MATH 200 LECTURE NOTES

DAN ROGALSKI

1. Crash course on groups

These notes are for a graduate course in algebra which assumes you have seen an undergraduate course in algebra already. Generally a first undergraduate course in algebra concentrates on groups, so basic group theory is the material which we will review most quickly. The purpose of this first section is to remind you of the basic definitions, examples, and theorems about groups.

Definition 1.1. Let G be a set with a binary operation $*$. Then G is a *group* with respect to that operation if

(1) $*$ is associative.
(2) There is an element $e \in G$ such that $e * a = a = a * e$ for all $a \in G$.
(3) For all $a \in G$ there is an element $b \in G$ such that $a * b = e = b * a$.

For your info, a structure satisfying only axiom (1) is a *semigroup*, and a structure satisfying only (1) and (2) is a *monoid*. We will refer to these weaker structures only in passing.

The operation $*$ is usually called the *multiplication* in G, e is the *identity element*, and the $b \in G$ such that $a * b = e = b * a$ is called the *inverse* of a. If we need to emphasize the operation in the group G, we write it as the pair $(G, *)$. But usually the operation is clear and we omit the $*$, writing $a * b$ simply as ab. We also usually write 1 for e, as the identity element in many standard groups of numbers under multiplication is already called that. We write the inverse of a as a^{-1}.

We have referred to "the" identity and "the" inverse of a. This is appropriate since they are uniquely determined: if e', e are identity elements, then $e' = e'e = e$. If b, b' are both inverses of a, then $b = be = b(ab') = (ba)b' = eb' = b'$.

We use throughout the following standard names for the traditional number systems one uses in mathematics: the natural numbers $\mathbb{N} = \{0, 1, 2 \dots\}$ (our convention is that 0 is a natural number); the integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$; the rational numbers $\mathbb{Q} = \{p/q \,|\, p, q \in \mathbb{Z} \text{ and } q \neq 0\}$; the real numbers $\mathbb{R}$; and the complex numbers $\mathbb{C} = \{a + bi | a, b \in \mathbb{R}\}$ (where $i^2 = -1$). We take the existence of the real numbers $\mathbb{R}$ as a given; in an analysis course you see how they can be constructed

from the rational numbers through a limiting process. Later on we will introduce formal concepts which recover the construction of $\mathbb{Q}$ from $\mathbb{Z}$ and the construction of $\mathbb{C}$ from $\mathbb{R}$.

We can get some simple examples of groups from these familiar number systems.

Example 1.2. $(\mathbb{Q}-\{0\},\cdot)$, $(\mathbb{R}-\{0\},\cdot)$, and $(\mathbb{C}-\{0\},\cdot)$ are all groups under multiplication. The associative property is a basic fact about multiplication in these number systems. It is easy to check that 1 is an identity element and that $a^{-1}=1/a$ exists for all nonzero a. On the other hand, $(\mathbb{Z}-\{0\},\cdot)$ is a monoid but not a group, as only 1 and -1 have multiplicative inverses in $\mathbb{Z}$.

Example 1.3. $(\mathbb{Z},+)$, $(\mathbb{Q},+)$, $(\mathbb{R},+)$, and $(\mathbb{C},+)$ are all groups under addition, with identity element 0 and where the inverse of a is $-a$. On the other hand, $(\mathbb{N},+)$ is not a group.

Given a group which is a familiar set with an operation usually called addition and written $+$, as in Example 1.3, all of our notational conventions are modified. As in the previous example, we always write the identity element as 0 and the inverse of a as $-a$, and refer to it as the *additive inverse* to stress this. Of course we also always write $a+b$ and do not omit the symbol for the operation—writing ab for the sum would be way too confusing. Given a group in the abstract, however, that is, something that satisfies the definition but without any further knowledge about it and its operation, we will use the multiplicative notation.

A somewhat more interesting example comes from considering modular arithmetic.

Example 1.4. Fix $n\geq 1$. We can define an equivalence relation on $\mathbb{Z}$ by $a\sim b$ if $a\equiv b \mod n$, that is, $b-a=nq$ for some $q\in\mathbb{Z}$. This partitions $\mathbb{Z}$ into n equivalence classes, called *congruence classes*. We write the congruence class containing a as $\overline{a}$, so formally $\overline{a}=\{a+nq|q\in\mathbb{Z}\}$. If we need to emphasize what n is we might also write this as $\overline{a}_n$. Another common notation for the congruence class of a is $[a]$ or $[a]_n$.

The set $\mathbb{Z}_n=\{\overline{a}|a\in\mathbb{Z}\}=\{\overline{0},\overline{1},\ldots,\overline{n-1}\}$ is a group under the operation $+$ of addition of congruence classes, defined by $\overline{a}+\overline{b}=\overline{a+b}$. The identity element is $\overline{0}$ and the (additive) inverse of $\overline{a}$ is $\overline{-a}$. We call $(\mathbb{Z}_n,+)$ the additive group of integers modulo n.

The addition rule $\overline{a}+\overline{b}=\overline{a+b}$ can be viewed in two ways, both of which are useful. One should show that it is *well-defined*, because when we write $\overline{a}$ we are referring to the class by one of its *representatives* a, but we could equally well refer to it by a different representative, say $a+nq$, since $\overline{a+nq}=\overline{a}$. Whenever an operation is defined by referring to representatives of sets, one needs to check that choosing different representatives would not lead to a different result. In this case, one needs that if $\overline{a'}=\overline{a}$ and $\overline{b'}=\overline{b}$, then $\overline{a+b}=\overline{a'+b'}$, which is an easy exercise in arithmetic.

We can also think of $\overline{a} + \overline{b} = \overline{a+b}$ as an addition rule *on sets*; we add each of the elements of $\overline{a}$ to each of the elements of $\overline{b}$, and take the entire set that results; this set is another congruence class which is $\overline{a+b}$, as the reader may check. We will come back to this point shortly when we review factor groups.

To give a more explicit example of the above, suppose $n = 5$. Then $\overline{2} = \{\dots, -8, -3, 2, 7, 12, \dots\}$ and $\overline{3} = \{\dots, -7, -2, 3, 8, 13, \dots\}$. By definition $\overline{2} + \overline{3} = \overline{5} = \overline{0} = \{\dots, -10, -5, 0, 5, 10, \dots\}$. If we take any element of $\overline{2}$ and add it to an element of $\overline{3}$, then $\overline{0}$ is the unique congruence class that contains the result. Hence $\overline{0}$ is also the set arising from adding each of the elements in $\overline{2}$ to each of the elements in $\overline{3}$ and collecting the results.

One way of getting interesting further examples of groups is to start with a monoid M, where elements need not have inverses, and simply remove the elements without inverses.

Lemma 1.5. *Let M be a monoid. Then the subset*

$$G(M) = \{a \in M | \textit{there exists } b \in M \textit{ such that } ab = 1 = ba\}$$

of M is a group under the restriction of the operation of M to the subset $G(M)$.

Proof. If $a, b \in G(M)$, say with $ac = 1 = ca$ and $bd = 1 = db$, then $(ab)(dc) = a(bd)c = a1c = ac = 1$, and similarly $(dc)(ab) = 1$, so that $ab \in G(M)$. This shows that the binary operation of M does restrict to give a binary operation on the subset $G(M)$. It is clear that associativity still holds after restricting to a subset, and 1 is in $G(M)$ (since $(1)(1) = 1$) and still behaves as an identity for the subset. Finally, inverses exist for all elements in $G(M)$ by construction since if $a \in G(M)$, say a has an inverse c, then c has the inverse a so that $c \in G(M)$ also. □

We can recover Example 1.2 using Lemma 1.5, for instance. Each of $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ is a monoid under multiplication with identity 1. In each case 0 is the only element without a multiplicative inverse, so throwing it away we get a group.

Here are some other examples of groups that arise naturally by applying this construction.

Example 1.6. Let F be a field. We will define this notion later when we study rings; if you have forgotten the definition, for now simply take F to be $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ when fields are mentioned. Let $M_n(F)$ be the set of all $n \times n$ matrices whose entries are elements in F. We write an element A of $M_n(F)$ as (a_{ij}), which indicates the matrix whose entry in the ith row and jth column is $a_{ij} \in F$. Now $M_n(F)$ is a monoid under matrix multiplication, defined by $(a_{ij})(b_{ij}) = (c_{ij})$ where $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$. The identity element is the identity matrix $I = (e_{ij})$ where $e_{ij} = 1$ if $i = j$ and $e_{ij} = 0$ if $i \neq j$.

Applying the construction above, we get that the subset

$$G(M_n(F)) = \{A \in M_n(F) | \text{ there exists } B \in M_n(F) \text{ s.t. } AB = I = BA\}$$

is a group under matrix multiplication. It is called the *general linear group* over F and written as $\mathrm{GL}_n(F)$. By a standard result in linear algebra, an element of $M_n(F)$ has a multiplicative inverse if and only if it is a nonsingular matrix, or equivalently has nonzero determinant, so we also have $\mathrm{GL}_n(F) = \{A \in M_n(F) | \det(A) \neq 0\}$.

Let $f : X \to Y$ be a function between two sets. Recall that we say f is *injective* if $f(x) = f(y)$ implies $x = y$ for $x, y \in X$. We say that f is *surjective* if for all $y \in Y$ there exists $x \in X$ such that $f(x) = y$. Finally a function f is *bijective* if it is injective and surjective.

Example 1.7. Let X be any set. Consider the set $\mathrm{Fun}(X, X)$ of all functions from X to itself. If $f : X \to X$ and $g : X \to X$ are functions, then $f \circ g : X \to X$ is the function with $[f \circ g](x) = f(g(x))$. Note that we will use the standard notation for composition in this course, sometimes called *right to left* composition because in the expression $f \circ g$, the function g is performed first, and then the function f. This is the most natural definition because of the standard convention of writing $f(x)$ for the image of x under f, that is, the function name is written on the left of the argument. There is nothing inevitable about that choice and in fact some authors choose the opposite convention, in which case they also choose left to right composition.

Now $\mathrm{Fun}(X, X)$ is a monoid, where the operation is the composition $\circ$. The identity element is the *identity function* $1_X : X \to X$ where $1_X(x) = x$ for all $x \in X$. Thus

$$G(\mathrm{Fun}(X, X)) = \{f : X \to X | \text{there is } g \text{ such that } f \circ g = 1_X = g \circ f\}$$

is a group under composition called the *symmetric group* on X and written $\mathrm{Sym}(X)$. The functions with multiplicative inverses under composition are precisely the bijective functions, so we also have $\mathrm{Sym}(X) = \{f : X \to X | f \text{ is bijective}\}$. The functions in $\mathrm{Sym}(X)$ are also called *permutations* of X and $\mathrm{Sym}(X)$ is called the *permutation group* of X.

As a special case, when $X = \{1, 2, \ldots, n\}$ is the set of the first n positive numbers, we write the group $\mathrm{Sym}(X)$ as S_n and call it the *nth symmetric group.*

Example 1.8. Let $\mathbb{Z}_n = \{\overline{0}, \overline{1}, \ldots, \overline{n-1}\}$ be the set of congruence classes modulo n, as in Example 1.4. There is also a multiplication of congruence classes, where we put $\overline{a}\,\overline{b} = \overline{ab}$. Again it is straightforward to check that this definition is independent of the choice of representatives for the

congruence classes. This is an associative operation with identity element $\overline{1}$, so $\mathbb{Z}_n$ is a monoid under multiplication. Note that $\overline{a}\overline{b} = \overline{b}\overline{a}$ for all a, b. Thus the subset

$$U_n = \{\overline{a} \in \mathbb{Z}_n | \text{there is } \overline{b} \in \mathbb{Z}_n \text{ such that } \overline{a}\overline{b} = \overline{1}\}$$

is a group under multiplication, called the *units group* of $\mathbb{Z}_n$.

We can say more about exactly which congruence classes are in U_n. If $\overline{a}\overline{b} = \overline{ab} = \overline{1}$, then $ab = 1 + nq$ for some $q \in \mathbb{Z}$. Thus $ab - nq = 1$ and it follows that $\gcd(a, n) = 1$. Conversely, if $\gcd(a, n) = 1$, then since the gcd is a $\mathbb{Z}$-linear combination we get $ba + qn = 1$ for some $b, q \in \mathbb{Z}$. Then $\overline{b}\overline{a} = \overline{ba} = \overline{1}$. We conclude that $U_n = \{\overline{a} \in \mathbb{Z}_n | \gcd(a, n) = 1\}$.

We now review some of the most basic properties of a group. Given a set X, we write $|X|$ for the cardinality of the set, as usual. In particular, for a group G, the number $|G|$ is called the *order* of the group. For example, consider the group U_n. Recall that *Euler φ function* is $\varphi : \mathbb{N} \to \mathbb{N}$ where $\varphi(n)$ is the number of integers a with $1 \leq a \leq n$ such that $\gcd(a, n) = 1$. Thus by definition we have that $|U_n| = \varphi(n)$. For a specific example, note that $U_{12} = \{\overline{1}, \overline{5}, \overline{7}, \overline{11}\}$ and $\phi(12) = 4$. The study of finite groups, i.e. those with finite order, tends to have a rather different flavor than the study of infinite groups. We will focus much of our attention on finite groups below.

Let G be a group. Two elements $a, b \in G$ are said to *commute* if $ab = ba$. If all pairs of elements in a group commute, we say that G is *abelian*; otherwise G is *non-abelian*. A more obvious name for the abelian property would be commutative, and in fact that is the name given to the analogous property in ring theory. In group theory the term abelian was chosen to honor the mathematician Niels Henrik Abel, whose work on the unsolvability of the quintic equation was a precursor to the development of group theory. All of the examples of groups given so far are abelian except for $\mathrm{GL}_n(F)$, which is non-abelian if $n \geq 2$, and $\mathrm{Sym}(X)$, which is nonabelian as long as X has at least three elements. In general, non-abelian groups are much more difficult to understand. For example, we will see that abelian groups with finitely many elements can all be described rather easily. The structure of finite non-abelian groups, on the other hand, attracted the intense efforts of many mathematicians in the latter half of the twentieth century, especially to try to classify finite simple groups. That project was declared complete in the 1980's but the details are so technical that they are accessible only to specialists.

1.1. **Subgroups and further examples.**

Definition 1.9. Let G be a group. A nonempty subset $H \subseteq G$ is a *subgroup* if (i) $ab \in H$ for all $a, b \in H$; and (ii) $a^{-1} \in H$ for all $a \in H$. When H is a subgroup of a group G we sometimes indicate this by writing $H \leq G$.

In words, a subset of a group is a subgroup if it is closed under products and closed under inverses. Some people prefer to use the following alternate definition: H is a subgroup if (i)$'$: $ab^{-1} \in H$ for all $a, b \in H$. It is easy to check that this single condition (i)$'$ is equivalent to (i) and (ii). Having only one condition is more elegant, though in practice the work required to check this single condition usually amounts to the same as checking (i) and (ii) separately.

If H is a subgroup of G, then we claim that H is itself a group under the same operation restricted to H. Note that condition (i) guarantees that the binary operation of G restricts to a binary operation on H, which is necessarily also associative. Since H is nonempty, picking any $a \in H$ we have $a^{-1} \in H$ by (ii) and hence $1 = aa^{-1} \in H$ by (i), so $1 \in H$ and clearly 1 is still an identity element for H. Finally, (ii) ensures that every $a \in H$ has an inverse element in H, so H is a group as claimed. The reader may check conversely that a subset of G is a group under the restricted binary operation precisely when it is a subgroup as defined above.

In the next examples we define some new interesting groups as subgroups of the groups we have defined so far.

Example 1.10. Let F be a field and let $G = \mathrm{GL}_n(F)$. Define

$$\mathrm{SL}_n(F) = \{A \in \mathrm{GL}_n(F) | \det(A) = 1\}.$$

Then $\mathrm{SL}_n(F)$ is a subgroup of $\mathrm{GL}_n(F)$ called the *special linear group*. To check that it is a subgroup, if $A, B \in \mathrm{SL}_n(F)$, so that $\det(A) = \det(B) = 1$, just note that $\det(AB^{-1}) = \det(A)\det(B^{-1}) = \det(A)\det(B)^{-1} = 1$ so that $AB^{-1} \in \mathrm{SL}_n(F)$ as well.

Example 1.11. Let I be the identity matrix in $\mathrm{GL}_2(\mathbb{C})$. We also define

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \qquad \text{and} \qquad C = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

in $\mathrm{GL}_2(\mathbb{C})$. Let Q_8 be the subset of $\mathrm{GL}_2(\mathbb{C})$ consisting of the 8 matrices $\{\pm I, \pm A, \pm B, \pm C\}$.

The matrices A, B, and C are easily checked to satisfy the following rules for multiplication: $A^2 = B^2 = C^2 = -I$; $AB = C = -BA$; $BC = A = -CB$; and $CA = B = -AC$. Using these rules it easily follows that Q_8 is closed under taking products and inverses, and so is a subgroup of $\mathrm{GL}_2(\mathbb{C})$. You could also check that these 8 matrices are exactly those matrices in $\mathrm{GL}_2(\mathbb{C})$ that are either diagonal or anti-diagonal; have determinant 1; and have nonzero entries taken from the set

$\{1, -1, i, -i\}$. These properties are preserved under multiplication and taking inverses, so this set of matrices must be a subgroup for that reason. In fact Q_8 is also a subgroup of $\mathrm{SL}_2(\mathbb{C})$.

Often instead of thinking of Q_8 as a subgroup of $\mathrm{GL}_2(\mathbb{C})$, one thinks of it abstractly as a group with 8 elements $\{\pm 1, \pm i, \pm j, \pm k\}$ with multiplication rules $i^2 = j^2 = k^2 = -1, ij = k = -ji, jk = i = -kj, ki = j = -ik$. This is the traditional notation that is borrowed from the ring of quaternions invented by Hamilton, which we will describe later in the ring theory section. One could also just define Q_8 by these multiplication rules, but checking associativity directly is messy. Defining it as a subgroup of $\mathrm{GL}_2(\mathbb{C})$, as we did, has the advantage that associativity of the operation comes for free.

Example 1.12. Let n be a positive integer with $n \geq 3$. Define $\theta = 2\pi/n$. We define

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \text{and} \qquad S = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

inside the group $\mathrm{GL}_2(\mathbb{R})$. A matrix $A \in \mathrm{GL}_2(\mathbb{R})$ gives a linear transformation of the real plane $\mathbb{R}^2$ via the formula $v \mapsto Av$ for column vectors $v \in \mathbb{R}^2$. Under this correspondence R gives the counterclockwise rotation of the plane about the origin by θ radians, and S is the reflection of the plane about the y-axis.

Direct calculation shows that the matrices R and S satisfy the rules $R^n = I$; $S^2 = I$; and $SR = R^{-1}S$. Using these relations it is straightforward to see that the set of matrices

$$D_{2n} = \{R^i S^j \,|\, 0 \leq i \leq n-1, 0 \leq j \leq 1\}$$

is a subgroup of $\mathrm{GL}_2(\mathbb{R})$, consisting of $2n$ distinct elements. It is called the *dihedral group of* $2n$ *elements*. (Warning: some authors call this group D_n. We prefer to have the subscript label the number of elements in the group.)

The dihedral group arises naturally as a group of symmetries. If one takes a regular n-gon in the plane centered at the origin, such that the y-axis is an axis of symmetry for it, then the elements of D_{2n} are exactly those linear transformations of the plane which send the points of the n-gon bijectively back to itself. These transformations are also called *rigid motions* of the n-gon.

Similarly as in the example Q_8 above, when working with the group D_{2n} abstractly, it is useful simply to take it to be a group with $2n$ distinct elements of the form $\{a^i b^j | 0 \leq i \leq n-1, 0 \leq j \leq 1\}$ satisfying the rules $a^n = 1, b^2 = 1, ba = a^{-1}b$. This is essentially the point of view of a *presentation* of a group, which we will define and study more formally in a later section.

1.2. **Cosets and Factor Groups.** The following notation for products of subsets of a group is quite convenient.

Definition 1.13. Let G be a group and let X and Y be any subsets of G. Then we define $XY = \{xy \,|\, x \in X, y \in Y\}$.

When we apply the product notation to a subset with a single element x, we write the subset as x rather than the more formally correct $\{x\}$. As an example we have the following.

Definition 1.14. Let H be a subgroup of a group G. Given any $x \in G$, then $xH = \{xh | h \in H\}$ is the *left coset* of H with representative x. Similarly, $Hx = \{hx | h \in H\}$ is the *right coset* of H with representative x.

Note that cosets are named after which side of H the representative x is on. We will generally focus on left cosets. The theory of right cosets is completely analogous, and the reader can easily formulate and prove analogous versions for right cosets of the following results.

As always, the notation changes in a group G with addition operation $+$: for subsets X and Y the "product" becomes $X + Y = \{x + y | x \in X, y \in Y\}$. Given a subgroup H of G and $x \in G$, the corresponding left coset with representative x is written $x + H = \{x + h | h \in H\}$.

Here are the important basic facts about the left cosets in a general (multiplicative) group.

Proposition 1.15. *Let $H \leq G$, i.e. let H be a subgroup of a group G. For any $x, y \in G$, we have*

(1) *$xH = yH$ if and only if $y^{-1}x \in H$ if and only if $x^{-1}y \in H$.*
(2) *Either $xH = yH$ or else $xH \cap yH = \emptyset$.*
(3) *$|xH| = |H|$.*

Proof. Define a relation on elements of G by $x \sim y$ if $x^{-1}y \in H$. Then for any $x \in G$, $x^{-1}x = 1 \in H$, so $x \sim x$. If $x \sim y$, then $x^{-1}y \in H$. Since H is closed under inverses, $(x^{-1}y)^{-1} = y^{-1}x \in H$ and $y \sim x$. Finally, if $x \sim y$ and $y \sim z$, so $x^{-1}y \in H$ and $y^{-1}z \in H$, then $(x^{-1}y)(y^{-1}z) = x^{-1}z \in H$ since H is closed under products, and so $x \sim z$. We have shown that $\sim$ is an equivalence relation on G. Therefore G is partitioned into disjoint equivalence classes. Given $x \in G$, the equivalence class containing x is

$$[x] = \{y \in G | x \sim y\} = \{y \in G | x^{-1}y \in H\} = \{xh | h \in H\} = xH.$$

Thus the equivalence class containing x is precisely the left coset with representative x. Now (2) follows from the fact that the equivalence classes partition G, and (1) follows from the definition of the equivalence relation.

Now define a function $\theta : H \to xH$ by $\theta(h) = xh$. The function θ is injective, since if $\theta(h_1) = \theta(h_2)$, then $xh_1 = xh_2$, and multiplying by x^{-1} on the left yields $h_1 = h_2$. The function θ is also clearly surjective. Thus θ is a bijection and $|xH| = |H|$. □

Lagrange's Theorem, one of the most fundamental results in group theory, is an immediate consequence of the observations in the previous result. If H is a subgroup of a group G, we write $|G : H|$ for the number of distinct left cosets of H in G. We call $|G : H|$ the *index* of H in G.

