12.10. *S_n is generated by $(12), (23), \ldots, (n-1n)$.*

Proof. By induction on n. The case $n = 2$ is obvious. Now, suppose it holds for S_{n-1}. We first show that all transpositions are generated by $(12), (23), \ldots, (n-1n)$. We can inductively assume it holds for any transposition not involving n. Thus, we need only consider transpositions of the form (an), where $a \neq n$. If $a = n-1$ then we're done. Otherwise, we know by the inductive hypothesis that $(an-1)$ is in our generated set, and

$$(n-1n)(an-1)(n-1n) = (an)$$

as required.

Now, let $\sigma \in S_n$, and give it a disjoint cycle decomposition $\sigma = c_1 \cdots c_n$. Without loss of generality, only c_1 contains n, and hence $c_2 \cdots c_n$ is in our generated subgroup by the inductive hypothesis. Writing

$$c_1 = (na_1a_2 \cdots a_r)$$

we get by the inductive hypothesis that $(a_1a_2 \cdots a_r)$ is in our subgroup, along with all transpositions and hence

$$(a_1a_2 \cdots a_r)(a_rn) = c_1$$

implies the desired result. □

Lemma 8.12.11. *Suppose p is prime. Then S_p can be generated by any p-cycle and transposition.*

Proof. Without loss of generality, let $\sigma = (123 \cdots p)$ be the p-cycle and $\tau = (1n)$ our transposition, where $n \neq 1$. By raising our σ to the right power and re-ordering, we can assume that $n = 2$. We get

$$\sigma\tau\sigma^{-1} = (23)$$

and

$$\sigma(23)\sigma^{-1} = (34)$$

Continuing this process, we see that $(nn+1)$ is in the generated subgroup for any $1 \leq n < p$. The result then follows by the previous lemma. □

Lemma 8.12.12. *Suppose $f \in \mathbb{Q}[x]$ is irreducible of prime degree p, and has exactly two non-real roots. Then* $\mathrm{Gal}(f) = S_p$.

Proof. We know for certain that $\mathrm{Gal}(f) \subset S_p$, and that $p \mid |\mathrm{Gal}(f)|$. Thus, by Sylow's theorems, there exists a p-cycle $(123\cdots p) \in \mathrm{Gal}(f)$. The complex conjugate is in $\mathrm{Gal}(f)$, and by assumption must be of order two and hence a transposition. The result then follows by lemma 8.12.11. □

Theorem 8.12.13. *The general quintic equation is not solvable over* $\mathbb{Q}$.

Proof. The polynomial $f(x) = x^5 - 4x + 2$ is irreducible over $\mathbb{Q}$ by the Eisenstein criteria, and can be shown fairly easily to have three real roots. □

Just as Galois would have wanted.

I end off this section by finally paying off a debt from back in the ring theory chapter.

Theorem 8.12.14 (Fundamental Theorem of Algebra)**.** $\mathbb{C}$ *is algebraically closed.*

Proof. We know that $\mathbb{C} = \mathbb{R}(i)$, where i is a root of $x^2 + 1$. By direct computation, one can show that every complex number has a square root, and hence that every degree two polynomial in $\mathbb{C}[x]$ is reducible. We know that every finite extension of $\mathbb{C}$ is separable, and hence any splitting field of a polynomial $f \in \mathbb{C}[x]$ is a Galois extension of $\mathbb{C}$, and hence a Galois extension of $\mathbb{R}$. Call this splitting field K_f. Let $G = \mathrm{Gal}(K_f/\mathbb{R})$. Since we have complex conjugation, we know that G has a Sylow-2 subgroup, call it H. K_f^H must then be an extension of $\mathbb{R}$ of odd degree, since it is Galois and $\mathrm{Gal}(K_f^H/\mathbb{R}) = \mathrm{Gal}(K_f/\mathbb{R})/H$. By the primitive element theorem, there exists some $a \in K_f^H$ such that $K_f^H = \mathbb{R}(a)$. Then $\min(a, \mathbb{R})$ must be of odd degree, so since every polynomial in $\mathbb{R}[x]$ of odd degree has a real root we conclude that $a \in \mathbb{R}$ and hence $K_f^H = \mathbb{R}$. But this in turn implies that $H = G$, and hence that $[K_f : \mathbb{R}]$ is a power of two and therefore $[K_f : \mathbb{C}]$ is a power of two or equals one. But in the former case since $\mathrm{Gal}(K_f/\mathbb{C})$ would have a subgroup of order a factor of two less than $\mathrm{Gal}(K_f/\mathbb{C})$ by Sylow's theorems, we'd get that $\mathbb{C}$ has a degree two extension, which is impossible. Thus, $[K_f : \mathbb{C}] = 1 \Rightarrow K_f = \mathbb{C}$. □