Theorem 1.16. *(Lagrange's Theorem) Let G be a group and let $H \leq G$ be a subgroup. Then*

$$|G| = |H||G : H|.$$

In particular, if G is finite, then $|H|$ divides $|G|$.

Proof. By the previous proposition, G is partitioned by the distinct left cosets of G. Also, each left coset xH has size $|xH| = |H|$. Therefore G is the disjoint union of $|G : H|$ subsets, each of which has size $|H|$. The result follows. □

Definition 1.17. Let G be a group. For $x, g \in G$, the *conjugate* of x by g is gxg^{-1}. Note that g and x commute (i.e. $xg = gx$) if and only if $gxg^{-1} = x$. We also write ${}^g x = gxg^{-1}$ and think of g as "acting" on x on the left by conjugation. We use the same notation for subsets, so ${}^g X = \{gxg^{-1} | x \in X\}$.

Definition 1.18. A subgroup H of G is *normal* if ${}^g H = gHg^{-1} \subseteq H$ for all $g \in G$. In this case we write $H \trianglelefteq G$.

Example 1.19. Let $G = \mathrm{GL}_n(F)$ for some field F. Then $H = \mathrm{SL}_n(F)$ is a normal subgroup of G. For if $A \in G$ and $B \in H$, so $\det(A) \neq 0$ and $\det(B) = 1$, then $\det(ABA^{-1}) = \det(A)\det(B)\det(A)^{-1} = \det(B) = 1$. Thus $ABA^{-1} \in H$.

Example 1.20. If G is abelian, then any subgroup H of G is normal, since $ghg^{-1} = gg^{-1}h = h$ for all $g \in G$ and $h \in H$.

Proposition 1.21. *Let $H \leq G$. The following are equivalent:*

(1) *$H \trianglelefteq G$, i.e. ${}^g H \subseteq H$ for all $g \in G$.*
(2) *${}^g H = H$ for all $g \in G$.*
(3) *$gH = Hg$ for all $g \in G$.*
(4) *Every right coset of H is also a left coset of H.*

Proof. (1) $\implies$ (2). By definition we have ${}^gH \subseteq H$, or $gHg^{-1} \subseteq H$. Multiplying by g^{-1} on the left and g on the right gives $H \subseteq g^{-1}Hg$. Applying this to the element g^{-1} gives $H \subseteq gHg^{-1}$. Thus $H = gHg^{-1} = {}^gH$.

(2) $\implies$ (3). Multiplying $gHg^{-1} = H$ on the right by g gives $gH = Hg$.

(3) $\implies$ (4). This is trivial.

(4) $\implies$ (1). Given the right coset Hg, we know it is equal to xH for some x. Now $g \in Hg = xH$ and of course $g \in gH$, so $gH \cap xH \neq \emptyset$. By Proposition 1.15, $gH = xH$. Thus $gH = Hg$. Since g was arbitrary, we have (3). Now (3) implies (2) by multiplying $gH = Hg$ on the right by g^{-1}, and (2) trivially implies (1). □

Example 1.22. Let H be a subgroup of a group G such that $|G : H| = 2$. In this case, there are only two left cosets. Since one of the them is $H = 1H$, there other must be $G - H$. Similarly, the right cosets must be $H = H1$ and its complement $G - H$. We see that any right coset is a left coset, so $H \trianglelefteq G$ by the preceding proposition. We conclude that *every subgroup of index 2 is normal.*

We can now define the quotient of a group by a normal subgroup.

Proposition 1.23. *Let $H \trianglelefteq G$. The set $G/H = \{$the distinct left cosets of H in $G\}$ is a group under the operation $(aH) * (bH) = abH$. The identity element is $1H = H$ and $(aH)^{-1} = a^{-1}H$. Moreover, $|G/H| = |G : H|$.*

The group G/H is called the *factor group* or *quotient group* of G by H. We often read G/H as "G mod H".

Proof. The main content of the proposition is that the operation is well defined. To see this, suppose that $a'H = aH$ and $b'H = bH$, so we have chosen other representatives for these cosets. Then $a' = a' \in a'H = aH$ and so $a' = ah_1$ for some $h_1 \in H$. Similarly $b' = bh_2$ for some $h_2 \in H$. Now $h_1 b \in Hb = bH$ since H is normal, by Proposition 1.21. Thus $h_1 b = bh_3$ for some $h_3 \in H$. We now get $a'b' = ah_1 bh_2 = abh_3 h_2 \in abH$. By Proposition 1.15, this forces $a'b'H = abH$. Thus the product operation is well defined.

Once we have a well defined operation, it is trivial to check that it is associative (because the operation of G is) and that the identity and inverses are as indicated, so that G/H is a group. We have $|G/H| = |G : H|$ since this is the number of left cosets, which are the elements of G/H by definition. □

As stated, we defined the operation on left cosets in G/H by using representatives: take two cosets, multiply their representatives, and take the coset containing that product. Similarly as in

Example 1.4, we could also think of this as a *product of sets*. Namely, in the setup of Proposition 1.23, we could define $(aH) * (bH)$ to be the product $(aH)(bH)$, using our usual product of subsets of a subgroup. Since G is associative, product of subsets is associative. Hence $(aH)(bH) = a(Hb)H = a(bH)H = abHH = abH$, using that H is a normal subgroup. In this way we recover the formula for the product in G/H.

Example 1.24. Let $G = (\mathbb{Z}, +)$. Then $H = n\mathbb{Z} = \{qn | q \in \mathbb{Z}\}$ is clearly a subgroup of G, and it is normal automatically since G is abelian. The factor group G/H consists of additive cosets $\{a + H | a \in \mathbb{Z}\}$, with addition operation in G/H defined by $(a + H) + (b + H) = (a + b) + H$. The coset $a + H = a + n\mathbb{Z}$ is precisely the congruence class $\overline{a}$, and the addition operation on cosets is precisely the usual addition on congruence classes, $\overline{a} + \overline{b} = \overline{a + b}$. In this way the factor group $\mathbb{Z}/n\mathbb{Z}$ is identified with the group $(\mathbb{Z}_n, +)$ of integers mod n under addition.

Example 1.25. Consider the dihedral group $G = D_{2n} = \{1, a, a^2, \ldots, a^{n-1}, b, ab, \ldots, a^{n-1}b\}$, where $a^n = 1, b^2 = 1, ba = a^{-1}b$. Recall that a corresponds to a rotation and b to a reflection of real two space. Thus $H = \{1, a, a^2, \ldots, a^{n-1}\}$ is a subgroup of G called the *rotation subgroup*; it consists of those elements of G which are rotations. Since $|H| = n$ is is clear that $|G : H| = 2$ and so H has just two cosets, H and $bH = \{b, ab, \ldots, a^{n-1}b\}$ which consists of all of the reflections. Since H has index 2 in G, it is automatic that $H \trianglelefteq G$ by Example 1.22, so we can define the factor group $G/H = \{H, bH\}$. This factor group has multiplication rules $(H)(H) = H$, $(H)(bH) = bH$, $(bH)(H) = (bH)$, and $(bH)(bH) = H$, which exactly express the facts that a product (i.e. composition) of two rotations is a rotation; a product of a rotation and a reflection is a reflection; and a product of two reflections is a rotation.

1.3. **Products of subgroups and normalizers.** Suppose that H and K are subgroups of a group G. The product $HK = \{hk | h \in H, k \in K\}$ need not be a subgroup of G.

Example 1.26. Let $G = D_6$, which we think of as the set of 6 distinct elements $\{1, a, a^2, b, ab, a^2b\}$ with multiplication rules $a^3 = 1, b^2 = 1, ba = a^{-1}b = a^2b$. Let $H = \{1, b\}$, $K = \{1, ab\}$. Since $b^2 = 1$ and $(ab)^2 = abab = aa^{-1}bb = b^2 = 1$, it is easy to see that H and K are subgroups of G. However, $HK = \{1, b, ab, a^2\}$ consists of 4 distinct elements, and this cannot be a subgroup of G by Lagrange's Theorem, since 4 is not a divisor of 6.

We will now investigate some conditions under which the product HK of two subgroups will be a subgroup again.

Definition 1.27. Let H be a subgroup of G. The *normalizer* of H in G is

$$N_G(H) = \{g \in G \,|\, {}^gH = gHg^{-1} = H\}.$$

Here are some basic facts about this definition.

Lemma 1.28. *Let $H \leq G$.*

(1) *$H \trianglelefteq G$ iff $N_G(H) = G$.*

(2) *$N_G(H) \leq G$.*

(3) *$H \trianglelefteq N_G(H)$.*

(4) *$N_G(H)$ is the unique largest subgroup K of G such that $H \trianglelefteq K$.*

Proof. (1) This is by definition of normal.

(2) If $g, h \in N_G(H)$, then $ghH(gh)^{-1} = ghHh^{-1}g^{-1} = gHg^{-1} = H$, so $gh \in N_G(H)$. Multplying $gHg^{-1} = H$ on the left by g^{-1} and on the right by g gives $H = g^{-1}Hg$, so $g^{-1} \in N_G(H)$.

(3) Clearly $H \subseteq N_G(H)$. Then $H \trianglelefteq N_G(H)$ follows by the definition of normal.

(4) By (3), $N_G(H)$ is such a K. If $H \trianglelefteq K$, Then every $k \in K$ satisfies $kHk^{-1} = H$, so $k \in N_G(H)$, and thus $K \subseteq N_G(H)$. □

We can now give a useful sufficient condition under which a product of two subgroups is again a subgroup.

Proposition 1.29. *Let $H \leq G$ and $K \leq G$.*

(1) *$HK \leq G$ if and only if $HK = KH$.*

(2) *If $K \leq N_G(H)$, then $HK \leq G$.*

(3) *If $H \leq N_G(K)$, then $HK \leq G$.*

Proof. (1) Suppose that $HK \leq G$. Note that $H \subseteq HK$ and $K \subseteq HK$. Since HK is a subgroup of G containing H and K, closure under products gives $(K)(H) \subseteq HK$. Given $x \in HK$, then $x^{-1} \in HK$ since HK is a subgroup. Thus we can write $x^{-1} = hk$ with $h \in H, k \in K$. Now $x = (hk)^{-1} = k^{-1}h^{-1} \in KH$. Thus $HK \subseteq KH$. So $KH = HK$.

Conversely, suppose that $KH = HK$. Given $h_1, h_2 \in H$ and $k_1, k_2 \in K$, we have $k_1h_2 \in KH = HK$ so $k_1h_2 = h_3k_3$ some $h_3 \in H$, $k_3 \in K$. Now $(h_1k_1)(h_2k_2) = h_1(k_1h_2)k_2 = h_1(h_3k_3)k_2 = (h_1h_3)(k_3k_2) \in HK$, so HK is closed under products. Next, $(h_1k_1)^{-1} = k_1^{-1}h_1^{-1} \in KH = HK$ so HK is closed under inverses. Hence HK is a subgroup of G.

(2) For all $k \in K$ we have $kHk^{-1} = H$ or equivalently $kH = Hk$. Then $KH = \bigcup_{k\in K} kH = \bigcup_{k\in K} Hk = HK$ and so part (1) applies to show that HK is a subgroup.

(3) This is proved in the same way as (2). □

One doesn't always need the full strength of the preceding proposition; often the following result suffices.

Corollary 1.30. *Let $H \leq G$ and $K \leq G$.*

(1) *If either $H \trianglelefteq G$ or $K \trianglelefteq G$ then $HK \leq G$.*

(2) *If both $H \trianglelefteq G$ and $K \trianglelefteq G$ then $HK \trianglelefteq G$.*

Proof. (1) If $H \trianglelefteq G$ then $N_G(H) = G$ so certainly $K \subseteq N_G(H)$ and Proposition 1.29(2) applies. Similarly, if $K \trianglelefteq G$ then Proposition 1.29(3) applies.

(2) We know that $HK \leq G$ by (1). If $g \in G$ then $gHKg^{-1} = gHg^{-1}gKg^{-1} = HK$, so $HK \trianglelefteq G$. □

1.4. **Fundamental homomorphism theorems.**

Definition 1.31. If G and H are groups, a function $\phi : G \to H$ is a *homomorphism* if $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in G$. If a homomorphism ϕ is a bijection, it is called an *isomorphism*. An isomorphism $\phi : G \to G$ is called an *automorphism* of G.

Homomorphisms are the functions that relate the multiplicative structure of two groups. The word is used for the analogous maps between many other kinds of algebraic structures as well, such as rings and modules, as we will see later. An isomorphism between two groups perfectly matches up the objects of one with those of the other in such a way that the multiplication operations correspond. You should think of isomorphic groups as being essentially the same group, just that the elements have been renamed. When there exists an isomorphism $\phi : G \to H$, we say that G and H are *isomorphic* and write $G \cong H$. It is easy to check that $\phi^{-1} : H \to G$ is also an isomorphism in this case. Also, if $\phi : G \to H$ and $\psi : H \to K$ are homomorphisms of groups, then $\psi \circ \phi : G \to K$ is easily seen to be a homomorphism; if ϕ and ψ are isomorphisms, then so is $\psi \circ \phi$.

By definition a homomorphism $\phi : G \to H$ preserves the product structure of the two groups. It also automatically preserves the identity element and inverses. Namely, $\phi(1) = \phi(1 \cdot 1) = \phi(1)\phi(1)$; so multiplying on the left by $\phi(1)^{-1}$ gives $1 = \phi(1)$. Then for any $a \in G$, we have $1 = \phi(1) = \phi(aa^{-1}) = \phi(a)\phi(a^{-1})$, which implies that $\phi(a^{-1}) = (\phi(a))^{-1}$.

Some results in linear algebra or calculus can be elegantly phrased in terms of homomorphisms. For example we have the multiplicativity of the determinant.

Example 1.32. Let F be a field. Then $\phi : \mathrm{GL}_n(F) \to F^\times$ given by $\phi(A) = \det A$ is a homomorphism of groups, since $\det(AB) = \det(A)\det(B)$ for any two matrices A and B.

As another example, we have the rules for exponents:

Example 1.33. Let $\phi : (\mathbb{R}, +) \to (\mathbb{R}^\times, \cdot)$ be defined by $\phi(x) = e^x$. Then ϕ is a homomorphism, since $\phi(x + y) = e^{x+y} = e^x e^y = \phi(x)\phi(y)$.

We will be more concerned with examples internal to group theory.

Example 1.34. Let H be a subgroup of G. Then the inclusion map $i : H \to G$ is a homomorphism of groups. If $H \trianglelefteq G$ then the natural surjection $\pi : G \to G/H$ given by $\pi(g) = gH$ is a homomorphism of groups.

Example 1.35. Let $g \in G$. Let $\phi_g : G \to G$ be defined by $\phi_g(a) = gag^{-1}$. Then ϕ_g is an automorphism of the group G called a *conjugation automorphism.*

To see this, first it is easy to verify that ϕ_g is a homomorphism, since $\phi_g(ab) = gabg^{-1} = gag^{-1}gbg^{-1} = \phi_g(a)\phi_g(b)$. Then we see that ϕ_g is a bijection since $\phi_{g^{-1}}$ is the inverse function.

We now present the fundamental homomorphism theorems, which will be used frequently later. The most important one is the first one, appropriately often called the "first isomorphism theorem".

Definition 1.36. Let $\phi : G \to H$ be any homomorphism. Then $K = \ker \phi = \{a \in G | \phi(a) = 1\} = \phi^{-1}(1)$ is called the *kernel* of ϕ, and $L = \phi(G)$ is referred to as the *image* of ϕ.

It is an easy exercise to show that the image L is a subgroup of H, and the kernel K is a *normal* subgroup of G.

Theorem 1.37. *(1st isomorphism theorem) Let $\phi : G \to H$ be a homomorphism. Let $K = \ker \phi$ and $L = \phi(G)$. Then there is an isomorphism of groups $\overline{\phi} : G/K \to L$ given by $\overline{\phi}(gK) = \phi(g)$.*

Proof. We have remarked that $K = \ker \phi$ is a normal subgroup of G, so the factor group G/K makes sense. Also, L is a subgroup of H, so it is certainly a group in its own right. As usual, since we are trying to define the function $\overline{\phi}$ on a factor group by referring to the coset representative, we must check that this function is well defined. Suppose that $gK = hK$. Then $g^{-1}h \in K$, so $\phi(g^{-1}h) = \phi(g^{-1})\phi(h) = \phi(g)^{-1}\phi(h) = 1$ since $K = \ker \phi$. This implies that $\phi(g) = \phi(h)$ and so $\overline{\phi}$ is indeed well defined.

Now that we know that $\overline{\phi}$ is well-defined, the rest is routine. The function $\overline{\phi}$ is a homomorphism since $\overline{\phi}(gKhK) = \overline{\phi}(ghK) = \phi(gh) = \phi(g)\phi(h) = \overline{\phi}(gK)\overline{\phi}(hK)$. It is a surjective function because an element of L has the form $\phi(g)$ for $g \in G$, and then $\phi(g) = \overline{\phi}(gK)$. Finally, if $\overline{\phi}(gK) = \overline{\phi}(hK)$ then $\phi(g) = \phi(h)$, so $\phi(g^{-1}h) = 1$ and $g^{-1}h \in \ker \phi = K$. Then $gK = hK$, so $\overline{\phi}$ is injective. We have shown now that $\overline{\phi}$ is bijective and hence it is an isomorphism. $\square$

The 1st isomorphism theorem shows that any homomorphism leads to an isomorphism between 2 closely related groups, a factor group of the domain and a subgroup of the codomain.

Example 1.38. Consider the homomorphism $\phi : \mathrm{GL}_n(F) \to F^\times$ of Example 1.32, where $\phi(A) = \det(A)$. Then ϕ is surjective, for given a nonzero scalar λ, the diagonal matrix B_λ whose diagonal entries are $\lambda, 1, 1, \ldots, 1$ satisfies $\phi(B_\lambda) = \lambda$. Thus the first isomorphism theorem says that ϕ induces an isomorphism $\mathrm{GL}_n(F)/K \to F^\times$, where $K = \ker\phi$. Now K consists of those matrices A such that $\det(A) = 1$, since 1 is the identity element of $F^\times$. Thus K is the subgroup of $\mathrm{GL}_n(F)$ we called the special linear group $\mathrm{SL}_n(F)$. We conclude that $\mathrm{GL}_n(F)/\mathrm{SL}_n(F) \cong F^\times$.

Example 1.39. Let $\phi : (\mathbb{R}, +) \to (\mathbb{R}^\times, \cdot)$ be the homomorphism $\phi(x) = e^x$ from Example 1.33. Then from real analysis we know that the image of ϕ is all positive real numbers $\mathbb{R}_{>0}$. Thus $\mathbb{R}_{>0}$ must be a subgroup of $(\mathbb{R}^\times, \cdot)$ (which is also obvious). The kernel of ϕ is trivial, because e^x is well-known to be one-to-one. Thus the first isomorphism theorem simply tells us that restricting the codomain of ϕ we obtain an isomorphism $(\mathbb{R}, +) \to (\mathbb{R}_{>0}, \cdot)$. The inverse map is obviously the map $\psi : (\mathbb{R}_{>0}, \cdot) \to (\mathbb{R}, +)$ given by $y \mapsto \ln(y)$.

Example 1.40. Let $\phi : (\mathbb{Z}_4, +) \to (\mathbb{Z}_4, +)$ be defined by $\phi(\overline{a}) = \overline{2a}$. It is easy to check that this is a well defined homomorphism whose kernel and image are both equal to $K = \{\overline{0}, \overline{2}\}$. The first isomorphism theorem states that $\mathbb{Z}_4/K \cong K$.

Earlier, we studied a product of subgroups and gave some conditions under which it will again be a subgroup. The 2nd isomorphism theorem is an important tool for better understanding such products.

Theorem 1.41. *Suppose that $N \trianglelefteq G$ and $H \leq G$. Then $N \cap H \trianglelefteq H$ and $H/(N \cap H) \cong HN/N$.*

Proof. When one is attempting to prove that a factor group is isomorphic to another group, like here, it is often cleanest to use the 1st isomorphism theorem– it can avoid having to check directly that a function defined on cosets is well-defined (because that work was already done in the proof of the 1st isomorphism theorem).

We note first that HN is indeed a subgroup of G, because $N \trianglelefteq G$, using Corollary 1.30. Then also $N \trianglelefteq HN$ and so the factor group HN/N makes sense.

Now we define a function $\phi : H \to HN/N$ by $\phi(h) = hN$. A general element of HN/N is of the form hxN for $h \in H, x \in N$. Since $xN = N$ we have $hxN = hN = \phi(h)$. Thus ϕ is surjective. If $h \in \ker\phi$ then $\phi(h) = hN = N$ which happens if and only if $h \in N$. Thus $\ker\phi = H \cap N$.

Now by the first isomorphism theorem, ϕ induces an isomophism $\overline{\phi} : H/(N \cap H) \to HN/N$ with formula $\overline{\phi}(h(N \cap H)) = hN$. We also get that $H \cap N \trianglelefteq H$ automatically as $H \cap N$ is the kernel of a homomorphism. $\square$

Here is an example of the 2nd isomorphism theorem in an additive setting. In an additive group G we write the "product" of two subgroups H and K as $H + K = \{h + k | h \in H, k \in K\}$.

Example 1.42. Let $G = (\mathbb{Z}, +)$. For any $n \geq 1$ write $n\mathbb{Z} = \{na | a \in \mathbb{Z}\}$ for the set of all integer multiples of n. It is clearly a subgroup of G and is automatically normal since G is abelian.

Now consider the group $n\mathbb{Z} + m\mathbb{Z}$. By the theory of the greatest common divisor, the elements of the form $na + mb$ with $a, b \in \mathbb{Z}$ are exactly the multiples of $d = \gcd(m, n)$, i.e. $n\mathbb{Z} + m\mathbb{Z} = d\mathbb{Z}$. Similarly, the elements of $n\mathbb{Z} \cap m\mathbb{Z}$ are exactly the common multiples of n and m, which are the multiples of the least common multiple $\ell = \operatorname{lcm}(m, n)$. So $n\mathbb{Z} \cap m\mathbb{Z} = \ell\mathbb{Z}$.

Now the 2nd isomorphism theorem says that $(n\mathbb{Z} + m\mathbb{Z})/m\mathbb{Z} \cong n\mathbb{Z}/(n\mathbb{Z} \cap m\mathbb{Z})$. We can also write this as $d\mathbb{Z}/m\mathbb{Z} \cong n\mathbb{Z}/\ell\mathbb{Z}$.

Now one may check that $d\mathbb{Z}/m\mathbb{Z}$ is a finite group with m/d elements. So our equation says in particular that $m/d = n/\ell$, or $\ell d = mn$. This is the familiar statement that $\operatorname{lcm}(m, n) \gcd(m, n) = mn$.

Here is another example of the 2nd isomorphism theorem.

Example 1.43. Consider the general linear group $G = \mathrm{GL}_n(F)$ for a field F, and its normal subgroup the special linear group $H = \mathrm{SL}_n(F)$. Let D be the set of diagonal matrices with nonzero entries. It is easy to see that D is a subgroup of $\mathrm{GL}_n(F)$ (but it is not normal unless $n = 1$). By the second isomorphism theorem we have $DH/H \cong D/(D \cap H)$.

Note that for any $A \in \mathrm{GL}_n(F)$, where $\lambda = \det(A)$, if $B_\lambda \in D$ is the diagonal matrix whose entries are $\lambda, 1, 1 \ldots, 1$, then $A = B_\lambda((B_\lambda)^{-1}A)$ expresses A as an element of DH, since $\det((B_\lambda)^{-1}A) = \det((B_\lambda)^{-1}) \det(A) = \lambda^{-1} \det(A) = 1$. So $DH = G$ and $DH/H = G/H$. We saw earlier that this group is isomorphic to $F^\times$. So we get that $D/(D \cap H) \cong F^\times$. This is also easy to prove directly using the determinant map and the 1st isomorphism theorem.

The remaining isomorphism theorems show how we can understand a factor group—in particular, its subgroups and factor groups—in terms of the original group.

Theorem 1.44. *(Correspondence theorem) Let K be a normal subgroup of G and let $\pi : G \to G/K$ be the natural quotient map with $\pi(g) = gK$. There is a bijective correspondence*

$$\mathcal{S} = \{H \mid K \leq H \leq G\} \to \mathcal{T} = \{N \mid N \leq G/K\}$$

Given by $H \mapsto \pi(H) = H/K$. Under this bijective correspondence $H \trianglelefteq G$ if and only if $H/K \trianglelefteq G/K$.

Proof. Since $\pi(H)$ is the image of a subgroup under a homomorphism, $\pi(H) = H/K$ is a subgroup of G/K and so π does give a function $\mathcal{S} \to \mathcal{T}$. Suppose that L is a subgroup of G/K. We can define $H = \pi^{-1}(L)$, where π^{-1} means the inverse image, i.e. $\pi^{-1}(L) = \{h \in G | \pi(h) \in L\}$. One checks that H is a subgroup of G containing K. Thus π^{-1} gives a map $\mathcal{T} \to \mathcal{S}$. Because π is a surjective function, it is immediate that $\pi(\pi^{-1}(L)) = L$ for any subgroup (in fact any subset) of G/K. It is always true that $H \subseteq \pi^{-1}(\pi(H))$ for any subgroup (in fact subset) of G. But if $K \leq H$, then $\pi^{-1}(\pi(H))$ consists of elements $a \in G$ such that $\pi(a) = aK \in H/K$, or $aK = hK$ for some $h \in H$. Then $h^{-1}a \in K$ and so $a \in hK \subseteq H$. So $H = \pi^{-1}(\pi(H))$. This shows that we do have a bijection as required.

The fact that normal subgroups correspond is an easy consequence of the definitions. □

Here is the final isomorphism theorem, which shows we don't have to think about a "factor group of a factor group", because we can identify it with a factor of the original group.

Theorem 1.45. *(3rd isomorphism theorem) Let $K \trianglelefteq G$ and $G' = G/K$. Then any normal subgroup of G' has the form H/K for a unique $H \trianglelefteq G$ with $K \subseteq H$, and $(G/K)/(H/K) \cong G/H$.*

Proof. We know from the correspondence theorem that the normal subgroups of G/K are in one-one correspondence with normal subgroups H of G with $K \leq H \leq G$ under the map $\pi : G \to G/K$. Thus every normal subgroup of G/K does have the form $\pi(H) = \{hK | h \in H\} = H/K$ for a unique such H with $H \trianglelefteq G$.

Now we define a homomorphism $\phi : G/K \to G/H$ by $\phi(aK) = aH$. To show this is well-defined, note that if $aK = bK$ then $a^{-1}b \in K$. So $a^{-1}b \in H$ which means $aH = bH$. Now ϕ is obviously surjective. If $aK \in \ker \phi$ then $aH = H$ and so $a \in H$. Thus $\ker \phi = \{hK | h \in H\} = H/K$ and by the 1st isomorphism theorem, $(G/K)/(H/K) \cong G/H$ as required. □

Example 1.46. Let $G = (\mathbb{Z}, +)$. We apply the correspondence and 3rd isomorphism theorems to factor groups of G.

First let us recall the classification of subgroups of G. We have the trivial subgroup $\{0\}$ of $\mathbb{Z}$. We often abuse notation and write this subgroup as 0. Suppose that $H \leq \mathbb{Z}$ is a nontrivial subgroup. Then if $a \in H$, its additive inverse $-a \in H$ as well. So H has some positive element. Let $n = \min\{a \in H | a > 0\}$. If $a \in H$ then by the usual division with remainder in $\mathbb{Z}$, $a = qn + r$ for some $q, r \in \mathbb{Z}$ with $0 \leq r < n$. But since $n \in H$, qn (the qth multiple of n) is in H. Thus $r = a - qn \in H$. By the definition of n, this forces $r = 0$ and hence $a = qn$. Thus $H \subseteq n\mathbb{Z} = \{qn | q \in \mathbb{Z}\}$. Conversely,

since $n \in H$ we easily get that $n\mathbb{Z} \subseteq H$ since H is a subgroup. We conclude that $H = n\mathbb{Z}$ for some $n \geq 1$. It is also trivial to see that $n\mathbb{Z}$ really is a subgroup of $\mathbb{Z}$ for all $n \geq 1$.

Thus the subgroups of $\mathbb{Z}$ are 0 together with the subgroups $n\mathbb{Z}$ for all $n \geq 1$. Since $\mathbb{Z}$ is abelian, these are all normal subgroups and so the possible factor groups of $\mathbb{Z}$ are $\mathbb{Z}/0 \cong \mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$, the integers modulo n under $+$, for all $n \geq 1$.

Given a nontrivial factor group of $\mathbb{Z}$, $\mathbb{Z}/n\mathbb{Z}$ for some $n \geq 1$, then the correspondence theorem tells us the subgroups of $\mathbb{Z}/n\mathbb{Z}$ are in bijective correspondence to subgroups of $\mathbb{Z}$ which contain $n\mathbb{Z}$. These are the $d\mathbb{Z}$ such that d is a divisor of n. Thus the subgroups of $\mathbb{Z}/n\mathbb{Z}$ are the groups $d\mathbb{Z}/n\mathbb{Z}$ where d is a divisor of n. There is one for each divisor d of n.

Moreover, by the 3rd isomorphism theorem, $(\mathbb{Z}/n\mathbb{Z})/(d\mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z}/d\mathbb{Z}$. This tells us exactly what factor groups of factor groups look like up to isomorphism.

1.5. **Generators and cyclic groups.**

Definition 1.47. Let $X \subseteq G$ where G is a group, and X is any subset. The *subgroup of G generated by X* is the intersection of all subgroups of G which contain X. We write $\langle X \rangle$ for this group.

It is easy to see that an arbitrary intersection of subgroups of G is again a subgroup. Thus $\langle X \rangle$ is indeed a subgroup of G, and so it must be the uniquely minimal subgroup of G containing X, as it is contained in all others. We claim that a more explicit way of describing $\langle X \rangle$ is as $\langle X \rangle = \{x_1^{\pm 1} \dots x_k^{\pm 1} | x_i \in X\}$. In other words, this is the set of all finite products of elements in X and their inverses. It is easy to see that the set of all such products is a subgroup of G. On the other hand, any subgroup of G containing X must contain all such products. Hence $\langle X \rangle$ is indeed the set of such products as claimed.

When X is finite, say $X = \{x_1, \dots, x_n\}$, we write $\langle x_1, \dots, x_n \rangle$ for $\langle X \rangle$. In particular, when $X = \{x\}$ we just write $\langle x \rangle$.

Definition 1.48. A group G is *cyclic* if $G = \langle a \rangle$ for some $a \in G$. In this case g is called a *generator* of G. A subgroup H of G is called cyclic if it is cyclic as a group in its own right, i.e. if $H = \langle a \rangle$ for some g in G.

We will see momentarily that cyclic groups are easy to understand, as they have quite a simple structure.

We first need to review notation for powers and define the order of an element. Given $a \in G$, where G is a group, we define $a^n \in G$ for all $n \geq 1$ as the product of n copies of a, i.e. $a^n = \overbrace{aa \dots a}^{n}$. When $n = 0$, we let $a^0 = 1$, where 1 is the identity of G, by convention. We have already defined

a^{-1} to be the inverse of a. Then for any $n < 0$ we let $a^n = (a^{-1})^{|n|}$, the product of $|n|$ copies of a^{-1}. A simple case-by-case analysis shows that the usual rules for exponents hold, that is

$$a^m a^n = a^{m+n} \text{ for all } m, n \in \mathbb{Z}. \tag{1.49}$$

In an additive group, as always, we change our notation as powers are not appropriate. So if the operation in G is $+$, for $n \geq 1$ instead of a^n we write $na = \overbrace{a + a + \cdots + a}^{n}$ and call it the nth multiple of a. We have $0a = 0$ and for $n < 0$, $na = |n|(-a)$. Then (1.49) becomes $na + ma = (n + m)a$ for all $m, n \in \mathbb{Z}$.

Now consider a cyclic subgroup $\langle a \rangle$ of an arbitrary group G, where we use the multiplicative notation by default. By the explicit description of the subgroup generated by a subset we found above, $\langle a \rangle$ consists of products of finitely many copies of a or a^{-1}. Thus $\langle a \rangle = \{a^i | i \in \mathbb{Z}\}$. The structure of this group is closely related to the following notion.

Definition 1.50. Let G be a group and let $a \in G$. The *order* of a, written $|a|$ or $o(a)$, is the smallest $n > 0$, if any, such that $a^n = 1$. If no such n exists we put $|a| = \infty$.

Theorem 1.51. *Let $a \in G$ for a group G. Let $\langle a \rangle$ be the cyclic subgroup of G generated by a.*

(1) *If $|a| = \infty$ then $a^i = a^j$ if and only if $i = j$, and $\langle a \rangle \cong (\mathbb{Z}, +)$.*

(2) *if $|a| = n < \infty$ then $a^i = a^j$ if and only if $i \equiv j \mod n$, and $\langle a \rangle \cong (\mathbb{Z}_n, +)$.*

Proof. We have noted that $\langle a \rangle = \{a^i | i \in \mathbb{Z}\}$. Define $\phi : (\mathbb{Z}, +) \to \langle a \rangle$ by $\phi(i) = a^i$. The rules for exponents in (1.49) show that ϕ is a homomorphism of groups. It is clear that ϕ is surjective, so by the first isomorphism theorem we have $\mathbb{Z}/\ker\phi \cong \langle a \rangle$.

(1) Suppose $o(a) = \infty$. If $a^i = a^j$, say with $i \leq j$, we have $a^{j-i} = 1$. This contradicts that a has infinite order unless $i = j$. But this means that ϕ is injective so ϕ is an isomorphism and $\mathbb{Z} \cong \langle a \rangle$.

(2) Suppose instead that $o(a) = n < \infty$. Then $\ker\phi$ is a nonzero subgroup of $\mathbb{Z}$ whose smallest positive element is n, by the definition of order. As we saw in Example 1.46, this means that $\ker\phi = n\mathbb{Z}$ and so $\mathbb{Z}/n\mathbb{Z} \cong \langle a \rangle$ by the 1st isomorphism theorem. We can identify $\mathbb{Z}/n\mathbb{Z}$ with the group $\mathbb{Z}_n$ of integers mod n, as we saw in Example 1.24. Now $a^i = a^j$ if and only if $a^j a^{-i} = a^{j-i} = 1$, if and only if $j - i \in \ker\phi = n\mathbb{Z}$, or equivalently $i \equiv j \mod n$. □

Corollary 1.52. *Let G be a finite group. If $a \in G$, then the order $|a|$ divides $|G|$.*

Proof. Since G is finite, $|a|$ is finite (else the powers of a are all distinct, which is impossible). We have $|\langle a \rangle| = |\mathbb{Z}_n| = n = |a|$ by the theorem. By Lagrange's Theorem, the order of the subgroup $\langle a \rangle$ must divide $|G|$. □

All results about the properties of cyclic groups can be proved just for the specific additive groups $\mathbb{Z}$ and $\mathbb{Z}_n$ if we wish, and then transferred to general cyclic groups via the isomorphisms in Theorem 1.51. For example, we have the following classification of subgroups of a cyclic group.

Proposition 1.53. *Let $G = \langle a \rangle$ be a cyclic group.*

(1) *If $|a| = \infty$, then every nonidentity element of G has infinite order. The subgroups of G are $\{1\}$ and the subgroups $\langle a^n \rangle = \{a^{in} | i \in \mathbb{Z}\}$ for each $n \geq 1$, and they are all cyclic.*

(2) *If $|a| = n < \infty$ then $|G| = n$ and the subgroups of G are $\langle a^{n/d} \rangle$ for each divisor d of n, where $|\langle a^{n/d} \rangle| = d$. In particular there is a unique subgroup of G of order d for each divisor d of n, and these subgroups are also cyclic.*

Proof. (1) We know that $\phi : (\mathbb{Z}, +) \to \langle a \rangle$ given by $\phi(i) = a^i$ is an isomorphism. We have shown that the subgroups of $(\mathbb{Z}, +)$ are 0 and the subgroups $n\mathbb{Z} = \langle n \rangle$ for each $n \geq 1$, as discussed in Example 1.46. It is obvious that all nonzero elements of $\mathbb{Z}$ have infinite additive order. Now statement (1) follows from transferring all of this information to $\langle a \rangle$ via ϕ.

(2) Similarly as in (1), we have an isomorphism $\phi : (\mathbb{Z}/n\mathbb{Z}, +) \to \langle a \rangle$ given by $\phi(\bar{i}) = a^i$. Now we have seen using the correspondence theorem, in Example 1.46, that the subgroups of $\mathbb{Z}/n\mathbb{Z}$ are exactly the groups $d\mathbb{Z}/n\mathbb{Z}$ for divisors d of n. Note that $d\mathbb{Z}/n\mathbb{Z}$ is the cyclic subgroup $\langle d + n\mathbb{Z} \rangle$ of $\mathbb{Z}/n\mathbb{Z}$. Transferring this information to $\langle a \rangle$, we get that the subgroups of $\langle a \rangle$ are those of the form $\langle a^d \rangle$ for divisors d of n, and there is exactly one of these for each divisor d. Since $|a| = n$, it is straightforward to see that $|a^d| = n/d$. Finally, as d runs over divisors of n, so does n/d, and replacing d by n/d gives statement (2). □

1.6. **Automorphisms.** One way that groups arise very naturally is as sets of symmetries of objects under composition. What one means by a symmetry depends on the setting but usually it is a bijection that preserves the essential features. For example, the dihedral group D_{2n} is the group of symmetries of a regular n-gon; here a symmetry is an orthogonal (distance preserving) bijective map of the plane that maps the n-gon back onto itself.

An automorphism of a group is a kind of self-symmetry that preserves the essential feature of a group—its product. Correspondingly, the set of automorphisms of a group will themselves form a group of symmetries.

Definition 1.54. Let G be a group. The set $\mathrm{Aut}(G)$ of all automorphisms of G is called the *automorphism group* of G. It is itself a group under composition.

It is very easy to check that the composition of two automorphisms is also an automorphism, and that the inverse function of an automorphism is agan an automorphism. Thus $\mathrm{Aut}(G)$ really is a group.

We already remarked earlier that for any $g \in G$, there is an automorphism $\theta_g : G \to G$ given by $\theta_g(x) = gxg^{-1}$. In other words, θ_g is "conjugation by g". Note that $\theta_g \circ \theta_h = \theta_{gh}$ and $(\theta_g)^{-1} = \theta_{g^{-1}}$. Thus $\mathrm{Inn}(G) = \{\theta_g | g \in G\}$ is a subgroup of $\mathrm{Aut}(G)$. The elements of $\mathrm{Inn}(G)$ are called *inner* automorphisms. They are in some sense the most obvious automorphisms of a group, the ones that are derived in a natural way from the multiplication in the group itself.

This is a good time as any to introduce the center of a group and centralizers of elements, since the center appears in the next theorem.

Definition 1.55. If $g \in G$, then the *centralizer* of g is $C_G(g) = \{x \in G | gx = xg\}$. The *center* of the group G is $Z(G) = \{x \in G | gx = xg \text{ for all } g \in G\}$.

In other words, the centralizer is the set of all elements which commute with the element g. A quick argument shows that $C_G(g)$ is a subgroup of G. Since the powers of g all commute with each other by (1.49), we always have $\langle g \rangle \subseteq C_G(g)$. The center is the set of all elements which commute with all other elements. One also easily check directly that $Z(G)$ is a subgroup of G. Alternatively, one notes that $Z(G) = \bigcap_{g \in g} C_G(g)$, and thus $Z(G)$ is a subgroup since it is an intersection of subgroups. In fact $Z(G) \trianglelefteq G$, since $gxg^{-1} = x$ for all $x \in Z(G)$ and all $g \in G$. The group G is abelian if and only if $G = Z(G)$.

Note that if G is abelian, then $\theta_g = 1$ for all $g \in G$ and so $\mathrm{Inn}(G) = \{1\}$ is trivial. More generally, we can relate $\mathrm{Inn}(G)$ to the center of G as follows:

Lemma 1.56. *Let G be a group. Then there is an isomorphism $\phi : G/Z(G) \to \mathrm{Inn}(G)$ given by $\phi(gZ(G)) = \theta_g$.*

Proof. Define $\psi : G \to \mathrm{Inn}(G)$ by $\psi(g) = \theta_G$. Then ψ is a homomorphism by the fact that $\theta_g \circ \theta_h = \theta_{gh}$, as we have already remarked. The map ψ is surjective by the definition of $\mathrm{Inn}(G)$. The kernel of ψ consists of those g such that $\theta_g = 1$. But $\theta_g(x) = gxg^{-1} = x$ holds for all x if and only if $g \in Z(G)$. Hence $Z(G) = \ker \psi$ and so there is an isomorphism $\overline{\psi} = \phi : G/Z(G) \to \mathrm{Inn}(G)$ with the desired formula, by the 1st isomorphism theorem. □

Thus if we understand the group G well (in particular if we know its center) there is not much mystery about $\mathrm{Inn}(G)$.

Lemma 1.57. *Let G be a group. Then $\mathrm{Inn}(G) \trianglelefteq \mathrm{Aut}(G)$.*

Proof. We have already remarked that $\mathrm{Inn}(G) \leq \mathrm{Aut}(G)$, so we just need to prove normality. Let $\theta_g \in \mathrm{Inn}(G)$ and let $\rho \in \mathrm{Aut}(G)$. Consider $\rho \circ \theta_g \circ \rho^{-1}$. Applying this to some x we have

$$\rho\theta_g\rho^{-1}(x) = \rho(g\rho^{-1}(x)g^{-1}) = \rho(g)x\rho(g^{-1}) = \rho(g)x\rho(g)^{-1} = \theta_{\rho(g)}(x).$$

Hence $\rho\theta_g\rho^{-1} = \theta_{\rho(g)} \in \mathrm{Inn}(G)$ and so $\mathrm{Inn}(G)$ is normal in $\mathrm{Aut}(G)$. □

Because of the lemma, it makes sense to define the factor group $\mathrm{Out}(G) = \mathrm{Aut}(G)/\mathrm{Inn}(G)$, which is called the *outer automorphism group*. It is the part of the automorphism group that tends to be harder to understand. We will give some examples of calculating automorphism groups in the next section.

Suppose that $K \leq H \leq G$ where $K \trianglelefteq H$ and $H \trianglelefteq G$. It is natural to hope that being a normal subgroup should be "transitive" in the sense that $K \trianglelefteq G$ in this situation, but this does not follow in general.

Example 1.58. Let $G = D_8$ be the dihedral group, where we write $G = \{a^ib^j | 0 \leq i \leq 3, 0 \leq j \leq 1\}$, with $a^4 = 1, b^2 = 1$, and $ba = a^{-1}b$. Then $H = \{1, a^2, b, a^2b\}$ is a subgroup of G, as is easy to check by direct calculation. Since $|G : H| = 2$, $H \trianglelefteq G$. Let $K = \{1, b\}$, which is a subgroup of G since $b^2 = 1$. The index $|H : K| = 2$ as well, so $K \trianglelefteq H$. However K is not normal in G, since $aba^{-1} = a^2b \notin K$.

Fortunately, in the next proposition we will see a useful situation where we are able to conclude that at a normal subgroup of a normal subgroup is normal, by strengthening the hypothesis of normality. Note that $H \trianglelefteq G$ is equivalent to $gHg^{-1} = H$ for all $g \in G$, or alternatively $\theta_g(H) = H$ for all inner automorphisms θ_g. So it is also interesting to consider those subgroups that are fixed by all automorphisms, not just inner ones.

Definition 1.59. A subgroup $H \leq G$ is *characteristic* if for all automorphisms $\sigma \in \mathrm{Aut}(G)$, $\sigma(H) = H$. We write $H \,\mathrm{char}\, G$ in this case.

Clearly from the remarks above, characteristic subgroups are normal.

Proposition 1.60. *Let $K \leq H \leq G$.*

(1) *If $K \,\mathrm{char}\, H$ and $H \trianglelefteq G$, then $K \trianglelefteq G$.*

(2) *If $K \,\mathrm{char}\, H$ and $H \,\mathrm{char}\, G$, then $K \,\mathrm{char}\, G$.*

Proof. (1) Suppose that $g \in G$. Since $H \trianglelefteq G$, we know that $\theta_g(H) = gHg^{-1} = H$. Thus the restriction $\rho = \theta_g|_H : H \to H$ is an automorphism of H, because it has the inverse $\theta_{g^{-1}}|_H$. Since $K \,\mathrm{char}\, H$, we have $\rho(K) = K$. But this says that $gKg^{-1} = K$. Thus $K \trianglelefteq G$.

(2) This is similar to (1) except that we start with an arbitrary automorphism of G instead of an inner automorphism θ_g. □

Example 1.61. Suppose that $H \trianglelefteq G$ where H is cyclic of finite order n. If K is any subgroup of H, say of order d, then we have seen that K is the unique subgroup of H of order d. If $\sigma \in \mathrm{Aut}(H)$, then $\sigma(K)$ is a subgroup of H of order d as well, so $\sigma(K) = K$. Thus $K \operatorname{char} H$. It follows from proposition 1.60 that $K \trianglelefteq G$.

For example, in $G = D_{2n}$ the rotation subgroup H is cyclic of order n and $H \trianglelefteq G$ since $|G : H| = 2$. Then if K is any subgroup of H, $K \trianglelefteq G$.