8.13 Transcendental Extensions*

This section is based on a similar one from [Lan05], along with chapter nine of [Mil22]. Earlier in this chapter, we started focusing almost exclusively on algebraic field extensions. Here, we study the exact opposite, which we call *transcendental* extensions.

Start with and field extension K/F. We can induce a partial order on algebraically independent subsets of K, which we denote Σ, by inclusion. It is pretty clear that, by Zorn's lemma, Σ has a maximal element $U \subset K$. We call this U a *transcendence basis* of K.

Definition 8.13.1. Let K/F be a field extension and $X \subset K$ a set. We say that $y \in K$ is algebraically dependent on X over F if there exists some $p \in F[z_1, \ldots, z_{n+1}]$ and $x_1, \ldots, x_n \in X$ such that $p(x_1, \ldots, x_n, z_{n+1}) \neq 0$ but $p(x_1, \ldots, x_n, y) = 0$. We say that a set Y is dependent on X if every element of Y depends on X.

Lemma 8.13.2. *Suppose $X = \{x_1, \ldots, x_m\}$ be a subset of K over F. If $y \in K$ is algebraically dependent on X over F but not $\{x_1, \ldots, x_{m-1}\}$, then x_m is algebraically dependent on $\{x_1, \ldots, x_{m-1}, y\}$ over F.*

Proof. Let $p \in F[z_1, \ldots, z_{m+1}]$ be the non-zero polynomial such that $p(x_1, \ldots, x_m, y) = 0$ and $p(x_1, \ldots, x_m, z) \neq 0 \in F[z]$. We may write p in the form

$$p(z_1, \ldots, z_{m+1}) = \sum_{i=0}^{n} g_i(z_1, \ldots, z_{m-1}, z_{m+1}) z_m^{n-i}$$

where $g_i \in F[z_1, \ldots, z_m]$. Plugging in, we get

$$p(x_1, \ldots, x_m, z) = \sum_{i=0}^{n} g_i(x_1, \ldots, x_{m-1}, z) x_m^{n-i}$$

Thus, some $g_i(x_1, \ldots, x_{m-1}, z) \neq 0$. Since y is not algebraically dependent on $X \backslash \{x_m\}$, $g_i(x_1, \ldots, x_{m-1}, y) \neq 0$. Thus, it follows that x_m is algebraically dependent on $\{x_1, \ldots, x_{m-1}, y\}$. □

Lemma 8.13.3. *Algebraic dependence of sets over F is transitive.*

Proof. Suppose $A, B, C \subset K$, C depends on B and B depends on A. Then $F(C)/F(B)/F(A)$ is a tower of algebraic field extensions. Hence, $F(C)/F(A)$ is algebraic. Pick any $c \in C$. Then there exists a polynomial $p \in F(A)[x]$ such that $p \neq 0$ and $p(c) = 0$. Note that any element of $F(A)$ may be written as a ration of polynomials in $F[A]$. Multiplying through by the denominators of each coefficient in p, we get the desired result. □

Lemma 8.13.4. *Suppose that $A = \{a_1, \ldots, a_m\}, B = \{b_1, \ldots, b_n\}$ are subsets of K such that A is algebraically independent over F, but algebraically dependent on B over F. Then $m \leq n$.*