1.7. **Direct products.** We will study direct products in more detail in a later section, but since direct products are very useful for building basic examples, it is good to have them at hand early on.

The direct product is a natural way of joining together two groups which apriori have no relationship to each other.

Definition 1.62. Let H and K be groups. We define the *direct product* of H and K to be $H \times K = \{(h, k) | h \in H, k \in K\}$, that is, the cartesian product of the sets H and K. The group operation in $H \times K$ is done coordinatewise, so $(h_1, k_1)(h_2, k_2) = (h_1 h_2, k_1 k_2)$ using the product of H in the first coordinate and the product of K in the second coordinate.

The group axioms for $H \times K$ follow immediately from the axioms for H and K. In particular, note that the identity element of $H \times K$ is $(1_H, 1_K)$ and that $(h, k)^{-1} = (h^{-1}, k^{-1})$.

If we understand the groups H and K well, it is usually quite easy to understand the properties of the group $H \times K$. For example, clearly $|G| = |H||K|$. If $g = (h, k) \in H \otimes K$, then $g^n = (h^n, k^n)$. This is equal to $(1, 1)$ if and only if $h^n = 1$ and $k^n = 1$. So if $|h| = \infty$ or $|k| = \infty$ then $|(h, k)| = \infty$. If h and k have finite order then $g^n = 1$ if and only if n is a multiple of $|h|$ and a multiple of $|k|$, and thus $|(h, k)| = \mathrm{lcm}(|h|, |k|)$.

There is no reason to restrict the definition to 2 groups above. We can define the product of a finite number of groups $G_1, G_2, \ldots, G_k$ in an analogous way, as the set of all k-tuples $(g_1, g_2, \ldots, g_k)$ with $g_i \in G_i$, with coordinatewise operations.

2. Free groups and presentations

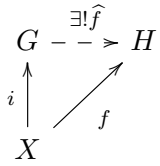
2.1. **Existence and uniqueness of the free group on a set.** We have informally described the dihedral group D_{2n} as a group with elements $\{a^i b^j | 0 \leq i \leq n-1, 0 \leq j \leq 1\}$ where $a^n = 1, b^2 = 1$

and $ba = a^{-1}b$. This is appropriate because we first defined it as a subgroup of the orthogonal group with $2n$ elements, and then showed it its elements can be described in terms of a rotation a and a reflection b as the $2n$ elements in the above set with the listed multiplication rules. Sometimes, however, we would like to define a group just by listing a set of elements (or even just a set of generators) and the rules that they should satisfy. One needs to be careful that there really is a group with the desired number of elements that satisfies those rules. The formalism of presentations, which we will describe in this section, allows one to make this precise.

We will first need to spend some time defining free groups. These are interesting groups we have not encountered yet that satisfy a certain universal property.

Definition 2.1. Let G be a group. We say that G is *free* on a subset $X \subseteq G$ if given a group H together with a function $f : X \to H$, there is a unique homomorphism $\widehat{f} : G \to H$ such that $\widehat{f}(x) = f(x)$ for all $x \in X$.

The universal property of a free group can be indicated by the following commutative diagram:



Here $i : X \to G$ is just the inclusion map of X into G, i.e. $i(x) = x$.

Commutative diagrams are convenient ways of visualizing properties that assert that certain compositions of functions are equal. The convention is that by saying the diagram is commutative or that it commutes, one means that all different paths that follow arrows from one object to another give equal compositions of functions. In the diagram above, that means that $\widehat{f} \circ i = f$ as functions $X \to H$, which is clearly the same as $\widehat{f}(x) = f(x)$ for all $x \in X$, the property stated in the definition of a free group. We have illustrated some other common conventions in the diagram above. Since the maps i and f are part of the given data, they are regular arrows, while the map $\widehat{f}$ is a dashed arrow because it is a map that is not given but whose existence is asserted by the property being illustrated. The exclamation point ! stands for "unique", so the notation $\exists!$ is read "there exists a unique" since the uniqueness of the function $\widehat{f}$ completing the diagram is part of the universal property.

The uniqueness is what makes a universal property so useful. It means in this case that we can define a homomorphism from a free group G on a set X to another group H simply by choosing any function $f : X \to H$. In other words, the elements in X are "free" to be sent anywhere we

please. There is then a unique extension of this function to a homomorphism of groups $\widehat{f} : G \to H$ which does the given map f on the subset X.

It is not at all obvious that any groups with such a property exist, but we will show that any set X can be embedded in a free group on that set. The case where X has one element is especially easy, as we have already seen that group before.

Example 2.2. Let G be an infinite cyclic group with generator $x \in G$. So $G = \langle x \rangle = \{x^i | i \in \mathbb{Z}\}$ where $x^i = x^j$ if and only if $i = j$. Then we claim that G is free on the one-element subset $X = \{x\}$. To prove this we check the definition directly. Let H be any other group and let $f : X \to H$ be a function. Since X has one element, such a function amounts to a choice of a single element $h \in H$ for which $f(x) = h$. Now we define $\widetilde{f} : G \to H$ by $\widehat{f}(x^i) = h^i$ for all $i \in \mathbb{Z}$. It is immediate that $\widehat{f}$ is a homomorphism by our rules for exponents in groups (1.49). Clearly also $\widehat{f}(x) = h = f(x)$ by construction. Finally, if $\phi : G \to H$ is any homorphism of groups for which $\phi(x) = f(x) = h$, then $\phi(x^i) = h^i$ for all i by the properties of homomorphisms, and so $\phi = \widehat{f}$. This shows the uniqueness of $\widetilde{f}$ and completes the claim that G is free on $\{x\}$.

Thus we have constructed a free group on a one-element set. Could there be an essentially different group which is also free on a one-element subset? The answer is no. In fact, free groups are determined up to isomorphism by the size of the set X. This is actually a general principle for objects in algebra that are called "free"— the object is uniquely determined up to isomorphism by the size of the subset it is free on.

Theorem 2.3. *Let G be a free group on a subset X and let G' be a free group on a subset X'. Suppose there is a bijection of sets $f : X \to X'$. Then there is a unique isomorphism of groups $\phi : G \to G'$ such that $\phi(x) = f(x)$ for all $x \in X$.*

Proof. Note that $f : X \to X'$ can be considered as a function $f : X \to G'$. Then by the universal property of G being free on X, there is a unique homomorphism $\phi : G \to G'$ such that $\phi(x) = f(x)$ for all $x \in X$. Once we prove that ϕ is an isomorphism of groups, we see from this that it will be unique.

Since f is a bijection, the inverse function $f^{-1} : X' \to X$ makes sense. Then similarly, using the universal property of G' on X', there is a unique homomorphism $\psi : G' \to G$ such that $\psi(x') = f^{-1}(x')$ for all $x' \in X'$.

Now $\psi \circ \phi : G \to G$ is a homomorphism, being a composition of two homomorphisms. By construction, we have $\psi \circ \phi(x) = \psi(f(x)) = f^{-1}(f(x)) = x$ for all $x \in X$. But the identity map

$1_G : G \to G$ is also a homomorphism $G \to G$ such that $1_G(x) = x$ for all $x \in X$. Since both 1_G and $\psi \circ \phi$ restrict on X to the inclusion function $i : X \to G$, by the uniqueness part of the universal property we must have $\psi \circ \phi = 1_G$. A symmetric argument using the universal property of G' gives $\phi \circ \phi = 1_{G'}$. We conclude that $\phi : G \to G'$ is an isomorphism of groups with inverse $\psi : G' \to G$. $\square$

Recall that two sets X, X' have the *same cardinality* if there is a bijection $f : X \to X'$. Notationally this is indicated by $|X| = |X'|$. The theorem shows that there is only one free group on a set of a given cardinality, up to isomorphism. So we can speak of "the" free group on n generators for a given finite number n, for example.

We now settle the trickier issue of showing that free groups exist, by giving a direct construction.

Definition 2.4. Let X be a set. We create an *alphabet* A of formal symbols consisting of the elements in X along with a new symbol x^{-1} for each $x \in X$. For example, if $X = \{x, y, z\}$ then the alphabet is $A = \{x, y, z, x^{-1}, y^{-1}, z^{-1}\}$. A *word in* X is a finite sequence of symbols in the alphabet A, written consecutively without spaces (like actual dictionary words). By convention we also have an "empty" word which we write as 1. The *length* of a word is the number of symbols it contains, where the empty word 1 has length 0.

Example 2.5. Let $X = \{x, y, z\}$. Then $w = xx^{-1}xyzyy^{-1}x$ is a word in X of length 8. For each $n \geq 0$, there are precisely 6^n distinct words of length n in X, since there are six symbols in the associated alphabet A to choose from for each of n spots.

Definition 2.6. Given a word in X, a *subword* is a some subsequence of consecutive symbols within the word. A word w in X is *reduced* if it contains no subwords of the form xx^{-1} or $x^{-1}x$ for $x \in X$.

For example, in the word $w = xx^{-1}xyzyy^{-1}x$ given above, $x^{-1}xyzy$ and $yy^{-1}x$ are subwords. This word is not reduced, for it contains $xx^{-1}, x^{-1}x$ and yy^{-1} as subwords. On the other hand, $xyx^{-1}zx^{-1}yxy^{-1}x$ is a reduced word.

Given a word w which is not reduced, say of length n, a *reduction* is the removal of some subword of w of the form xx^{-1} or $x^{-1}x$, squeezing the remaining symbols together to obtain a new word of length $n - 2$. If that word is also not reduced, we can perform some other reduction on it, and continue in this way. Obviously this process must stop at some point, leaving us with a reduced word we call the *reduction* of w, notated $\text{red}(w)$ (which could be the empty word 1).

Example 2.7. If $w = yxyy^{-1}x^{-1}x$, we can first remove yy^{-1} leaving $yxx^{-1}x$. Now we can remove xx^{-1}, leaving the reduced word yx. We could instead have started by removing the $x^{-1}x$ at the tail end of w, leaving $yxyy^{-1}$, and then removing yy^{-1} to obtain yx.

Proposition 2.8. *Given a word w on a set X, any possible sequence of reductions leads to the same reduced word* $\mathrm{red}(w)$ *(and thus* $\mathrm{red}(w)$ *is well-defined).*

This proposition seems intuitively reasonable, but it certainly needs proof. We leave it to the reader as an exercise so as not to interrupt the flow of the discussion here.

Definition 2.9. Given a set X, we define $F(X)$ as follows. As a set, $F(X)$ consists of all reduced words in X, that is words from the associated alphabet A, which do not contain any subwords of the form $x_i x_i^{-1}$ or $x_i^{-1}x_i$. The product in $F(X)$ is defined as $v * w = \mathrm{red}(vw)$ for $v, w \in F(X)$, where vw means the concatenation of the two words. (Note that although v and w are reduced, vw may not be, which requires passing to the reduction $\mathrm{red}(vw)$ to obtain another element of the set $F(X)$. We are also relying on Proposition 2.8 here to be sure that $\mathrm{red}(vw)$ is a well-defined element of $F(X)$.)

Example 2.10. If $X = \{x, y\}$, then in $F(X)$ we have $(xyx) * (x^{-1}y^{-1}x) = \mathrm{red}(xyxx^{-1}y^{-1}x) = xx$.

Theorem 2.11. *Let X be a set and let $F(X)$ be the set defined above. Identify X with the subset of $F(X)$ consisting of length 1 words on the symbols in X.*

(1) *$F(X)$ is a group under the operation $*$.*

(2) *$F(X)$ is free on the subset X.*

Proof. (1) It is not immediately obvious in this case that $*$ is associative. Note that if $u, v, w \in F(X)$ are reduced words, then $(u * v) * w = \mathrm{red}(\mathrm{red}(uv)w)$, while $u * (v * w) = \mathrm{red}(u\,\mathrm{red}(vw))$. Both of these expressions are obtained by applying some sequence of reductions to uvw. Thus they are equal to $\mathrm{red}(uvw)$ by the uniqueness of the reduced word obtained through applying reductions, as stated in Proposition 2.8. So $*$ is indeed associative. The trivial word 1 is clearly an identity element for $F(X)$, since $1 * w = \mathrm{red}(1w) = \mathrm{red}(w) = w$ and similarly $w * 1 = w$, for any $w \in F(X)$. Finally, if $w = x_1^{e_1} \dots x_n^{e_n}$ is some reduced word, where each $x_i \in X$, and $e_i = \pm 1$, then it is easy to check that $x_n^{-e_n} \dots x_1^{-e_1}$ is also a reduced word and gives an inverse for w under $*$.

(2) If H is any group and $f : X \to H$ is some function, we define $\widehat{f} : F(X) \to H$ by $\widehat{f}(x_1^{e_1} \dots x_n^{e_n}) = f(x_1)^{e_1} \dots f(x_n)^{e_n}$, for any reduced word $x_1^{e_1} \dots x_n^{e_n} \in F(X)$, where $e_1 = \pm 1$ and $x_i \in X$. Suppose that $v, w \in F(X)$ and that $v * w = vw$, in other words the concatenation of v and w

is already reduced. In this case from the definition of $\widehat{f}$ we easily get $\widehat{f}(v*w) = \widehat{f}(vw) = \widehat{f}(v)\widehat{f}(w)$. In the general case, when calculating $v * w = \text{red}(vw)$, note that all of the reductions happen along the "join" between the two words. In other words, there is a word u such that $v = v'u$ and $w = u^{-1}w'$, and $v * w = \text{red}(vw) = v'w'$. Since the products $v'w'$, $v'u$ amd $u^{-1}w'$ are already reduced, we obtain

$$\widehat{f}(v*w) = \widehat{f}(v'w') = \widehat{f}(v')\widehat{f}(w') = \widehat{f}(v')\widehat{f}(u)\widehat{f}(u^{-1})\widehat{f}(w') = \widehat{f}(v'u)\widehat{f}(u^{-1}w') = \widehat{f}(v)\widehat{f}(w).$$

(Here, the product $\widehat{f}(u)\widehat{f}(u^{-1})$ has the form $f(x_1)^{e_1} \dots f(x_n)^{e_n} f(x_n)^{-e_n} \dots f(x_1)^{-e_1}$, which is trivial in H). Thus $\widehat{f}$ is a homomorphism. This homomorphism certainly satisfies $\widehat{f}(x) = f(x)$ for $x \in X$. Finally, any element of $F(X)$ is equal to a product in $F(X)$ of elements of X and their inverses. It is clear from this that any homomorphism is determined by its action on the elements of X, so that $\widehat{f}$ is the unique homomorphism extending f. □

Note that in a free group $F(X)$, for a given $x \in X$ the word $\overbrace{xx\dots x}^{n}$ is equal to the product of n copies of x in $F(X)$. So we can write this as x^n from now on. Similarly, we write $\overbrace{x^{-1}x^{-1}\dots x^{-1}}^{n}$ as x^{-n}. By abuse of notation we will also call expressions involving powers of the elements in X and their inverses words. For example we can refer to $x^2yx^{-2}y$ as a word in $\{x, y\}$, with the understanding that this stands for the word $xxyx^{-1}x^{-1}y$.

We have seen that a free group on a set with one element is just an infinite cyclic group. To close this section we remark that free groups on sets X with at least two elements, on the other hand, are very large and have some counterintuitive properties.

Example 2.12. The free group $G = F(X)$ on a set $X = \{x, y\}$ with two elements contains a subgroup H which isomorphic to a free group on a countably infinite set. We claim that one such example is $H = \langle y, xyx^{-1}, x^2yx^{-2}, \dots \rangle$. If $Z = \{z_0, z_1, z_2, \dots, \}$ is a countably infinite set, note that by the universal property we certainly get a unique homomorphism $\phi : F(Z) \to H$ with $\phi(z_i) = x^iyx^{-i}$ for all i. Because the image of ϕ contains a set of generators for H, $\phi(F(Z)) = H$. One can show furthermore that ϕ is injective (we leave this as an exercise), so that $F(Z) \cong H$ as claimed. Moreover, this means that G also contains subgroups isomorphic to free groups on any finite number of generators, for $H_n = \langle y, xyx^{-1}, \dots x^{n-1}yx^{-n+1} \rangle$ will be isomorphic to a free group on n elements.

It is at least true that if $F(X) \cong F(Y)$ for some sets X and Y, then $|X| = |Y|$. This can be seen by noting that the set of groups H such that there is a surjective homomorphism $\phi : F(X) \to H$ is the same as the set of groups that can be generated by a subset of at most $|X|$ elements. But

for each X one can exhibit a group that is generated by $|X|$ elements but cannot be generated by a set of smaller cardinality.

A group is called *free* if it is isomorphic to $F(X)$ for some set X. There is also the following interesting theorem, which we will not prove in this course:

Theorem 2.13. *(Nielsen-Schreier) Every subgroup of a free group is also free.*

2.2. **Presentations.** Suppose that H is any group, and that $H = \langle X \rangle$ for some subset X, i.e. that H is generated as a group by the subset X. We can use that same X to define a free group $F(X)$ which is free on the set X. Then by the universal property of the free group, there is a unique homomorphism $\phi : F(X) \to H$ with $\phi(x) = x$ for all $x \in X$. Since the elements in H are expressions of the form $x_1^{e_1} \dots x_n^{e_n}$ with $x_i \in X$ and $e_i = \pm 1$, it is clear that all of these elements are in the image of ϕ, so ϕ is surjective. By the first isomorphism theorem, $H \cong F(X)/N$ for some $N \trianglelefteq F(X)$. We have thus shown that *every group is isomorphic to a factor group of some free group*. We will now how such a description is especially useful when we can also give an explicit generating set for the normal subgroup N.

The comments above also give another way of thinking about the "freeness" of the free group. Note that because the elements of $F(X)$, namely reduced words in X, are products in $F(X)$ of the length one words x and x^{-1} with $x \in X$, the free group on X is also generated by its subset X. Since any other group generated by X is isomorphic to $F(X)/N$, we can think of $F(X)$ as the *most general* group which is generated by a set X.

We are now ready to define presentations.

Definition 2.14. Let $F(X)$ be a free group on a set X and let $W \subseteq F(X)$ be some set of elements in $F(X)$ (that is, some set of reduced words in X). Let N be the intersection of all normal subgroups of $F(X)$ which contain W. The notation $\langle X|W \rangle$ is called a *presentation* and by definition it is equal to the group $F(X)/N$. We call the elements in X *generators* and the elements in W *relations*.

By definition N above is the intersection of all normal subgroups of $F(X)$ containing W. It can also be described as the unique smallest normal subgroup of $F(X)$ containing W, because an intersection of normal subgroups is again normal. There is an explicit description of the elements of N in terms of the generators in W, but it is awkward, and not needed in order to work with the presentation.

It is often useful to find a presentation which is isomorphic to a given known group. Let us do this carefully now for D_{2n}.

Example 2.15. Consider the dihedral group $D_{2n} = \{1, a, a^2, \ldots, a^{n-1}, b, ab, a^2b, \ldots, a^{n-1}b\}$. From the original construction of D_{2n} as a set of transformations of the plane, we know that the $2n$ listed elements are distinct and that a and b satisfy the relations $a^n = 1$, $b^2 = 1$, and $ba = a^{-1}b$. Note that the last relation can also be written as $b^{-1}aba = 1$, by multiplying on the left by $b^{-1}a$.

Consider the presentation $G = \langle x, y | x^n, y^2, y^{-1}xyx\}$. We claim that this presented group is isomorphic to D_{2n}.

Step 1. By the universal property of the free group, there is a unique homomorphism $\phi : F(x, y) \to D_{2n}$ such that $\phi(x) = a$ and $\phi(y) = b$.

Step 2. One checks that $\phi(w) = 1$ for all words $w \in W$. This is immediate in this case because these correspond to relations among the generators $a, b \in D_{2n}$ we already know. Namely $\phi(x^n) = a^n = 1$, $\phi(y^2) = b^2 = 1$, and $\phi(y^{-1}xyx) = b^{-1}aba = 1$.

Step 3. By definition $G = F(x, y)/N$, where N is the smallest normal subgroup of $F(x, y)$ containing the set of relations $W = \{x^n, y^2, y^{-1}xyx\}$. Since $\ker \phi$ is a normal subgroup of $F(X)$ and by the previous step $W \subseteq \ker \phi$, we obtain $N \subseteq \ker \phi$. This implies that ϕ factors through $F(x, y)/N$, that is there is an induced homomorphism $\overline{\phi} : F(x, y)/N \to D_{2n}$ such that $\overline{\phi}(vN) = \phi(v)$ for all $v \in F(x, y)$.

Step 4. Note that $\{a, b\}$ generates D_{2n} and since the image of $\overline{\phi}$ is a subgroup, this forces $\overline{\phi}(G) = D_{2n}$. So $\overline{\phi}$ is surjective.

Step 5. We claim that $|G| \leq 2n$. This is the only step that can be tricky and where the details vary from example to example. The idea is to use the relations to show that an arbitrary reduced word in x, y must be equal mod N one of a few special words.

Let us write the coset $vN \in F(x, y)/N$ as $\overline{v}$. We know that $\overline{y^{-1}xyx} = 1$, or equivalently $\overline{yx} = \overline{x^{-1}}\,\overline{y}$. This equation also implies $\overline{y}\overline{x^{-1}} = \overline{xy}$. Similarly, we also have $\overline{y^{-1}}\overline{x^e} = \overline{x^{-e}}\overline{y^{-1}}$ for $e = \pm 1$. Using these relations, we can move each $\overline{y}$ or $\overline{y^{-1}}$ that occurs in $\overline{v}$ to the right of the x and x^{-1} terms, flipping the exponents of x, until finally we obtain $\overline{v} = \overline{x^i y^j}$ for some $i, j \in \mathbb{Z}$. But since $\overline{x}^n = 1$ (as $x^n \in N$), and similarly $\overline{y}^2 = 1$, we can actually get $\overline{v} = \overline{x^i y^j}$ with $0 \leq i \leq n-1$ and $0 \leq j \leq 1$. This shows that every element of G/N is equal to one of at most $2n$ cosets, so $|G| \leq 2n$. (This argument does not show that all of the elements $\overline{x^i y^j}$ with $0 \leq i \leq n-1$ and $0 \leq j \leq 1$ are actually distinct in G, so apriori we just have an inequality as claimed).

Step 6. Since $\overline{\phi} : G \to D_{2n}$ is a surjective homomorphism from a group G with $|G| \leq 2n$ onto a group with $2n$ elements, this forces $|G| = 2n$ and $\overline{\phi}$ is injective, hence an isomorphism.

Steps 1-3 of the example above are routine and so we don't need to be so explicit about them in every example. They can summed up by a universal property for a presentation which generalizes

the universal property of the free group itself. If w is a word in X, H is a group, and $f : X \to H$ is some function, we write $\mathrm{eval}_f(w)$ for the element of H obtained by substituting $f(x_i) \in H$ for x_i everywhere in the word w, and think of this as "evaluating" the word at the given elements of H. In other words, when w is reduced, $\mathrm{eval}_f(w)$ is just $\widehat{f}(w)$ where $\widehat{f} : F(X) \to H$ is the unique homomorphism of groups extending f, we see saw by the proof of the universal property of $F(X)$ in Theorem 2.11(2).

Theorem 2.16. *Let $\langle X|W\rangle$ be a presented group and let H be another group. Given a function $f : X \to H$ which has the property that $\mathrm{eval}_f(w) = 1$ for all $w \in W$, there is a unique homomorphism of groups $\psi : \langle X|W\rangle \to H$ with the property that $\psi(x) = f(x)$ for all $x \in X$.*

The proof of the theorem is similar to what was done in steps 1-3 of the preceding example and so we leave it to the reader. The upshot is that defining homomorphisms from presentations is easy: we can send the generators anyplace we like as long as the relations evaluate to 1; and then there is a unique homomorphism from the presentation that does that.

Remark 2.17. Some other notations for the relations in a presentation are in common use. Rather than writing $\langle x_1, \ldots, x_n|w_1, \ldots, w_m\rangle$, one might write $\langle x_1 \ldots, x_n|w_1 = 1, \ldots, w_m = 1\rangle$ to emphasize that the relations become equal to 1 in the presented group. Also more general than a relation of the form $w = 1$, it is common to allow relations of the form $w_1 = w_2$ which set two words equal. Such a relation should be interpreted to mean $w_2^{-1}w_1 = 1$.

For example, the presentation for D_{2n} is often written as $\langle x, y|x^n = 1, y^2 = 1, yx = x^{-1}y\rangle$.

Example 2.18. Here is an example where we start with a presentation to show that it is hard to predict from a glance at the relations what kind of group it is, for example what its order is.

Let $G = \langle x, y|xyx, yxy\rangle$. By definition this is $F(x, y)/N$ where N is the smallest normal subgroup of $F(x, y)$ containing xyx and yxy. Write $vN = \overline{v} \in F(x, y)N$ for $v \in F(x, y)$, as in the earlier example. Now notice that $\overline{xyxy} = \overline{x}$ since $\overline{yxy} = 1$ but also $\overline{xyxy} = \overline{y}$ since $\overline{xyx} = 1$. Thus $\overline{x} = \overline{y}$ in G. Moreover, this also means that $1 = \overline{xyx} = \overline{x}^3$ in G.

The upshot of these calculations is that for any $v \in F(x, y)$, since modulo N we can replace any y by x, we get $\overline{v} = \overline{x^i}$ for some $i \in \mathbb{Z}$. Then since $\overline{x}^3 = 1$, we even get $\overline{v} = \overline{x^i}$ with $0 \leq i \leq 2$. So $|G| \leq 3$.

To see that G actually has order 3 and is not smaller, it is enough to find a surjection from G onto a group of order 3. Let H be cyclic of order 3, where $H = \langle h\rangle$ so $|h| = 3$. There is a unique homomorphism $\phi : G \to H$ with $\phi(x) = h$ and $\phi(y) = h$, since both xyx and yxy evaluate to $h^3 = 1$

under the evaluation of x to h and y to h. Since ϕ is clearly surjective, this forces $|G| = 3$ and ϕ is an isomorphism. So G is cyclic of order 3.

Let us also do an example of a presentation of a infinite group.

Example 2.19. Consider $\mathbb{Z}^2 = \{(a, b)|a, b \in \mathbb{Z}\}$ under the operation of vector addition. It is easy to see that this is an abelian group. We claim that $G = \langle x, y|yx = xy\rangle$ is a presentation of $\mathbb{Z}^2$. Define a function $f : \{x, y\} \to \mathbb{Z}^2$ by $f(x) = (1, 0)$ and $f(y) = (0, 1)$. Since $\mathbb{Z}^2$ is additive, the relation $yx = xy$ evaluates under f to $(1, 0) + (0, 1) = (0, 1) + (1, 0)$, which is certainly true since $\mathbb{Z}^2$ is abelian. Thus there is a unique homomorphism of groups $\phi : \langle x, y|yx = xy\rangle \to \mathbb{Z}^2$ which restricts to f. The homomorphism ϕ is surjective because the set $\{(1, 0), (0, 1)\}$ generates $\mathbb{Z}^2$.

Now for $v \in G$ we write $\overline{v}$ for the image vN of v in $G = F(x, y)/N$, where N is the smallest normal subgroup of $F(x, y)$ containing $y^{-1}x^{-1}xy$. The relation $\overline{yx} = \overline{xy}$ tells us that $\overline{y}^j$ and $\overline{x}^i$ also commute for all $i, j \in \mathbb{Z}$. Thus for an arbitrary word $v \in F(x, y)$, by pushing all powers of y to the right we get $\overline{v} = \overline{x}^i\overline{y}^j$ for $i, j \in \mathbb{Z}$.

We have see that $G = \{\overline{x}^i\overline{y}^j|i, j \in \mathbb{Z}\}$. Now note that $\phi(\overline{x}^i\overline{y}^j) = (i, j) \in \mathbb{Z}^2$. This means that the elements $\overline{x}^i\overline{y}^j$ must be distinct for distinct ordered pairs (i, j), and that ϕ is injective and hence an isomorphism of groups.

We will see more examples of presentations of groups and how they are useful later on.

3. Group actions

3.1. **Definition and basic properties of actions.** Many groups can be naturally thought of as symmetries of other objects, such as the dihedral group which is the group of symmetries of a regular polygon. Each group element gives a way of permuting the points of the object while preserving its essential structure. We can think of a group element as "acting on" the object of which it is a symmetry, in the sense that applying the group element moves each point to another point. The idea of a group acting on a set is an abstraction of this. It will turn out to be an essential tool in the applications of groups as well as in understanding the structure of groups themselves.

Definition 3.1. Let X be a set and G a group. A (left) group action of G on X is a rule assigning an element $g \cdot x$ to each $x \in X$ and $g \in G$, where we think of $g \cdot x$ as the result of g acting on x. Formally this is a function $f : G \times X \to X$ where $f(g, x) = g \cdot x$. To be a group action this must satisfy

(i) $1 \cdot x = x$ for all $x \in X$.

(ii) $g \cdot (h \cdot x) = (gh \cdot x)$ for all $g, g, h \in G$, $x \in X$.

In words, the axioms for a group action say that the identity element acts trivially on all elements, and the result of acting by two group elements in succession is the same as the result of acting all at once by their product. As another consequence of the axioms, note that if $g \cdot x = y$, then $g^{-1} \cdot y = g^{-1} \cdot (g \cdot x) = g^{-1}g \cdot x = 1 \cdot x = x$. In other words, g^{-1} "undoes" whatever g does to points in X. When the context is clear, we often write gx instead of $g \cdot x$ unless this would lead to confusion.

We now give a series of examples. Usually verifying that the axioms of an action are satisfied is routine, and so we leave it to the reader without further comment.

Example 3.2. Let $G = S_n$ and $X = \{1, 2, \ldots, n\}$. Then G acts on X, where given $\sigma \in G$ and $i \in X$, $\sigma \cdot i = \sigma(i)$.

Example 3.3. Let $X = \mathbb{R}^n$, where we think of elements of X as column vectors, and $G = \mathrm{GL}_n(\mathbb{R})$. Then G acts on X by $A \cdot v = Av$ for $A \in G$ and $v \in X$. This is just the usual action of matrices on column vectors. We can also think of G as the group of linear symmetries of n-space.

By taking X to be related to the group G itself we obtain interesting actions which will play a key role in investigating the structure of groups further.

Example 3.4. Let G be a group and let $X = G$. Then G acts on X by left multiplication, where $g \cdot x = gx$ for $g, x \in G$. Note that axiom (ii) is just the associative property of G.

Example 3.5. Let G be a group and let $X = G$. Then G acts on X by conjugation, where $g \cdot x = {}^g x = gxg^{-1}$ for $g, x \in G$. (This is a case where it would be confusing to write this action as gx; the exponent notion ${}^g x$ is a convenient alternative).

Example 3.6. Given any action of G on X, if H is a subgroup of G then clearly we can restrict the action of G on X to an action of H on X with the same formula. For example, if G acts on itself by left multiplication, we can also consider the action of H on G by left multiplication.

Example 3.7. Let G be a group and let $H \leq G$ be a subgroup. Let $X = \{gH | g \in G\}$ be the set of left cosets of H in G. Then G acts on X by left multiplication: $g \cdot xH = gxH$. As usual, one must check that this formula for the action is well-defined.

Example 3.8. Let H be a subgroup of G. Let $X = \{xHx^{-1} | x \in G\}$ be the set of all conjugates of the subgroup H. Then G acts on X by conjugation: $g \cdot K = gKg^{-1}$ for $g \in G$, $K \in X$.

Example 3.9. There are many variations of the example above which take different sets of subgroups. For example, we could take $X = \{\text{subgroups of } G\}$ or $X = \{\text{subgroups of } G \text{ with order } d\}$. Really any set of subgroups which is closed under conjugation would suffice.

Group actions can be thought of in an alternate way which is conceptually very important. Let G act on X. Then we can define a function $\phi : G \to \mathrm{Sym}(X)$ where $\phi(g) = \phi_g$, with $\phi_g(x) = g \cdot x$ for $x \in X$. First of all, ϕ_g is indeed a bijection and hence an element of $\mathrm{Sym}(X)$, for $\phi_{g^{-1}} = (\phi_g)^{-1}$ since as we remarked earlier, g^{-1} undoes what g does. Then ϕ is a homomorphism of groups: since $\phi_{gh}(x) = gh \cdot x = g \cdot (h \cdot x) = \phi_g(\phi_h(x))$ for all x, we have $\phi_{gh} = \phi_g \circ \phi_h$ as functions.

Conversely, suppose that G is a group and X is a set, and we are given a homomorphism $\phi : G \to \mathrm{Sym}(X)$. Then we can define an action of G on X by $g \cdot x = [\phi(g)](x)$: first, $1 \cdot x = \phi(1)(x) = 1_X(x) = x$ since any homorphism sends 1 to 1, and second $g \cdot (h \cdot x) = \phi(g)(\phi(h)(x)) = [\phi(g) \circ \phi(h)](x) = \phi(gh)(x) = gh \cdot x$.

A quick calculation shows that these processes are inverse to each other, in other words if we start with an action and define the homorphism ϕ, the action obtained from ϕ is the original one; and if we start with a homomorphism ϕ and use it to define an action, the associated homomorphism is the original ϕ. Thus we have proved

Proposition 3.10. *For a fixed group G and set X, there is a bijection between actions of G on X and homomorphisms $\phi : G \to \mathrm{Sym}(X)$.*

This gives us two ways of thinking about what a group action is, both of which are useful. The definition focuses more on how a group element acts on the elements of X one at a time. The homomorphism version considers how each element of G acts on X as a whole.

One immediate application is known as Cayley's Theorem:

Theorem 3.11. *Every finite group G with $|G| = n$ is isomorphic to a subgroup of S_n.*

Proof. Let G act on itself by left multiplication. Let $\phi : G \to \mathrm{Sym}(G)$ be the corresponding homomorphism; thus writing $\phi(g) = \phi_g$, we have $\phi_g(h) = gh$. If $g \in \ker \phi$, then $\phi_g(h) = gh = h$ for all $h \in G$, which clearly forces $g = 1$. Thus ϕ is injective. Hence G is isomorphic to its image $\phi(G)$, which is a subgroup of $\mathrm{Sym}(G)$. Since G has n elements, clearly $\mathrm{Sym}(G) \cong S_n$. □

Cayley's Theorem suggests that we will understand all finite groups if we can sufficiently understand the symmetric groups and their subgroups. This sounds more promising than it actually is. Finite groups are very complicated in general, and Cayley's Theorem simply means that the structure of subgroups of symmetric groups must be horrendously complicated as well. In fact we

will usually get much more interesting information from other group actions than the action of G on itself by left multiplication.

Remark 3.12. We defined the notion of a "left" action of a group on a set. There is an analogous notion of a right action of a group G on a set X as well. This is a rule associating an element $x * g \in X$ to each $g \in G$ and $x \in X$, where $x * 1 = x$ and $(x * g) * h = x * (gh)$ for all $x \in X$, $g, h \in G$. Left and right actions are not quite the same concept; however, given a right action of G on X one can define a left action of G on X by $g \cdot x = x * g^{-1}$ for all $g \in G, x \in X$. This left action has all of the same information as the right action. For this reason we will not have any need to consider right actions below.

3.2. **Orbits and Stabilizers.** Let G act on a set X. We define a relation on X by $x \sim y$ if $y = gx$ for some $g \in G$. Note that $x \sim x$ since $x = 1x$. If $x \sim y$ with $y = gx$ then $x = g^{-1}y$ so that $y \sim x$. Finally, if $x \sim y$ and $y \sim z$, say with $y = gx$ and $z = hy$, then $z = hy = hgx$ and so $x \sim z$. We have proved that $\sim$ is an equivalence relation on X.

Given any equivalence relation on X, it partitions X into disjoint equivalence classes, where we write the class containing x as $\mathcal{O}_x$ and call it the *orbit* of x. By definition,

$$\mathcal{O}_x = \{y \in X | y = gx \text{ for some } g \in G\}.$$

Since the equivalence classes partition X, for each x and y either $x \sim y$ and $\mathcal{O}_x = \mathcal{O}_y$, or else $\mathcal{O}_x \cap \mathcal{O}_y = \emptyset$. We say that the action of G on X is *transitive* if there is only one orbit; so for any $x, y \in X$ there is $g \in G$ such that $gx = y$. For example, S_n clearly acts transitively on $\{1, 2, \ldots, n\}$.

Given an action of G on X, the *stabilizer* of $x \in X$ is $G_x = \{g \in G | gx = x\}$. It is easy to check that this is a subgroup of G. There is a close relationship between orbits and stabilizers, as we see now.

Theorem 3.13. *(Orbit-Stabilizer theorem) Let G act on a set X.*

(1) *Given $x \in X$, $|\mathcal{O}_x| = |G : G_x|$.*
(2) *if $gx = y$ for $x, y \in X$ and $g \in G$, then $G_y = gG_xg^{-1}$.*

Proof. (1) Let $S = \{gG_x | g \in G\}$ be the set of left cosets of G_x in G. Then $|S| = |G : G_x|$ by definition. Define a function $f : S \to \mathcal{O}_x$ by $f(gG_x) = gx$. To check that this is well-defined, note that if $gG_x = hG_x$, then $g^{-1}h \in G_x$ and so $g^{-1}hx = x$. Then acting on both sides by g we get $hx = gx$. It is obvious that f is surjective. If $gx = hx$, then $g^{-1}hx = x$ and so $g^{-1}h \in G_x$; hence $gG_x = hG_x$. This shows that f is also injective. Hence f is a bijection and so the cardinalities $|S| = |G : G_x|$ and $|\mathcal{O}_x|$ are equal.

(2) Note that if $h \in G_x$, then since $x = g^{-1}y$, we have $ghg^{-1}y = ghx = gx = y$. Thus $gG_xg^{-1} \subseteq G_y$. The same argument applied to $g^{-1}y = x$ with the roles reversed shows that $g^{-1}G_yg \subseteq G_x$. Multiplying by g on the left and g^{-1} on the right gives $G_y \subseteq gG_xg^{-1}$. Thus $G_y = gG_xg^{-1}$ as claimed. □

Many applications of group actions by finite groups arise from the following corollary.

Corollary 3.14. *Let G act on a set X. If $|G| < \infty$, then every orbit $\mathcal{O}$ of X is finite and $|\mathcal{O}|$ divides $|G|$.*

Proof. If $\mathcal{O} = \mathcal{O}_x$ then we have $|\mathcal{O}_x| = |G : G_x|$ by the Orbit-Stabilizer theorem. Since G is finite, the subgroup $|G_x|$ divides $|G|$ by Lagrange's theorem, and $|G : G_x| = |G|/|G_x|$ is also a divisor of $|G|$. □

We gave a number of examples of group actions earlier. Let us consider what the orbits look like for some of them and what information the orbit-stabilizer theorem tells us.

Example 3.15. Let $G = S_n$ act on $X = \{1, 2, \ldots, n\}$ as in Example 3.2. As we already remarked, this is a transitive action and has one orbit X. Hence $|X| = n = |G : G_i|$ for each $i \in X$, and so G_i is a subgroup of index n. Explicitly, G_i is the subgroup of permutations that fix the number i. This is clearly identified with the group of arbitrary permutations of the remanining $n-1$ numbers, and so each G_i is isomorphic as a group to S_{n-1}. It is clear that all of the G_i are different, though by Theorem 3.13 they are all conjugate in S_n.

Example 3.16. Let G act on $X = G$ by left multiplication as in Example 3.4. This is again a transitive action, since if $g, h \in G$, then $kh = g$ where $k = gh^{-1}$. There is one orbit and all stabilizers are trivial: $G_g = \{1\}$ for all g.

A bit more interesting is to restrict this action to some subgroup H of G, as in Example 3.6, so that H acts on G by left multiplication. Now the orbit $\mathcal{O}_g$ is clearly equal to the right coset Hg, and so there are $|G : H|$ orbits, each of size $|H|$. The stabilizers are again all trivial.

3.3. **Applications of orbit stabilizer.**

3.3.1. *Producing normal subgroups.* Given an action of G on X, we have seen that we can express it in terms of a homomorphism $\phi : G \to \mathrm{Sym}(X)$ instead. The kernel of this homomorphism $K = \ker \phi$ is a normal subgroup of G which we naturally call the *kernel* of the action. Since $\phi(g) = \phi_g$ where $\phi_g(x) = g \cdot x$, we see that $g \in K$ if and only if $\phi_g = 1_X$ or equivalently $g \cdot x = x$ for all x. Thus $K = \bigcap_{x \in X} G_x$ is the intersection of the stabilizer subgroups of all elements in X.

This is the part of G that is not "doing anything" in the action. In fact, if we wanted we could mod out by K and define an induced action of G/K on X by $gK \cdot x = g \cdot x$.

Taking kernels of actions is a useful way of producing normal subgroups in a group G, by finding an action of G on a set X and taking the kernel.

Theorem 3.17. *Let G be a group with subgroup H such that $|G : H| = m < \infty$.*

(1) *G has a normal subgroup K with $K \subseteq H$ and with $|G : K|$ dividing $m!$.*

(2) *If $|G| < \infty$ and m is the smallest prime dividing $|G|$, then $H \trianglelefteq G$.*

Proof. (1) Let G act on the set X of left cosets of H by $g \cdot xH = gxH$. Consider the corresponding homomorphism $\phi : G \to \mathrm{Sym}(X)$. Since $|X| = |G : H| = m$, $\mathrm{Sym}(X) \cong S_m$. In particular, $|\mathrm{Sym}(X)| = m!$. By the 1st isomorphism theorem, if $K = \ker \phi$ then $G/K \cong \phi(G)$. Also, by Lagrange's theorem, $|\phi(G)|$ divides $|\mathrm{Sym}(X)| = m!$. Thus K is a normal subgroup of G with $|G/K| = |G : K|$ dividing $m!$. Note that if $k \in K$ then in particular $k \cdot H = kH = H$, and so $k \in H$. Thus $K \subseteq H$.

(2) Suppose now that $m = p$ is prime and is the smallest prime dividing the order of G. Note that $p! = p(p-1)!$ and that all prime factors of $(p-1)!$ must be smaller than p. This implies that $\gcd(p!, |G|) = p$. Now $|G : K| = |\phi(G)|$ is a divisor of both $|G|$ and $p!$. Hence it divides p. Since $|G : H| = p$ already and $K \subseteq H$, we must have $K = H$. Thus $H \trianglelefteq G$. □

One can be more explicit about the subgroup K constructed in the previous result. Let G act on left cosets of H and consider the stabilizer subgroup $G_{xH} = \{g \in G | gxH = xH\}$ of some coset xH. We have $gxH = xH$ if and only if $x^{-1}gx \in H$ if and only if $g \in xHx^{-1}$. Thus each stabilizer subgroup $G_{xH} = xHx^{-1}$ is a conjugate of H. (This could also have been proved by using Theorem 3.13(2).) As observed above, the kernel of the action $K = \bigcap_{x \in G} G_{xH}$ is the intersection of all stabilizer subgroups, so $K = \bigcap_{x \in G} xHx^{-1}$. This subgroup is sometimes called the *core* of H. It is the unique largest subgroup of H which is normal in G.

Example 3.18. Suppose that G is a finite group with $|G| = p^m$ for some prime p. Such a group is called a *p-group*. If $H \leq G$ with $|G : H| = p$, then $H \trianglelefteq G$ by Theorem 3.17. We will study p-groups in more detail later on.

Example 3.19. We will construct later a group G with $|G| = 60$ such that G is *simple*, that is, where the only normal subgroups of G are G and $\{1\}$. Suppose that H is a subgroup of this simple group G, with $|G : H| = m$. Then by the theorem, G has a normal subgroup K contained in H

with $|G:K| \leq m!$. If $m \leq 4$ we get $|G:K| \leq 24$ and hence $\{1\} \subsetneq K \subseteq H \subsetneq G$, a contradiction. We conclude that the smallest possible index of a proper subgroup of G is 5.

3.3.2. *Products of subgroups.* Another application of the orbit-stabilizer theorem is the following formula for the size of a product of subgroups.

Lemma 3.20. *Let $H \leq G$ and $K \leq G$, with H and K finite. Then $|HK| = |K||H|/|K \cap H|$.*

Proof. Let G act on the left cosets of K as usual. We may restrict this action to H, so that H acts on left cosets of K by $h \cdot xK = hxK$. Now consider the orbit containing the coset $K = 1K$. This orbit is $\mathcal{O}_K = \{hK | h \in H\}$. The stabilizer of the coset K is

$$H_K = \{h \in H | hK = K\} = \{h \in H | h \in K\} = H \cap K.$$

By the orbit-stabilizer theorem we have $|\mathcal{O}_K| = |H|/|H \cap K|$. On the other hand, note that each element of $\mathcal{O}_K$ is itself a coset with $|K|$ elements, and the union of all of the elements in the cosets in $\mathcal{O}_K$ is HK. Thus $|HK| = |\mathcal{O}_K||K|$. Then $|HK| = |H||K|/|H \cap K|$. □

Note that if either H or K is normal in G, then the formula in the lemma easily follows from the 2nd isomorphism theorem. But it is occasionally useful to be able to know this formula holds regardless of whether or not HK is even a subgroup of G.

3.3.3. *Applications to counting.* Next we discuss an application of the orbit-stabilizer theorem to combinatorics. This section is optional reading and will not be covered in lecture, and you are not responsible for it on exams.

Sometimes when G acts on a set X we are especially interested in the *number* of orbits, and would like to know this information without first finding all of the orbits explicitly. There is an orbit-counting formula that is often very helpful in this regard.