Proof. Suppose $m \geq n$. Re-order the elements of A, B so that the common elements are the firsts k entries (potentially none). That is, we get $B = \{a_1, \ldots, a_r, b_{r+1}, \ldots, b_n\}$. By assumption, a_{r+1} is algebraically dependent on B, but not $\{a_1, \ldots, a_r\}$. Thus, there exists some minimal $r' \geq r + 1$ such that a_{r+1} depends on $\{a_1, \ldots, a_r, b_{r+1}, \ldots, b_{r'}\}$ but not $\{a_1, \ldots, a_r, b_{r+1}, \ldots, b_{r'-1}\}$. By lemma 8.13.2, it follows that $b_{r'}$ is algebraically dependent on $B_1 = B \backslash \{b_{r'}\} \cup \{\alpha_{r+1}\}$, and hence all of B is algebraically dependent on B_1. Thus, by lemma 8.13.3, A is algebraically dependent on B_1. Repeating this process, we eventually get $A = B_i$, and hence $n = m$. □

Theorem 8.13.5. *Every transcendence basis of K/F has the same cardinality.*

Proof. The case where K/F has a finite transcendence basis follows immediate by lemma 8.13.4. Now, suppose that K/F does not have a finite transcendence basis. Let $X = \{x_i\}_{i \in I}, Y = \{y_j\}_{j \in J}$ be a pair of transcendence bases of K, and assume without loss of generality that $|I| \leq |J|$. It suffices to prove that $|J| \leq |I|$. For each $j \in J$, let $U_j \subset I$ be a finite subset of such that y_j is algebraically dependent on $X' = \{x_i\}_{i \in U_j}$ over F. Then Y is algebraically dependent on X' over F, and K is algebraically dependent on Y over F, so K is algebraically dependent on X' over F. Thus, by the maximality of X, $X' = X$, from which the result follows. □

Note. If S is a transcendence basis of K over F then K is algebraic over $F(S)$, and that any algebraic subset of K is contained in a transcendence basis. These are simple consequences of the definitions and Zorn's lemma respectively.

Definition 8.13.6. The transcendence dimension of a field extension is the cardinality of any transcendence basis.

The last thing we'll cover here is the decomposition of any field extension into a transcendental and algebraic extension. Of course, we need to explain what we mean by this first.

Definition 8.13.7. A field extension K/F is called purely transcendental if there exists a transcendence basis S of K over F such that $F(S) = K$.

It's pretty clear that the desired decomposition is to take a transcendence basis S of K/F, then decompose into the tower $K/F(S)/F$, which will be a purely transcendental followed by an algebraic field extension. There's one last thing I want to mention about transcendence and algebraic elements before ending off.

Theorem 8.13.8. *Let K/F be a field extension. The subset L of elements in K algebraic over F is a field.*

Proof. It's clear that $1, 0 \in L$. Now, suppose that $a, b \in L$. Then $F(a+b), F(ab) \subset F(a, b)$ are both finite and hence algebraic extensions, so $a + b, ab \in F$. Since $-a, a^{-1} \in F(a)$, they are in L, giving the desired result. □

8.14 Infinite Galois Groups*

We end off by talking briefly about the Galois groups of non-finite field extensions, following a similar section in [Lan05]. To do this, we'll primarily need to build up some tools from group theory.

Theorem 8.14.1. *Let I be a partially ordered set, $\{G_i\}_{i\in I}$ a collection of groups, and for each pair $i \leq j \in I$ take a homomorphism $\varphi_{ij} : G_i \to G_j$ such that if $i \leq j \leq k$ then $\varphi_{ik} = \varphi_{jk} \circ \varphi_{ij}$. This collection of groups and maps has a limit in the category **Grp**.*

Proof. We define our limit in the following manner. Start with

$$G = \{(g_i)_{i\in I} \mid g_i \in G_i, i \leq j \Rightarrow \varphi_{ij}(g_i) = g_j\}$$

We claim that this is a subgroup of $\prod_{i\in I} G_i$. Indeed, pick any $(g_i)_{i\in I}, (h_i)_{i\in I} \in G$. Then for any $i \leq j$ we get

$$\varphi_{ij}(g_i h_i) = \varphi_{ij}(g_i)\varphi_{ij}(h_i) = g_j h_j$$

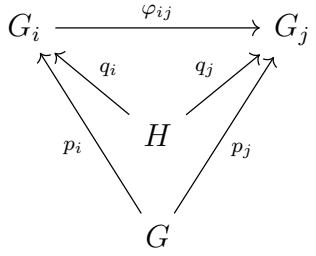
and

$$\varphi_{ij}(g_i^{-1}) = \varphi_{ij}(g_i)^{-1} = g_j^{-1}$$

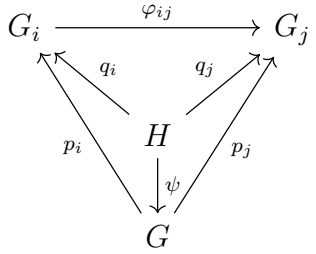
as required. We can define homomorphisms $p_i : G \to G_i$ by projection. We claim that G along with $\{p_i\}_{i\in I}$ is the desired limit. First suppose that $g = (g_i)_{i\in I} \in G$, and $i \leq j$. Then

$$(\varphi_{ij} \circ p_i)(g) = \varphi_{ij}(g_i) = g_j = p_j(g)$$

as required. Now, suppose that H is a group and $\{q_i\}_{i\in I}$ a set of homomorphisms such that the diagram

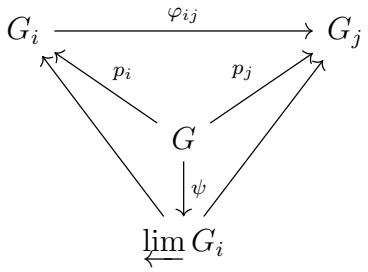


commutes for any $i \leq j$. We need to find a (unique) homomorphism $\psi : H \to G$ such that



commutes as well. In particular, we just need to satisfy the condition $p_i \circ \psi = q_i$ for any $i \in I$. For some $h \in H$, let $(g_i)_{i\in I} = \psi(h)$. Then our condition implies that $g_i = q_i(h)$. This fully defines the homomorphism, as required. □

In this case, we denote G by $G = \varprojlim G_i$. Now, let K/F be an infinite Galois extension. Let $\{L\}_{i\in I}$ be the set of all finite sub-extensions which are Galois. Set $G = \mathrm{Gal}(K/F), H_i = \mathrm{Gal}(K/L_i), G_i = \mathrm{Gal}(L_i/F)$. Then by proposition 8.7.6, we get a standard projection homomorphism $p_i : G \to G_i$ given by the projection map $G \mapsto G/H_i \cong G_i$. We can also define a partial order on the set $\{G_i\}_{i\in I}$ by inclusion, and hence define our $\varphi_{ij} : G_i \to G_j$ to be the inclusion maps. We can set these all up to be consistent, so we get a unique homomorphism $\psi : G \to \varprojlim G_i$ such that the following diagram commutes



where the unlabelled arrows are associated projection maps.

Theorem 8.14.2. *ψ is an isomorphism, so $G \cong \varprojlim G_i$.*

Proof. First, suppose $\sigma \in \ker(\psi)$. Then $p_i(\sigma) = \mathrm{Id}_{G_i}$ for every $i \in I$, and hence σ is fixed on every finite Galois sub-extension of K. Since every element of K is in some such sub-extension, it follows that $\sigma = \mathrm{Id}_K$ and hence ψ is injective. Now, pick any $g \in \varprojlim G_i$. We

will define $\sigma : G \to G$ in the following manner. First, note that by our prior construction, $g = (\sigma_i)_{i \in I}$, where each $\sigma_i \in G_i$ and if $i \leq j$ then $\varphi_{ij}(\sigma_i) = \sigma_j$. If $a \in K$, let L_i, L_j be any Galois field extensions containing a. Let $L_k = L_i L_j$, which by proposition 8.8.3 is Galois. Then by commutativity of the diagram

$$\varphi_{ik}(\sigma_i)(a) = \sigma_k(a) = \varphi_{jk}(\sigma_j)(a)$$

But of course each φ map is just the identity on its source, so $\sigma_i(a) = \sigma_j(a)$. Thus, we can define $\sigma : K \to K$ by $\sigma(a) = \sigma_i(a)$. This restricts, by assumption, to an automorphism on every finite Galois sub-extension of K, and hence is an automorphism on K. Since clearly $\psi(\sigma) = g$, ψ is surjective. □

Note. There's a lot of questions of choice of maps here that I'm glossing over. [Lan05] deals with them a bit more thoroughly, though perhaps at the expense of the idea of the theorem being less clear.