Theorem 3.21. *Let a finite group G act on a finite set X. We define $\chi(g) = |\{x \in X | gx = x\}|$ for each $g \in G$. Then the number of orbits of the action is*

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

Proof. Consider the set $G \times X$ and its subset $S = \{(g, x) | gx = x\}$. Note that by considering one g at a time, we have $|S| = \sum_{g \in G} \chi(g)$. On the other hand, we can consider one x at a time. The set of $g \in G$ for which $gx = x$ is the stabilizer subgroup G_x. Thus $|S| = \sum_{x \in X} |G_x|$. We also know from

the orbit-stabilizer theorem that the orbit $\mathcal{O}_x$ containing x has size $|\mathcal{O}_x| = |G : G_x| = |G|/|G_x|$. Now we get

$$\frac{1}{|G|}\sum_{g\in G}\chi(g) = \sum_{x\in X}\frac{|G_x|}{|G|} = \sum_{x\in X}\frac{1}{\mathcal{O}_x}.$$

For each orbit $\mathcal{O}$, there are $|\mathcal{O}|$ terms in sum of the form $1/|\mathcal{O}|$ as x ranges over $x \in \mathcal{O}$. Thus in the final sum we get a contribution to the sum of 1 for each orbit, and so the sum is equal to the number of orbits. □

The formula is sometimes called "Burnside's counting formula" though it is not due to Burnside, but was known to Cauchy many years before Burnside popularized it.

The reason the formula is useful is that it is often easier to compute $\chi(g)$ for group elements g than it is to find the orbits and their sizes directly, especially if $|G|$ is much smaller than $|X|$. Note that $\chi(g)$ can be interpreted as the number of fixed points of g.

Example 3.22. One has an unlimited collection of black and white pearls and one wants to string r of them into a necklace. How many different necklaces are possible? Note that 2 necklaces are the same if they look alike after one is rotated or possibly flipped over.

The key to solving this problem is to interpret it in terms of a group action. We think of each necklace of r beads as sitting on a plane, arranged in a circle with center the origin. Then the dihedral group D_{2r} acts on the collection of all necklaces. By definition, two necklaces are considered the same if and only if they are in the same orbit of this action. So the solution to the problem is the number of orbits of this action.

The full set of possible necklaces (without considering which are deemed the same) is a set X where for each position we can choose one of 2 colors of pearls. Thus $|X| = 2^r$.

By the orbit counting formula, the number of orbits is

$$\frac{1}{|D_{2r}|}\sum_{g\in D_{2r}}\chi(g) = \frac{1}{2r}\sum_{g\in D_{2r}}\chi(g).$$

The fact that we chose pearls of two colors is not important, and the same method we present below would also work to count the number of necklaces with some larger number of different possible colors.

It is not difficult to develop from the expression above an explicit formula that works for all r, though the cases where r is even or odd are slightly different. For simplicity we work out the case when $r = 6$ only here, to demonstrate the method.

We have to consider the elements g of D_{12} one at a time and calculate how many fixed points they will have in their actions on the set of necklaces. Suppose first that g is a rotation. The

rotation subgroup $R = \langle a \rangle$ is cyclic of order 6. If $g = a$, then it is clear that if action by g leaves the necklace type fixed, since each pearl gets sent to its neighbor, all pearls must have the same color. So there are only 2 fixed necklaces. The same is true for $g = a^5 = a^{-1}$. If $g = a^2$ or a^4, then each pearl gets moved two places. This divides the 6 pearls into two groups of 3 which are permuted cyclically by this action. There are then $2^2 = 4$ necklaces that are fixed, since the pearls in each group can be chosen black or white independently. Similarly if $g = a^3$ there are $2^3 = 8$ fixed necklaces. Of course, when $g = 1$ all 2^6 necklaces are fixed. Finally, if g is a reflection, then either the axis of reflection goes through the centers of two pearls and flips the other pearls in two pairs—in this case there are 2^4 fixed necklaces; or the axis of reflection goes between the pearls and flips all of the pearls in three pairs—in this case there are 2^3 fixed necklaces. There are 3 reflections of each type. The final answer is $(1/12)(2^6 + 2^3 + 2(2^2) + 2(2) + 3(2^4) + 3(2^3)) = 13$ possibilities.

3.3.4. *The class equation.* Consider the action of G on itself by conjugation: $g \cdot x = gxg^{-1} = {}^g x$. The orbit of x, $\mathcal{O}_x = \{gxg^{-1} | g \in G\}$, is called a *conjugacy class* in this case and we write it as $\mathrm{Cl}(x)$ or $\mathrm{Cl}_G(x)$ if we need to emphasize in which group we are working. The stabilizer subgroup of x is

$$G_x = \{g \in G | gxg^{-1} = x\} = \{g \in G | gx = xg\} = C_G(x),$$

the centralizer of x in G. The orbit-stabilizer theorem now implies that $|\mathrm{Cl}(x)| = |G : C_G(x)|$. By Corollary 3.14, if G is finite than all conjugacy classes have order dividing the order of $|G|$. Note also that since conjugation preserves the order of an element (as conjugation gives an automorphism of the group), all members of a conjugacy class have the same order.

Example 3.23. Let G be a group and let $x \in G$. From the equation $|\mathrm{Cl}(x)| = |G : C_G(x)|$ we see that $\mathrm{Cl}(x)$ has one element if and only if $C_G(x) = G$. But the centralizer of x is the whole group G if and only if x is in the center, i.e. $x \in Z(G)$. We see that the elements that have conjugacy classes of size one are precisely the elements in the center of G. In particular, if G is abelian, then all conjugacy classes have size one.

Example 3.24. Let $G = D_{2n} = \{1, a, \ldots, a^{n-1}, b, ab, \ldots, a^{n-1}b\}$ be the dihedral group of order $2n$, where $n \geq 3$. Let us find the conjugacy classes of G. Let $x = a^i$ and consider $\mathrm{Cl}(x)$. If $g = a^j$ then $gxg^{-1} = x = a^i$ since g and x commute, while if $g = a^j b$ then $gxg^{-1} = a^j b a^i b^{-1} a^{-j} = a^{j-i} b b^{-1} a^{-j} = a^{j-i} a^{-j} = a^{-i}$. Hence $\mathrm{Cl}(a^i) = \{a^i, a^{-i}\}$. If $i = 0$ this is the one-element class $\{1\}$, and if n is even and $i = n/2$ then this is the one-element class $\{a^{n/2}\}$. Otherwise $\{a^i, a^{-i}\}$ is a class of two elements.

If $x = a^i b$ and $g = a^j$ then $gxg^{-1} = a^j a^i b a^{-j} = a^{i+j} a^j b = a^{2j+i} b$, while if $g = a^j b$ then $gxg^{-1} = a^j b a^i b b^{-1} a^{-j} = a^{j-i} a^j b = a^{2j-i} b$. We see that if n is odd then $\mathrm{Cl}(a^i b) = \{b, ab, \ldots, a^{n-1}b\}$ is the set of all reflections. If n is even, on the other hand, then the reflections break up into two conjugacy classes $\{b, a^2 b, \ldots, a^{n-2}b\}$ and $\{ab, a^3 b, \ldots, a^{n-1}b\}$, each of size n.

Since we understand the sizes of the conjugacy classes, we automatically get information about the centralizers of elements. Note that when n is odd, $Z(D_{2n}) = \{1\}$, while if n is even, $Z_{D_{2n}} = \{1, a^{n/2}\}$. This follows from the calculation of which conjugacy classes have size 1. If $\{a^i, a^{-i}\}$ is a conjugcacy class of size 2, then $|G : C_G(a^i)| = 2$, so $|C_G(a^i)| = n$. Clearly then $C_G(a^i) = \{1, a, \ldots, a^{n-1}\}$ is the rotation subgroup, since this is an abelian subgroup of order n containing a^i. If n is odd, then $|\mathrm{Cl}(a^i b)| = n$ and so $|C_G(a^i b)| = 2$. Thus in ths case $C_G(a^i b) = \langle a^i b\rangle = \{1, a^i b\}$ must be the cyclic subgroup of order 2 generated by the reflection $a^i b$. On the other hand, if n is even then we get that $|C_G(a^i b)| = 4$. Again this centralizer contains $\langle a^i b\rangle = \{1, a^i b\}$ but it also contains the non trivial center Z. Thus $C_G(a^i b)$ must be the product $\langle a^i b\rangle Z = \{1, a^i b, a^{n/2}, a^{i+n/2} b\}$ in this case, since this already contains 4 distinct elements.

Suppose that G is finite. The information given by the orbit-stabilizer theorem applied to the conjugaction action of G on itself is often organized into a form called the *class equation*, which is especially useful for deriving consequences about the center $Z(G)$. The equation is

$$|G| = |Z(G)| + \sum_x |G|/|C_G(x)|, \tag{3.25}$$

where the sum runs over one representative x of each conjugacy class of size bigger than 1. The equation is just a way of expressing that there are $|Z(G)|$ conjugacy classes of size 1, and picking one x from each conjugacy class of bigger size, the class $\mathrm{Cl}(x)$ has size $|\mathrm{Cl}(x)| = |G|/|C_G(x)|$. Then since G is the disjoint union of its conjugacy classes, the formula follows.

The class equation will be a key tool in proving the Sylow Theorems in the next section. Here is an immediate interesing application.

Theorem 3.26. *Let G be a group of order p^m for some prime p and $m \geq 1$. Then $|Z(G)|$ is a multiple of p. In particular, $Z(G)$ is nontrivial.*

Proof. Let $|G| = p^m$ where $m \geq 1$. Consider the class equation for G. Each term $|G|/|C_G(x)|$ in the sum is the size of an conjugacy class not of size 1. Since it is a divisor of $|G|$, it is a prime power p^i for some $i \geq 1$. Thus p divides every term in $\sum_x |G|/|C_G(x)|$. Since p also divides $|G|$, from the class equation we see that p divides $|Z(G)|$. □

The following fact is sometimes called the "G/Z-theorem". We leave it as an exercise.

Lemma 3.27. *Let G be a group with center $Z = Z(G)$. If G/Z is cyclic, then G is abelian.*

Ultimately, one of the goals of group theory is to classify groups of certain types. For example, given an integer n, one would like to be able to give a list of groups of that order such that every group of order n is isomorphic to exactly one group on the list. We would then say that we have classified groups of order n "up to isomorphism". This goal is attainable only for certain special values of n; in general, groups are too complicated and one must settle for less exact kinds of results.

We can use the results developed so far to classify groups of order p and p^2, where p is a prime.

Theorem 3.28. *Let G be a group and let p be a prime.*

(1) *If $|G| = p$ then $G \cong \mathbb{Z}_p$.*

(2) *If $|G| = p^2$ then either $G \cong \mathbb{Z}_{p^2}$ or $G \cong \mathbb{Z}_p \times \mathbb{Z}_p$.*

Proof. (1) Let x be any non-identity element of G. Then $|x|$ is a divisor of $|G| = p$ by Corollary 1.52, and $|x| \neq 1$. So $|x| = p$. This means that $|\langle x \rangle| = p$ and hence $\langle x \rangle = G$. But $\langle x \rangle \cong (\mathbb{Z}_p, +)$ by Theorem 1.51.

(2) First we show that G is abelian. By Theorem 3.26, p groups have a nontrivial center $Z = Z(G)$, and so $|Z| = p$ or $|Z| = p^2$. If $|Z| = p$, then $|G/Z| = p$. By part (1), the group G/Z is cyclic. Then by the G/Z-theorem (Lemma 3.27), G is abelian, contradicting $|Z| = p$. Thus $Z = G$ and $|Z| = p^2$.

Now suppose that G has an element x of order p^2. In this case $G = \langle x \rangle \cong (\mathbb{Z}_{p^2}, +)$, similarly as in part (1). Otherwise, since all elements have order dividing $|G|$, all nonidentity elements of G have order p. Let $x \neq 1$ and let $H = \langle x \rangle$. Then $|H| = p$. Pick $y \notin H$ and let $K = \langle y \rangle$. Then $|K| = p$ as well. $H \cap K$ is a subgroup of K and is not equal to K (since $y \notin H$), so by Lagrange's theorem, $|H \cap K| = 1$ and $H \cap K = \{1\}$.

Consider the function $\phi : H \times K \to G$ given by $\phi(h, k) = hk$. This is a homomorphism, because

$$\phi((h_1, k_1)(h_2, k_2)) = \phi((h_1h_2, k_1k_2)) = h_1h_2k_1k_2 = h_1k_1h_2k_2 = \phi((h_1, k_1))\phi((h_2, k_2)),$$

using that G is abelian. If $(h, k) \in \ker \phi$, then $hk = 1$, so $h = k^{-1} \in H \cap K = \{1\}$, forcing $h = k = 1$. Thus $\ker \phi$ is trivial and ϕ is injective. Now $|G| = p^2 = |H \times K|$. An injective function between sets of the same size is bijective. Thus ϕ is an isomorphism. Finally, $H \cong K \cong (\mathbb{Z}_p, +)$ by part (1), so $H \times K \cong \mathbb{Z}_p \times \mathbb{Z}_p$. □

It is also fairly easy to classify groups of order p^3 for a prime p. These are most easily described using semi-direct products, which are defined later. Groups of order p^n become complicated very quickly as n grows, and a full classification is known only for small n ($n \leq 7$).

4. Sylow Theorems

Lagrange's theorem shows that a subgroup H of a finite group G must have order dividing the order of the group. The converse question is much harder: given a divisor d of $|G|$, where G is a finite group, when must G have a subgroup of order d?

If one starts cataloguing examples of finite groups of small order, one would quickly see that the answer is not always. The alternating group A_4 has order 12 but no subgroup of order 6 (we will define A_4 in the next section and show this fact). This is the smallest possible such example. The full symmetric group S_4, which has order 24 (and of which A_4 is a subgroup) also has no subgroup of order 6.

On the other hand, the Sylow Theorems show that if d divides $|G|$ and $d = p^i$ is a power of a prime, then G does in fact have a subgroup of order d. This is the strongest positive result in this direction. The theorems will also give information about how many subgroups of order p^i one should expect when p^i is the largest power of p dividing $|G|$. These are the most powerful basic results for understanding the structure of finite groups.

Definition 4.1. Let p be a prime. A finite group G is a *p-group* if $|G| = p^m$ for some $m \geq 0$.

Definition 4.2. Let G be a finite group. Let p be a prime with $|G| = p^m k$ where $\gcd(p, k) = 1$; in other words, p^m is the largest power of p dividing $|G|$. A *Sylow p-subgroup* of G is a subgroup H with $|H| = p^m$.

We will see soon that Sylow p-subgroups always exist for any prime p dividing $|G|$. As a first step, we show an important result known as Cauchy's Theorem, in the special case of an abelian group.

Theorem 4.3. *(Cauchy's Theorem for abelian groups) Let G be a finite abelian group and let p be a prime divisor of $|G|$. Then G has an element of order p.*

Proof. We induct on the order of G, assuming the result is true for all groups of smaller order. If $|G| = 1$ the result is trivial, so the base case holds. Assume that $|G| \neq 1$ and pick any $1 \neq x \in G$. Consider the order $|x|$ of x. Suppose first that p divides $|x|$, say $|x| = pk$. Then it is easy to see that $|x^k| = p$. So we have found an element of order p. On the other hand, suppose that p does not divide $|x|$. Then $H = \langle x \rangle$ has order $|\langle x \rangle| = |x|$ which is relatively prime to p. It follows that the factor group G/H (which makes sense since G is abelian and hence all subgroups are normal) has order $|G/H| = |G|/|H|$, which is divisible by p. Since $|G/H| < |G|$, the induction hypothesis tells us that G/H has an element of order p, say yH. Consider $|y|$. If $y^n = 1$, then certainly

$(yH)^n = y^nH = 1H = H$. Thus n is a multiple of the order of yH in G/H, which is p. Now we again have an element y of order which is a multiple of p, with $|y| = n = p\ell$, say. Then $|y^\ell| = p$. □

4.1. **Sylow Existence.** We now prove that Sylow subgroups exist. Because more or less the same argument works, we show in fact that there exist groups of any prime power order dividing the order of the group.

Theorem 4.4. *(Sylow existence) Let G be a finite group with $|G| = p^m k$, where p is prime and $\gcd(p, k) = 1$. Then for all $0 \leq i \leq m$, the group G has a subgroup of order p^i. In particular, G has a Sylow p-subgroup, that is, a subgroup H with $|H| = p^m$.*

Proof. We induct on the order of G. Assume we know the result for all groups of order smaller than $|G|$. There is nothing to do when $m = 0$, so assume that $m \geq 1$ and p divides $|G|$.

Consider the class equation $|G| = |Z(G)| + \sum_x |G|/|C_G(x)|$, where x runs over a set of representatives for the conjugacy classes of size bigger than 1. Suppose first that p does not divide $|Z(G)|$. Since p divides $|G|$, p must not divide one of the terms in the sum. So there is x such that $|G|/|C_G(x)|$ is not a multiple of p. This forces $|C_G(x)| = p^m\ell$ where $\gcd(p, \ell) = 1$. But $|C_G(x)| < |G|$ since $|G : C_G(x)| = |\operatorname{Cl}(x)|$ is at least 2, because x is in a conjugacy class of size bigger than 1. By induction, for any i we choose with $0 \leq i \leq m$, the subgroup $C_G(x)$ has a subgroup H with $|H| = p^i$. But of course H is a subgroup of G as well, of the desired order.

On the other hand, suppose that p does divide $|Z(G)|$. Since $Z(G)$ is an Abelian group, by Theorem 4.3, the abelian group $Z(G)$ has an element of order p, say x. Since $x \in Z(G)$, the cyclic subgroup generated by x satisfies $\langle x\rangle \trianglelefteq G$ and $|\langle x\rangle| = p$. So we can form the factor group $\overline{G} = G/\langle x\rangle$, where $|\overline{G}| = |G|/p = p^{m-1}k$. By the induction hypothesis, for each $0 \leq i \leq m-1$, $\overline{G}$ has a subgroup of order p^i. By the correspondence theorem, this subgroup has the form $H/\langle x\rangle$ for some subgroup H of G with $\langle x\rangle \leq H \leq G$. Moreover, since $|H/\langle x\rangle| = |H|/|\langle x\rangle| = p^i$, we must have $|H| = p^{i+1}$. This gives subgroups of G of orders p^j for all $1 \leq j \leq m$. But because it is trivial to find a subgroup of order $p^0 = 1$, we get subgroups of all orders p^j with $0 \leq j \leq m$ as needed. □

An immediate consequence is Cauchy's Theorem for a general (not necessarily abelian) finite group.

Corollary 4.5. *(Cauchy's Theorem) Let G be a finite group. Let p be a prime dividing $|G|$. Then G has an element of order p.*

Proof. By Theorem 4.4, G has a subgroup of order p, say H. Choosing any $x \neq 1$ in H, we must have $|x| = p$ by Lagrange's theorem. □

4.2. **Sylow conjugation and Sylow counting.** Now that we know that a finite group G has a Sylow p-subgroup for every prime p that divides its order, the next question is how many distinct such Sylow p-subgroups G has. The knowledge of this number, or at least knowing that this number lies among a small list of possibilities, often gives important information about the structure of G.

Given a Sylow p-subgroup P of G, there is an obvious way to potentially produce other Sylow p-subgroups: if $\sigma \in \mathrm{Aut}(G)$, then $\sigma(P)$ is clearly again a Sylow p-subgroup. We may not know about the structure of $\mathrm{Aut}(G)$, but at least we know that G has inner automorphisms, and so each conjugate xPx^{-1} of P will again be a Sylow p-subgroup. We will now see that all of the Sylow p-subgroups arise in this way from a given one through conjugation. In fact we can show that any p-subgroup is contained in a conjugate of any fixed Sylow p-subgroup.

Theorem 4.6. *(Sylow conjugates) Let G be a finite group and let p be a prime dividing $|G|$. let P be a Sylow p-subgroup of G. Suppose that Q is any p-subgroup of G. Then there is $g \in G$ such that $Q \subseteq gPg^{-1}$. In particular, if Q is a Sylow p-subgroup then $Q = gPg^{-1}$ for some $g \in G$.*

Proof. The key to this result is to consider a non-obvious group action and to which we apply the orbit-stabilizer theorem. Let G act on the set $X = \{gP | g \in G\}$ of left cosets of P by left multiplication; this is just the standard action of Example 3.7. Now restrict this action to the subgroup Q of G and let Q act on X.

Consider the orbit-stabilizer theorem for the action of Q on X. Every orbit has size dividing $|Q|$, which is therefore a power of p. On the other hand, $|X| = |G : P| = |G|/|P|$, which is not divisible by p, since P is a Sylow p-subgroup. Since X is the disjoint union of its orbits, it follows that some orbit of the Q-action has size which is not a multiple of p. The only possible conclusion is that there exists an orbit of size $p^0 = 1$.

Let $\{gP\}$ be such an orbit of size 1. Then for all $q \in Q$, we have $qgP = gP$. This is equivalent to $g^{-1}qg \in P$, or $q \in gPg^{-1}$, for all $q \in Q$. Thus $Q \subseteq gPg^{-1}$ for this g, proving the first statement.

Now apply this result to any Sylow p-subgroup Q of G. We get that $Q \subseteq gPg^{-1}$ for some g. But $|Q| = |gPg^{-1}|$ since both are Sylow p-subgroups. This forces $Q = gPg^{-1}$. □

The conclusion that "all Sylow p-subgroups of G are conjugate" is the easiest part of the preceding theorem to remember, but the more general first statement—that any p-subgroup is contained in a conjugate of a Sylow p-subgroup—is often useful as well.

The last Sylow theorem gives some numerical restrictions that the number of Sylow p-subgroups has to satisfy. These restrictions are often enough to calculate this number in simple examples, or at least narrow down the list of possibilities.

Theorem 4.7. *(Sylow counting) Let G be a finite group. Let p be a prime and write $|G| = p^m k$ where $\gcd(p, k) = 1$. Let n_p be the number of distinct Sylow p-subgroups of G. Then*

(1) *$n_p = |G : N_G(P)|$ for any Sylow p-subgroup P. In particular, n_p divides k.*

(2) *$n_p \equiv 1 \mod p$.*

Proof. (1) Fix a Sylow p-subgroup P and let $X = \{gPg^{-1} | g \in G\}$ be the set of conjugates of P. By Theorem 4.6, X is the set of all Sylow p-subgroups of G. Let G act on X by conjugation. Again by Theorem 4.6, this action is transitive, in other words the orbit $\mathbb{O}_P$ of P is all of X. Then by the orbit-stabilizer theorem, $|X| = |G : G_P|$ where G_P is the stabilizer of P. But $G_P = \{g \in G | gPg^{-1} = P\} = N_G(P)$ is the normalizer of P by definition. So $|X| = n_p = |G : N_G(P)|$. Since $P \subseteq N_G(P)$, $|G : N_G(P)|$ is a divisor of $|G : P| = k$.