The consequences of this are quite simple, namely that any infinite Galois extension is fully characterized by its finite Galois sub-extensions. If you know some topology and are interested in studying infinite Galois extensions further, it may be worth looking at chapter 7 of [Mil22].

Chapter 9

Commutative Algebra

Chapter 10

Homology

Bibliography

[rus99] Fom: Russel paradox for naive category theory, https://cs.nyu.edu/pipermail/fom/1999-May/003117.html, 1999, Accessed: 2024-31-07.

[m3m17] The yoneda lemma, https://www.math3ma.com/blog/the-yoneda-lemma, 2017, Accessed: 2024-26-07.

[Bad10] B. Badzioch, Math 619 : Abstract algebra i lecture notes, week 13, http://www.math.buffalo.edu/~badzioch/MTH619/Lecture_Notes.html, 2010.

[BS12] S. Burris and H. Sankappanavar, *A Course in Universal Algebra*, mellenium ed., 2012.

[ET20] D. Epelbaum and A. Trisal, Introductory category theory notes, https://web.math.ucsb.edu/~atrisal/category%20theory.pdf, 2020, Accessed: 2024-10-08.

[Fol99] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, second ed., John Wiley and Sons, 1999.

[Gha99] E. Ghate, Studying mobile context-aware social services in the wild, in *Cyclotomic Fields and Related Topics: Proceedings of the Summer School*, Bhaskaracharya Pratishthana, Pune, India, 1999, pp. 135–146.

[Gub21] N. Gubareni, *Introduction to Modern Algebra and Its Applications*, 1. ed., CRC Press, 2021.

[Jac09] N. Jacobson, *Basic Algebra I*, second ed., Dover Publications, Mineola, NY, 2009.

[Lan10] S. M. Lane, *Categories for the Working Mathematician*, second ed., Springer-Verlag, New York, NY, 2010.

[Lan05] S. Lang, *Algebra*, third ed., Springer-Verlag, New York, NY, 2005.

[Lan11] D. Lantz, Math 320 : Abstract algebra i, presentation on the structure theorem for finite abelian groups, https://math.colgate.edu/math320/dlantz/, 2011.

[Mil22] J. S. Milne, Fields and galois theory (v5.10), 2022, Available at www.jmilne.org/math/, p. 144.

[nLa24a] nLab authors, complete category, https://ncatlab.org/nlab/show/complete+category, August 2024, Revision 7.

[nLa24b] nLab authors, continuous functor, https://ncatlab.org/nlab/show/continuous+functor, August 2024, Revision 15.

[nLa24c] nLab authors, hom-functor preserves limits, https://ncatlab.org/nlab/show/hom-functor+preserves+limits, August 2024, Revision 11.

[nLa24d] nLab authors, limits of presheaves are computed objectwise, https://ncatlab.org/nlab/show/limits+of+presheaves+are+computed+objectwise, August 2024, Revision 4.

[nLa24e] nLab authors, representable functor, https://ncatlab.org/nlab/show/representable+functor, August 2024, Revision 57.

[Rom07] S. Roman, *Advanced Linear Algebra*, third ed., Springer-Verlag, New York, NY, 2007.

[Sil23] L. Silberman, Math 412: Advanced linear algebra lecture notes, https://personal.math.ubc.ca/~lior/teaching/2223/412_W23/, 2023.

[Uni13] T. Univalent Foundations Program, *Homotopy Type Theory: Univalent Foundations of Mathematics*, https://homotopytypetheory.org/book, Institute for Advanced Study, 2013.