(2) Now restrict the action of G on X by conjugation to the subgroup P, so P acts on the set of Sylow p-subgroups by conjugation. In this case the orbit-stabilizer theorem gives us different information. In particular, the size of every orbit of this action divides $|P|$ and thus must be a power of p. Note that $\{P\}$ is an orbit of this action, since $xPx^{-1} = P$ for all $x \in P$. Suppose conversely that $\{Q\}$ is a singleton orbit. Then $gQg^{-1} = Q$ for all $g \in P$, in other words, $P \subseteq N_G(Q)$. By Proposition 1.29, this means that PQ is a subgroup of G. Now $|PQ| = |P||Q|/|P \cap Q|$ by Lemma 3.20 (or the 2nd isomorphism theorem). Since $|P|, |Q|$, and $|P \cap Q|$ are all powers of p, $|PQ|$ must be a power of p. But $P \subseteq PQ$ and P is a Sylow p-subgroup, so this forces $PQ = P$. Thus $Q \subseteq P$. Since Q and P are both Sylow p-subgroups, $P = Q$.

We have shown that there is exactly one orbit of size one, namely $\{P\}$. All other orbits have size a power of p. Since X is the disjoint union of the orbits of the P-action, it follows that $|X| = n_p \equiv 1 \mod p$. □

One of the useful consequences of knowing the number of Sylow p-subgroups of a group G is that we can tell if a Sylow p-subgroup is normal or not.

Corollary 4.8. *Let G be a finite group and let p be a divisor of $|G|$. The following are equivalent:*

(1) *There is exactly one Sylow p-subgroup of G.*

(2) *G has a characteristic Sylow p-subgroup.*

(3) *G has a normal Sylow p-subgroup.*

Proof. (1) $\implies$ (2): If P the unique Sylow p-subgroup of G, then if $\sigma \in \mathrm{Aut}(G)$, $\sigma(P)$ is also a Sylow p-subgroup and hence $\sigma(P) = P$. So $P \operatorname{char} G$.

(2) $\implies$ (3): this is obvious.

(3) $\implies$ (1): If P is a Sylow p-subgroup of G with $P \trianglelefteq G$, then the number n_p of Sylow p-subgroups is $n_p = |G : N_G(P)| = |G : G| = 1$. □

4.3. **Examples of the use of the Sylow theorems.**

Example 4.9. Let us consider groups G with order $|G| = pq$, where $p < q$ are distinct primes. By the Sylow Existence Theorem (or Cauchy's Theorem), G has a subgroup P with $|P| = p$ and a subgroup Q with $|Q| = q$. The subgroup $P \cap Q$ is contained in P and Q and so has order dividing both p and q. Since p and q are distinct primes, $|P \cap Q| = 1$ so $P \cap Q$ is trivial. By Lemma 3.20, $|PQ| = |P||Q|/|P \cap Q| = pq = |G|$. Thus $PQ = G$.

Let n_q be the number of Sylow q-subgroups. By the information given by the Sylow counting theorem, n_q divides p and $n_q \equiv 1 \mod q$. Thus n_q is 1 or p, but since $p < q$, $p \equiv 1 \mod q$ is impossible. Thus $n_q = 1$, which gives $Q \trianglelefteq G$ by Corollary 4.8. (One could also show that $Q \trianglelefteq G$ in this case by observing that the index $|G : Q| = p$ is the smallest prime dividing the order of G.)

Consider now the number n_p of Sylow p-subgroups. The Sylow counting theorem gives n_p divides q and $n_p \equiv 1 \mod p$. Either $n_p = 1$ or $n_p = q$. In the latter case we must have $q \equiv 1 \mod p$, or p divides $(q - 1)$. We see that if p does not divide $(q - 1)$, then $n_p = 1$ and so $P \trianglelefteq G$ as well.

Suppose $P \trianglelefteq G$. We claim that in this case we have $G \cong P \times Q$. We will have a general result later about "recognizing internal direct products" which implies this, but for the moment let us just show it in this case directly. First, note that if $x \in P$ and $y \in Q$ then $xyx^{-1}y^{-1} = (xyx^{-1})y^{-1} \in Q$ since Q is normal, and $= x(yx^{-1}y^{-1}) \in P$ since P is normal. But $P \cap Q = 1$, so $xyx^{-1}y^{-1} = 1$, or $xy = yx$. This shows that the elements of P commute with the elements of Q. Now define $\phi : P \times Q \to G$ by $\phi(x, y) = xy$. Since P commutes with Q, if $x_1, x_2 \in P$ and $y_1, y_2 \in Q$ we have

$$\phi((x_1, y_1)(x_2, y_2)) = \phi(x_1x_2, y_1y_2) = x_1x_2y_1y_2 = x_1y_1x_2y_2 = \phi((x_1, y_1))\phi((x_2, y_2))$$

and so ϕ is a homomorphism. Since $PQ = G$, ϕ is surjective. Since $|P \times Q| = pq = |G|$, ϕ must automatically be injective as well and hence an isomorphism. Now note that since P and Q have prime order, they are cyclic and thus $P \cong (\mathbb{Z}_p, +)$ and $Q \cong (\mathbb{Z}_q, +)$. Thus $G \cong \mathbb{Z}_p \times \mathbb{Z}_q$. Moreover, we will prove later when we study direct products that $\mathbb{Z}_p \times \mathbb{Z}_q \cong \mathbb{Z}_{pq}$, in other words G must itself be cyclic.

We will also see later that in the case where P is not normal in G, the group G can still be described by a more general construction called a semi-direct product.

The example above already gives a classification result for groups of certain orders:

Proposition 4.10. *Suppose that $n = pq$ where p and q are primes with $p < q$ for which p does not divide $q - 1$. Then any group G of order n is cyclic and isomorphic to $(\mathbb{Z}_{pq}, +)$. Thus there is only one group of order n up to isomorphism.*

A useful exercise in reinforcing the techniques of group theory is to try to classify all groups of order n up to isomorphism for small n. Consider for example $n < 36$. So far, we know that groups of prime order p are cyclic, which handles $n = 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31$; groups of order p^2 for a prime p are cyclic or else isomorphic to $\mathbb{Z}_p \times \mathbb{Z}_p$, which handles $n = 4, 9, 25$; and now we know that for $n = 15 = (3)(5), 33 = (3)(11)$, and $35 = (5)(7)$, again all groups of order n are cyclic. We will develop enough techniques below to handle the remaining orders, except $n = 16$. Groups of order 16 are technically more complicated because $16 = 2^4$ is a large power of a prime. There happen to be 14 isomorphism classes of groups of order 16, so clearly the classification of those is more sensitive.

Rather than trying to classify all groups of order n, often one is looking for less exact information. Recall that a group G is *simple* if G is nontrivial and $\{1\}$ and G are the only normal subgroups of G. Having a normal subgroup allows one to take a factor group and apply inductive arguments, so because of their lack of normal subgroups simple groups tend to be the hardest groups to understand. The classification of finite simple groups was one of the major projects in algebra in the last century. One of the first questions in this project is which orders n can possibly be the order of a simple group. Because the Sylow theorems often allow us to show that a Sylow subgroup must be normal, they can be used to show that groups of certain orders n cannot be simple.

Example 4.11. Let G be a group of order p^2q where p and q are distinct primes. We will show that G must have either a normal Sylow p-subgroup or a normal Sylow q-subgroup. In particular, G cannot be simple.

Let n_p be the number of Sylow p-subgroups, and n_q the number of Sylow q-subgroups. From the Sylow counting theorem we have n_p divides q (so $n_p \in \{1, q\}$) and $n_p \equiv 1 \mod p$; and n_q divides p^2 (so $n_q \in \{1, p, p^2\}$) and $n_q \equiv 1 \mod q$.

If $n_p = 1$, then $P \trianglelefteq G$ for a Sylow p-subgroup P. Similarly, if $n_q = 1$ then $Q \trianglelefteq G$ for a sylow q-subgroup Q. So we will assume that $n_p = q$ and $n_q \in \{p, p^2\}$ and seek a contradiction. If $q < p$, then $q \not\equiv 1 \mod p$, so this is ruled out. Thus assume $p < q$. If $n_q = p$, then we again get a contradiction because $p \not\equiv 1 \mod q$. So we can assume $n_q = p^2$.

To finish, we rule out the possibility that $n_p = q$ and $n_q = p^2$ through the technique of "element counting". Each Sylow q-subgroup has order q, so if Q and Q' are distinct Sylow q-subgroups, then

$|Q \cap Q'|$ is a proper divisor of q and hence is equal to 1. This shows that any two distinct Sylow q-subgroups intersect trivially. Now consider which elements of G have order q. Every nontrivial element of a Sylow q-subgroup Q has order q, and any element x with $|x| = q$ generates a cyclic subgroup of order q. Thus the elements of order q are exactly the nontrivial elements contained in the Sylow q-subgroups. Thus there are $n_q(q-1)$ elements of order q, since each Sylow q-subgroup contains $q-1$ elements of order q once the identity is excluded, and none of these order q elements are common to two Sylow q-subgroups. Since we are assuming that $n_q = p^2$, this gives $p^2(q-1)$ elements of order q. That leaves $p^2q - (p^2)(q-1) = p^2$ elements in the group unaccounted for. Let P be any Sylow p-subgroup of G. Then $|P| = p^2$ and none of the elements in P can have order q, by Lagrange's theorem. This implies that P is exactly the elements in G which do not have order q. However, this means that there is exactly one Sylow p-subgroup, so $n_p = 1$, a contradiction.

Element counting, as in the example above, works best when the group order n has a prime factor q occuring to the first power in the prime factorization of n. For example, suppose in the example above we instead tried to count elements of order p to acheive a contradiction. Now since Sylow p-subgroups have order p^2, it is not true that any two distinct Sylow p-subgroups intersect trivially; they could intersect in a subgroup of order p. In addition, maybe a Sylow p-subgroup is cyclic and so has some elements of order p^2. So things are more complicated.

Here is an example which shows that if one's goal is just to show groups of a particular order are not simple, we can combine techniques from the Sylow theorems with other ideas, in particular taking the kernel of a group action.

Example 4.12. Let $|G| = p^3q$ for distinct primes p and q. We aim to show that G is not a simple group. Most of this can be done exactly as in Example 4.11, and so we don't repeat the details. In particular, we can assume that $n_p = q$ and $n_q \in \{p, p^2, p^3\}$ since otherwise some Sylow subgroup is normal; $q < p$ and $n_p = q$ contradict $n_p \equiv 1 \mod p$, so $p < q$; $n_q = p$ contradicts $n_q \equiv 1 \mod q$; and finally $n_q = p^3$ leads to a contradiction by counting elements of order q.

The only case that needs to be analyzed in a different way from Example 4.11 is $p < q$, $n_p = q$, and $n_q = p^2$. since $n_q \equiv 1 \mod q$, this means q divides $p^2 - 1 = (p-1)(p+1)$. Since q is prime, either q divides $p-1$ or q divides $p+1$. Since $p < q$, this quickly leads to a contradiction unless $q = p+1$. This can happen only if $p = 2$ and $q = 3$, so $|G| = 24$. In fact, there are groups of order 24 in which neither a Sylow 2-subgroup nor a Sylow 3-subgroup is normal, namely the symmetric group S_4.

Since the goal is just to prove that G is not simple, in this last case we consider a group action instead. We are assuming that there are 3 Sylow 2-subgroups. Let G act on the set of Sylow 2-subgroups by conjugation. This gives a homomorphism of groups $\phi : G \to S_3$. We know that all Sylow 2-subgroups are conjugate, so the action has one orbit. In particular this means that $\ker(G) \neq G$ since the action is not trivial trivial. Also, since $|G|/|\ker(G)| = |\phi(G)| \leq |S_3| = 6$, $\ker(G) \neq \{1\}$. Thus $\ker(G) \trianglelefteq G$ is a nontrivial proper normal subgroup and so G is not simple.

Here is an example where one can make use of the more precise information that $n_p = |G : N_G(P)|$ in the Sylow counting theorem, rather than just that n_p divides $|G : P|$.

Example 4.13. Let G be a group with $|G| = 105 = (3)(5)(7)$. We know that n_3 divides 35 and is congruent to 1 mod 3, so $n_3 \in \{1, 7\}$. Similarly we get $n_5 \in \{1, 21\}$ and $n_7 \in \{1, 15\}$. Thus the simple divisibility and congruence conditions coming from Sylow counting do not allow us to immediately conclude that any of n_3, n_5, or n_7 is equal to 1. However, we will see that in fact $n_5 = n_7 = 1$.

Consider n_3. If P is a Sylow 3-subgroup, then $n_3 = |G : N_G(P)| \in \{1, 7\}$ which means that $|N_G(P)| \in \{15, 105\}$. If $|N_G(P)| = 15$, then let Q be a Sylow 5-subgroup of $N_G(P)$. If $N_G(P) = 105$ then let Q be any Sylow 5-subgroup of G. Either way, we see that $Q \leq N_G(P)$ and so $H = PQ$ is a subgroup of G. By Lagrange's theorem, $|P \cap Q| = 1$. Thus $|PQ| = |P||Q|/(P \cap Q) = 15$.

Now by Proposition 4.10, every group of order 15 has normal Sylow 3 and 5-subgroups (and is in fact cyclic). Thus $Q \trianglelefteq H$ which means that $|N_G(Q)|$ is a multiple of 15. In turn, since $|G : N_G(Q)| = n_5$, we get n_5 divides 15. Since $n_5 \in \{1, 21\}$ we conclude that $n_5 = 1$ after all. Thus $Q \trianglelefteq G$.

In addition, now that we know that G has a normal sylow 5-subgroup Q, that means if R is a Sylow 7-subgroup then QR is a subgroup of G, with $|QR| = 35$. By Proposition 4.10 again, groups of order 35 have normal Sylow subgroups and are cyclic. So $|N_G(R)| \geq 35$ and since $n_7 = |G : N_G(R)| \in \{1, 15\}$ we also get $n_7 = 1$. So $R \trianglelefteq G$ as well.

We will give more examples later once we develop the techniques of semidirect products, when we will be in a better position to classify groups of other small orders.

5. Symmetric and Alternating groups

5.1. **Cycle notation in S_n.** In this section we discuss some of the important results for the symmetric groups. Since we have not yet done much with S_n we begin by reviewing some of the basic results and notation for these groups.

Recall that $S_n = \mathrm{Sym}(X)$ for $X = \{1, 2, \ldots, n\}$. One notation for an element $\sigma \in S_n$ is to give a $2 \times n$ matrix in which the ith column consists of i and $\sigma(i)$. Since the numbers in X can occur in any order in the bottom row, defining a unique permutation, it is clear that $|S_n| = n!$, the number of ways of ordering n distinct numbers.

Example 5.1.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 5 & 3 & 1 & 2 & 4 & 7 & 9 & 8 \end{pmatrix}$$

represents the element $\sigma \in S_9$ for which $\sigma(1) = 6$, $\sigma(2) = 5$, $\sigma(3) = 3$, etc.

For most purposes a much better notation for a permutation is the cycle notation we develop next. If $a_1, a_2, \ldots, a_k$ are k distinct numbers in X, then we can define an element $\sigma \in S_n$ such that $\sigma(a_i) = a_{i+1}$ for $1 \leq i \leq k-1$, $\sigma(a_k) = a_1$, and $\sigma(b) = b$ for all b such that $b \neq a_i$ for all i. Such a permutation is called a *k-cycle* and we have the special notation $(a_1 a_2 \ldots a_k)$ for σ. There is no preference for which element is listed first in the cycle notation, and any k-cycle can be written in k different ways: for example, $(123) = (231) = (312)$. Note that a 1-cycle (a) is the same as the identity element 1 in S_n. A 2-cycle (ab) is also called a *transposition*.

Example 5.2. Recall that the product in S_n is composition. As usual we omit notation for the product in most cases, but the reader must remember that functions are composed from right to left. On the other hand, the notation for a cycle is read left to right. For example, consider $\sigma = (12)(23)(123) \in S_3$. To find $\sigma(1)$, first applying (123) sends 1 to 2; then applying (23) to the element 2 sends it to 3; then applying (12) to the element 3 yields 3. So $\sigma(1) = 3$. The reader may check similarly that $\sigma(2) = 1$ and $\sigma(3) = 2$. So $\sigma = (132)$.

Two permutations τ, σ are called *disjoint* if for all $a \in X$, either $\tau(a) = a$ or $\sigma(a) = a$. Note that two cycles $(a_1 a_2 \ldots a_k)$ and $(b_1 b_2 \ldots b_l)$ are disjoint if and only if $a_i \neq b_j$ for all i, j, in other words all of the $k + l$ elements appearing in the notation are distinct.

We leave the proof of the following basic result to the reader.

Lemma 5.3. *Let $G = S_n$.*

(1) *If τ and σ are disjoint then $\tau\sigma = \sigma\tau$.*

(2) *Every permutation in S_n can be written as a product of pairwise disjoint cycles of length at least 2. This representation is unique up to the order in which we write the cycles in the product. We call this representation* disjoint cycle form.

Example 5.4. Consider the permutation $\sigma \in S_9$ in Example 5.1. It is easy to find its disjoint cycle form. One can start with any integer. Beginning with 1, following down the columns gives $1 \mapsto 6 \mapsto 4 \mapsto 1$. Since this completes a cycle we now start with 2 and get $2 \mapsto 5 \mapsto 2$. Similarly we have $3 \mapsto 3$, $7 \mapsto 7$, and $8 \mapsto 9 \mapsto 8$. The disjoint cycle form of σ is $(164)(25)(89)$. The order in which we write these cycles is immaterial because disjoint cycles commute, so $\sigma = (89)(164)(25)$ is also a disjoint cycle form, for example.

For some purposes it is useful to consider the variation of disjoint cycle form where 1-cycles are included. This is also unique if one insists that all numbers belong to some cycle. So in this case we would write the disjoint cycle form of σ as $(164)(25)(89)(3)(7)$. We call this disjoint cycle form *with* 1*-cycles*.

One advantage of disjoint cycle form is that when a permutation is written in this way its order in the group S_n may be calculated easily.

Lemma 5.5. *Let $\sigma \in S_n$ be a permutation with disjoint cycle form $\tau_1\tau_2 \dots \tau_k$ where each τ_i is a d_i-cycle. Then $|\sigma| = \operatorname{lcm}(d_1, \dots, d_k)$.*

Proof. First, it is easy to observe that the order in S_n of a d-cycle is d. Then since disjoint cycles commute we get $\sigma^m = \tau_1^m\tau_2^m \dots \tau_k^m$ for all $m \geq 1$. Now since the τ_i are pairwise disjoint permutations, so are the τ_i^m. It follows that $\sigma^m = 1$ if and only if $\tau_i^m = 1$ for all i. Now since $\tau_i^m = 1$ precisely when m is a multiple of the order d_i of τ_i, we get $|\sigma| = \operatorname{lcm}(d_1, \dots, d_k)$. □

Example 5.6. Suppose we want to find the smallest n such that S_n contains an element of order 12. Such a permutation σ would have disjoint cycle form $\tau_1 \dots \tau_k$ where τ_i is a d_i-cycle and $\operatorname{lcm}(d_1, \dots, d_k) = 12$. Observe that $(123)(4567) \in S_7$ has order $\operatorname{lcm}(3, 4) = 12$, while if $n \leq 6$ then it is impossible to find a set of integers that add to n and have least common multiple 12. Thus $n = 7$. More generally, if $m = p_1^{e_1} \dots p_k^{e_k}$ is the prime factorization of m, where the p_i are distinct primes and $e_i \geq 1$, one can prove that the smallest n such that S_n contains an element of order m is $n = p_1^{e_1} + \dots + p_k^{e_k}$.

5.2. **Conjugacy classes in S_n.** The disjoint cycle form of a permutation is also closely connected to its conjugacy class.

Definition 5.7. Given $\sigma \in S_n$, write $\sigma = \tau_1\tau_2 \dots \tau_k$ in disjoint cycle form with 1-cycles. The *cycle type* of σ is $1^{n_1}2^{n_2} \dots$ where there are n_d distinct d-cycles in the disjoint cycle form of σ. Since we include 1-cycles, note that $n = n_1 + 2n_2 + 3n_3 + \dots$. It is convenient to include 1-cycles so that it is clear which permutation group we are working in.

For example, the $\sigma \in S_9$ given in Example 5.1 has cycle type $1^2 \cdot 2^2 \cdot 3^1$.

Proposition 5.8. *Permutations $\sigma, \sigma' \in S_n$ are conjugate in S_n if and only if σ and σ' have the same cycle type. Thus each conjugacy class in S_n consists of all permutations of some cycle type.*

Proof. Let $\sigma, \tau \in S_n$. If $\sigma(i) = j$, then

$$\tau\sigma\tau^{-1}(\tau(i)) = \tau\sigma(i) = \tau(j).$$

This shows that if $\sigma = (a_1 a_2 \dots a_d)$ is some d-cycle, then $\sigma' = \tau\sigma\tau^{-1} = (\tau(a_i)\tau(a_2)\dots\tau(a_d))$ is also a d-cycle. Then if σ is written as a product of pairwise disjoint cycles, σ' will be a product of cycles of the same lengths, where each integer a is replaced by $\tau(a)$ throughout. So any conjugate $\sigma' = \tau\sigma\tau^{-1}$ of σ has the same cycle type as σ.

Conversely, if σ and σ' are two permutations with the same cycle type, we can pair up each cycle in σ with some cycle of the same length in σ', so that the pairing is one-to-one. Then clearly there is a permutation τ so that for each cycle $(a_1 a_2 \dots a_d)$ in σ, $(\tau(a_1)\tau(a_2)\dots\tau(a_d))$ is the paired cycle in σ'. Then by the calculation above, $\sigma' = \tau\sigma\tau^{-1}$ is a conjugate of σ. □

Example 5.9. Suppose that $\sigma = (135)(246)(78)(9) \in S_9$ and $\sigma' = (1)(568)(39)(247)$. Then σ and σ' are conjugate in S_9 by the proposition, since both have cycle type $1^1 \cdot 2^1 \cdot 3^2$. Note that there are multiple choices of τ such that $\tau\sigma\tau^{-1} = \sigma'$, depending on how we pair the cycles and also how we write the cycles. One choice in this case is to pair $(135) \to (247)$, $(246) \to (568)$, $(78) \to (39)$ and $(9) \to (1)$. Then $\tau = (125734689)$ will give $\tau\sigma\tau^{-1} = \sigma'$. Another possible pairing is $(135) \to (685)$ (since (685) is another notation for (568)), $(246) \to (247)$, $(78) \to (93)$ and $(9) \to (1)$. Then the corresponding is $\tau = (1679)(38)(2)(4)(5)$ which also satisfies $\tau\sigma\tau^{-1} = \sigma'$.

Note that a choice of cycle type of permutation in S_n is the same as a choice of decomposition of n as a sum of positive integers (the cycle lengths) with repeats allowed. This is called a *partition* of n. For example, if $n = 5$ then the possible partitions are $1+1+1+1+1$, $2+1+1+1$, $3+1+1$, $4+1$, 5, $2+2+1$, and $2+3$. The number of partitions of n is a function $p(n)$ well-studied in combinatorics. By Proposition 5.8, $p(n)$ is the number of conjugacy classes in S_n.

Example 5.10. It is not hard to count the number of elements in a conjugacy class in S_n. For example, let us consider a permutation σ of cycle type $1^1 \cdot 3^2$ in S_7. A permutation of this type has the form $(abc)(def)(g)$ where the numbers a—g are all different. Considering the cycle shape as fixed, there are 7! ways of writing the numbers 1 through 7 inside the parentheses. However, since each 3-cycle can be written 3 ways, we have to divide by (3)(3). In addition, switching the order

in which the two 3-cycles are listed does not change the permutation, and so we have to divide by 2. Thus $|\operatorname{Cl}(\sigma)| = 7!/(18) = 280$.

We also know that $|\operatorname{Cl}(\sigma)| = |S_n|/|C_{S_n}(\sigma)|$ which implies that $C_{S_n}(\sigma) = 18$. For instance let $\sigma = (123)(456) \in S_7$, which has the particular cycle type we are studying. The permutation σ obviously commutes with (123) and (456). In addition, if $\tau = (14)(25)(36)$, then by our formula for conjugating permutations we get $\tau\sigma\tau^{-1} = (456)(123) = \sigma$. So $\langle (123), (456), (14)(25)(36) \rangle \subseteq C_{S_n}(\sigma)$. One can check that these three elements do generate a subgroup of order 18, so in fact $C_{S_n}(\sigma) = \langle (123), (456), (14)(25)(36) \rangle$.

5.3. **The alternating group A_n.** Let $\sigma = (a_1 a_2 \dots a_d)$ be an d-cycle in S_n. Then an easy calculation shows that $\sigma = (a_1 a_2)(a_2 a_3) \dots (a_{d-1} a_d)$. Then since every $\sigma \in S_n$ is a product of (pairwise disjoint) cycles, σ can be written as a product of (generally non-disjoint) transpositions.

In general there are many different ways to write a permutation as a product of transpositions. For example, $(1234) = (12)(23)(34) = (34)(24)(14) = (34)(24)(13)(34)(13)$. However, what cannot change is the *parity* of the number of transpositions involved. So, for example, (1234) could never be expressed as a product of an even number of transpositions.

Theorem 5.11. *If $\sigma \in S_n$ satisfies $\sigma = \tau_1 \tau_2 \dots \tau_m$ and $\sigma = \rho_1 \rho_2 \dots \rho_k$ where all τ_i and ρ_i are transpositions, then either m and k are both even or m and k are both odd.*

There are many different proofs of this theorem; we will omit the proof here.

Definition 5.12. For each $n \geq 2$, The *alternating group* is the subset A_n of S_n consisting of those permutations that are equal to a product of an even number of transpositions. We call the permutations in A_n *even*. The permutations that are equal to a product of an odd number of transpositions (i.e. those in $S_n - A_n$) are called *odd*.

Lemma 5.13. *Let $n \geq 2$. Then $A_n \trianglelefteq S_n$ and $|S_n : A_n| = 2$, so that $|A_n| = n!/2$.*

Proof. Suppose that $\sigma = \tau_1 \tau_2 \dots \tau_m$ and $\sigma' = \rho_1 \rho_2 \dots \rho_k$ where the τ_i and ρ_k are transpositions, and m and k are even so that $\sigma, \sigma' \in A_n$. Then $\sigma\sigma' = \tau_1 \tau_2 \tau_m \rho_1 \rho_2 \dots \rho_k$ is a product of $m + k$ transpositions and thus $\sigma\sigma' \in A_n$. In addition, $\sigma^{-1} = \tau_m^{-1} \tau_{m-1}^{-1} \dots \tau_1^{-1} = \tau_m \tau_{m-1} \dots \tau_1$ is a product of m transpositions since a transposition is its own inverse. Thus $\sigma^{-1} \in A_n$. We see that A_n is a subgroup of S_n.

Next, note that every permutation in the coset $(12)A_n$ is odd. Conversely, if σ is odd, then $(12)\sigma$ is even, so $(12)\sigma \in A_n$; then $\sigma = (12)(12)\sigma \in (12)A_n$. We conclude that $(12)A_n$ consists precisely

of all of the odd permutations. Since every permutation is even or odd, we have $S_n = A_n \bigcup (12)A_n$ is a (disjoint) union of two cosets of A_n, forcing $|S_n : A_n| = 2$. Since $|S_n| = n!$, we get $|A_n| = n!/2$. Finally, since A_n has index 2 in S_n, A_n is automatically normal in S_n, i.e. $A_n \trianglelefteq S_n$. □

5.4. **Using A_n to produce normal subgroups of index 2.** Suppose that a group G gives a left action on a set X of size n. We have seen that this corresponds to a homomorphism of groups $\phi : G \to S_n$. We have now constructed a normal subgroup A_n of S_n of index 2. Suppose that the subgroup $\phi(G)$ of S_n is not contained in A_n. Then $\phi(G)A_n$ is a subgroup of S_n which is strictly larger than A_n and this forces $\phi(G)A_n = S_n$ by Lagrange's theorem. We then have $S_n/A_n = \phi(G)A_n/A_n \cong \phi(G)/(A_n \cap \phi(G))$ by the second isomorphism theorem. So $A_n \cap \phi(G) \trianglelefteq \phi(G)$ with $|\phi(G) : A_n \cap \phi(G)| = 2$. By subgroup correspondence, taking the inverse image we see that $\phi^{-1}(A_n) \trianglelefteq G$ with $|G : \phi^{-1}(A_n)| = 2$.

This method gives a way of finding normal subgroups of index 2 inside a group G in some cases. One just has to produce a homomorphism $\phi : G \to S_n$ for some symmetric group S_n, such that the image of ϕ is not contained in A_n. Here is an interesting application.

Proposition 5.14. *Suppose that G is a group with $|G| = 2m$ for some odd integer m. Then there is $H \trianglelefteq G$ with $|G : H| = 2$. Moreover, H is the unique subgroup of index 2 in G, and so H $\operatorname{char} G$.*

Proof. This is a rare instance in which one gets useful information from the left multiplication action. So let G act on itself by left multiplication, $g \cdot x = gx$. This gives a homomorphism of groups $\phi : G \to \mathrm{Sym}(G)$. Here, since $|G| = 2m$ we have $\mathrm{Sym}(G) \cong S_{2m}$. Now suppose that $g \in G$ is an element of order d. Then $\{1, g, g^2, \ldots, g^{d-1}\}$ are d distinct elements of g, so that for any $x \in G$, the elements $\{x, gx, g^2x, \ldots, g^{d-1}x\}$ are also distinct. Moreover, since the action of g on the left satisfies $g \cdot g^i x = g^{i+1}x$ for $0 \leq i \leq d-2$ and $g \cdot g^{d-1}x = g^d x = 1x = x$, we see that g permutes these d elements in a d-cycle. It follows that every element of G is permuted under the action of g in some d-cycle, so that the disjoint cycle form of $\phi(g)$ must be a product of pairwise disjoint d-cycles, necessarily $(2m)/d$ of them.

Suppose that d is even. Then $(2m)/d$ is a divisor of m and hence is odd. Moreover, a d-cycle is a product of $(d-1)$-transpositions and is thus an odd permutation. The disjoint cycle form of $\phi(g)$ thus is a product of an odd number of odd permutations and so is odd in S_{2m}. On the other hand, if d is odd, then $\phi(g)$ is a product of d-cycles, which are even, so $\phi(g)$ is even in S_{2m}.

Now let $H = \phi^{-1}(A_n) \leq G$. The group G does contain elements of even order; for example, by Cauchy's theorem G must have an element of order 2. Thus $\phi(G) \not\subseteq A_n$. As we saw in the comments before the proposition, we get from this that $H \trianglelefteq G$, $|G : H| = 2$, and $|H| = m$. This

shows that H exists. Moreover, from the previous paragraph we see that H consists precisely of the elements in G that have odd order. Suppose that H' is another subgroup of G with $|G : H'| = 2$. Then $|H'| = m$ is odd. Thus every element of H' must have order a divisor of m, which will be odd. Since H' consists of elements of odd order, $H' \subseteq H$. But then $H' = H$ since $|H'| = |H| = m$.

Finally, if $\rho \in \mathrm{Aut}(G)$, then $\rho(H)$ is also a subgroup of order m. So $\rho(H) = H$ and thus $H \operatorname{char} G$. □

5.5. A_n **is simple for** $n \geq 5$**.** Above we have completely understood the structure of the conjugacy classes in S_n. The conjugacy classes in A_n are closely related to those of S_n. Let us restrict the action of S_n on itself by conjugation to the action of A_n on S_n by conjugation. Of course in this case the orbits may be different in general. If $\sigma \in S_n$, its orbit $\mathcal{O}_\sigma$ under the A_n-action has size $|A_n|/|C_{A_n}(\sigma)|$ by the orbit stabilizer theorem. (Note we are not assuming $\sigma \in A_n$ here, but the notation $C_{A_n}(\sigma) = \{\tau \in A_n | \tau\sigma\tau^{-1} = \sigma\}$ still makes sense.) In addition, its S_n-orbit $\mathrm{Cl}_{S_n}(\sigma)$ has size $|S_n|/|C_{S_n}(\sigma)|$. We also have $C_{A_n}(\sigma) = C_{S_n}(\sigma) \cap A_n$ by definition. Using the 2nd isomorphism theorem, $C_{S_n}(\sigma)/(C_{S_n}(\sigma) \cap A_n) \cong (C_{S_n}(\sigma)A_n)/A_n$.

Now since $|S_n : A_n| = 2$, either $C_{S_n}(\sigma)A_n = A_n$ or else $C_{S_n}(\sigma)A_n = S_n$. In the first case we obtain $C_{S_n}(\sigma) \subseteq A_n$ and so $C_{S_n}(\sigma) = C_{A_n}(\sigma)$. Then the A_n-orbit of σ has size $|\mathcal{O}_\sigma| = |\mathrm{Cl}_{S_n}(\sigma)|/2$ by the calculations above. If this happens, because $\mathrm{Cl}_{S_n}(\sigma)$ is a union of A_n-orbits, the only possibility is that $\mathrm{Cl}_{S_n}(\sigma)$ is breaking up as a union of two A_n-orbits of equal size. Alternatively, if $C_{S_n}(\sigma)A_n = S_n$ the numerics above force $|C_{S_n}(\sigma) : C_{A_n}(\sigma)| = 2$ and $|\mathcal{O}_\sigma| = |\mathrm{Cl}_{S_n}(\sigma)|$, so that $\mathcal{O}_\sigma = \mathrm{Cl}_{S_n}(\sigma)$.

We conclude that every conjugacy class of S_n is either also an orbit of the action of A_n, or else breaks up as a union of two A_n-orbits of equal size. Now apply this to $\sigma \in A_n$. The orbit under A_n in this case is $\mathrm{Cl}_{A_n}(\sigma)$. We get that the conjugacy class of $\sigma \in A_n$ is either equal to its conjugacy class in S_n, or else contains half of the elements of its conjugacy class in S_n. Moreover, one can completely characterize which case happens for a given conjugacy class. We state the precise result here for completeness, but leave the proof to the reader as an exercise.

Lemma 5.15. *Let $\sigma \in S_n$ and suppose and consider $\mathcal{K} = \mathrm{Cl}_{S_n}(\sigma)$, the conjugacy class of σ. Restrict the action of S_n on itself by conjugation to the action of the subgroup A_n. Then either (i) $\mathcal{K}$ is also an A_n-orbit, or else (ii) $\mathcal{K}$ is the disjoint union of two A_n-orbits of equal size. Case (ii) occurs if and only if $C_{S_n}(\sigma) = C_{A_n}(\sigma)$, if and only the disjoint cycle type (with 1-cycles) of σ is of the form $n_1^1 n_2^1 n_3^1 \dots n_k^1$ for some distinct odd integers $n_1, \dots, n_k$.*

In words, for the conjugacy class of σ to split into two A_n-orbits, σ should be a product of cycles with distinct odd lengths when written in disjoint cycle form. 1-cycles must be included for this result to be correct.

Example 5.16. Consider conjugacy classes in A_5. If $\sigma = (123)$, writing it with 1-cycles as $(123)(4)(5)$ we see that its cycle type is $1^2 3^1$. Thus it is not of the special form in which case (ii) occurs in the lemma above and so we have case (i): $\mathrm{Cl}_{A_n}(\sigma) = \mathrm{Cl}_{S_n}(\sigma)$, which is the set of all 3-cycles in S_n, of which there are $(5)(4)(3)/3 = 20$. Similarly, if $\sigma = (12)(34)$ then its conjugacy class in A_n is the full class of all products of 2-disjoint 2-cycles in S_n; there are $5!/(2)(2)(2) = 15$ of these.

However, if $\sigma = (12345)$ then this has cycle type 5^1 and so Cl_{S_n}, which has $5!/5 = 24$ members, splits into two conjugacy classes in A_n each of size 12. It is easy to check that the complement of $\mathrm{Cl}_{A_n}((12345))$ in $\mathrm{Cl}_{S_n}((12345))$ is $\mathrm{Cl}_{A_n}((12354))$; in other words (12345) and (12354) are conjugate in S_n but not conjugate in A_n.

The analysis above completely determines the sizes of conjugacy classes in A_n. Including the trivial conjugacy class $\{1\}$, the order 60 group A_5 breaks up into conjugacy classes of size $1, 12, 12, 15,$ and 20.

Recall that a group G is *simple* if the only normal subgroups of G are the trivial subgroup $\{1\}$ and G itself. Based on our analysis of conjugacy classes in A_5, there is an easy proof that A_5 is simple.

Proposition 5.17. *A_5 is a simple group.*

Proof. Suppose that $N \trianglelefteq A_5$. If $x \in N$, then $gxg^{-1} \subseteq gNg^{-1} = N$ for all $g \in A_5$. This shows that $\mathrm{Cl}(x) \subseteq N$. As a consequence, N must be a disjoint union of conjugacy classes of A_5. On the other hand, by Lagrange's Theorem, $|N|$ is a divisor of $|A_5| = 60$.

The conjugacy classes of A_5 have sizes $1, 12, 12, 15,$ and 20. Obviously N contains the class $\{1\}$ of size 1. An easy check shows that there is no possible way to take some of these numbers, including 1, which sum to a proper divisor d of 60 with $1 < d < 60$. So either $N = \{1\}$ or $N = A_5$. □

Consider the alternating groups A_5 for $n < 5$. $A_1 = A_2 = \{1\}$, which is boring, and $A_3 = \{1, (123), (132)\}$ is cyclic of order 3. These groups are simple. On the other hand, let us see now that A_4 is not simple. Let $V = \{\{1\}, (12)(34), (13)(24), (14)(23)\} \subseteq A_4$. A quick calculation shows that V is a subgroup of A_4. Because V contains all of the possible permutations in S_4 of cycle type 2^2, V is a union of conjugacy classes of S_4. Thus $V \trianglelefteq S_4$ and so $V \trianglelefteq A_4$ also. The letter V is

traditional for this subgroup; V stands for "vier", the German word for 4. Since V is a group of order 4 whose elements all have order 2, by our classification of groups of order p^2 we must have $V \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. This is also easy to check directly.

We now show that $n = 4$ is the only outlier.

Theorem 5.18. *Let $n \geq 5$. Then A_n is a simple group.*

Proof. The proof goes by induction on n with $n = 5$ as the base case, which we handled in Proposition 5.17. Consider now $n > 5$ and assume that A_{n-1} is simple. Consider the natural action of A_n on $\{1, 2, \ldots, n\}$. It is easy to see that this is a transitive action; given $i, j \in \{1, 2, \ldots n\}$ with $i \neq j$, if we pick a third number k different from i and j then the 3-cycle $(ijk) \in A_n$ sends i to j. Consider $H_i = (A_n)_i$, the stabilizer subgroup of $i \in \{1, 2, \ldots, n\}$. This is the set of even permutations which fix i. This is the same as the set of even permutations of the set $\{1, 2, \ldots, i-1, i+1, \ldots, n\}$, which can be identified with A_{n-1}. Thus each stabilizer subgroup H_i is isomorphic to A_{n-1}. In addition, because the action is transitive, if $\sigma \in A_n$ is such that $\sigma(i) = j$ then $\sigma H_i \sigma^{-1} = H_j$ by Theorem 3.13(2). So all of these stabilizer subgroups are conjugate.

Let $N \trianglelefteq A_n$. We now consider two cases. First, suppose that $N \cap H_i \neq \{1\}$. Now $N \cap H_i \trianglelefteq H_i$, and since $H_i \cong A_{n-1}$, it is a simple group by the induction hypothesis. So the only conclusion in this case is $N \cap H_i = H_i$. But then choosing $\sigma \in A_n$ such that $\sigma(i) = j$, we have $H_j = \sigma H_i \sigma^{-1} \subseteq \sigma N \sigma^{-1} = N$. Thus N contains H_j for all j, and so N contains the subgroup generated by all of the H_j. However, any product of two 2-cycles involves at most 4 numbers and so fixes some number and is contained in some H_j. It follows that N contains all products of two 2-cycles, and hence $N = A_n$.

The other case is where $N \cap H_i = \{1\}$ for all i. It could be that $N = \{1\}$, in which case we are done, so suppose not. Pick $1 \neq \sigma \in N$. We claim that we can find $\tau \in A_n$ so that $1 \neq \sigma^{-1}\tau\sigma\tau^{-1} \in H_i$ for some i. If we do this, then since N is normal we see that $\sigma^{-1}(\tau\sigma\tau^{-1}) \in N$ and so $N \cap H_i \neq \{1\}$, and we get a contradiction. To prove the claim, by relabeling the integers and moving the largest cycle to the front, we can assume without loss of generality that the disjoint cycle form of σ either begins $(12)(34)\ldots$ or $(123\ldots d)\ldots$ for some $d \geq 3$. Taking $\tau = (345) \in A_n$, since τ fixes 1 and 2, one easily sees that $\sigma^{-1}\tau\sigma\tau^{-1} \in H_1$. To see that $\sigma^{-1}\tau\sigma\tau^{-1} \neq 1$, from our formula for conjugation we get that $\tau\sigma\tau^{-1}$ begins $(12)(45)\ldots$ or $(124\ldots)\ldots$, respectively. In either case this is not the same as σ, so $\sigma \neq \tau\sigma\tau^{-1}$, or $\sigma^{-1}\tau\sigma\tau^{-1} \neq 1$, verifying the claim. □

As already mentioned, classifying the finite simple groups up to isomorphism was one of the major projects in algebra in the latter half of the 20th century. This was announced as complete in the 1980's, though there is still ongoing work to streamline and explain the very technical proof, which

is spread over the publications of many mathematicians. The abelian simple groups are simply the cyclic groups of prime order p, so only the nonabelian case is interesting. The classification of nonabelian simple groups involves a number of infinite families of simple groups, of which the groups $\{A_n | n \geq 5\}$ are the easiest to handle. Some other infinite families arise naturally from matrix groups over finite fields. After the infinite families there are a small number of exceptional simple groups that don't belong to any family; these 26 groups are called the *sporadic* simple groups. The largest sporadic group is the *Fisher-Griess Monster*, named for its enormous size; it has approximately 8×10^{53} elements. Still, the largest prime factor q dividing the order of the monster group is 71, which is also the largest prime factor of the order of any of the sporadic groups. So even the largest of the sporadic groups tend to have orders which are products of many small primes.

One example of a family of simple groups coming from matrices are the *projective special linear groups*. Recall that for any field F, we have the general linear group $\mathrm{GL}_n(F)$ of $n \times n$ matrices with entries from F. This can't be simple because it always has the special linear group $\mathrm{SL}_n(F)$ of matrices with determinant 1, where $\mathrm{SL}_n(F) \trianglelefteq \mathrm{GL}_n(F)$. It also has a nontrivial center $Z = \{\lambda I | \lambda \in F^\times\}$ consisting of nonzero scalar multiples of the identity, and $Z \trianglelefteq \mathrm{GL}_n(F)$. Then $SZ = Z \cap \mathrm{SL}_n(F) \trianglelefteq \mathrm{SL}_n(F)$ and so $\mathrm{SL}_n(F)$ can't be simple either. One then defines the projective special linear group to be $\mathrm{PSL}_n(F) = \mathrm{SL}_n(F)/SZ$. Its name comes from the fact that it has a natural action on a projective space, rather than the Euclidean space F^n on which $\mathrm{SL}_n(F)$ usually acts.

The groups $\mathrm{PSL}_n(F)$ for $n \geq 2$ are simple except in a few exceptional small cases (similar to how A_n only becomes simple for $n \geq 5$). Namely, $\mathrm{PSL}_n(F)$ is simple if $n \geq 3$ for any F, and $\mathrm{PSL}_2(F)$ is simple as long as F has at least 4 elements. In particular, by taking F to be a field with finitely many elements, we get an infinite family of finite simple groups in this way.

We will study finite fields in detail later in the course. For each prime q there is a unique field with q elements, namely the ring $\mathbb{Z}_q$ of integers modulo q with the standard addition and multiplication of congruence classes. Then by the result above, $\mathrm{PSL}_2(\mathbb{Z}_q)$ is a finite simple group as long as $q \geq 5$. One may see that $\mathrm{PSL}_2(\mathbb{Z}_5)$ is isomorphic to A_5. However, $\mathrm{PSL}_2(\mathbb{Z}_7)$ is a new simple group of order 168. This is the next smallest possible order of a non-Abelian simple group after 60. Interestingly, $\mathrm{PSL}_3(\mathbb{Z}_2)$ also turns out to have 168 elements and it is isomorphic to $\mathrm{PSL}_2(\mathbb{Z}_7)$.

The reader can see Rotman's book, "An introduction to the theory of groups", for the proof that the projective special linear groups are simple. Rotman also gives an introduction to the Mathieu groups, which are some of the sporadic simple groups that arise as automorphism groups of very special combinatorial objects called Steiner systems.

6. Direct and semidirect products

6.1. **External and internal direct products.** In an earlier section we briefly recalled the definition of the direct product of two groups G and H. This is the easiest way to stick two groups together to form a new group. There is no reason to restrict this to two groups. If $H_1, \ldots, H_k$ are finite groups, with no assumed relationship to each other, we define $H_1 \times H_2 \times \cdots \times H_k$ to be the cartesian product of sets, $\{(h_1, h_2, \ldots, h_k) | h_i \in H_i\}$, with the product

$$(h_1, h_2, \ldots, h_k)(h_1', h_2', \ldots, h_k') = (h_1 h_1', \ldots, h_k h_k'),$$

where the product in the ith coordinate is done in the group H_i. It is easy to check that this is a group, with identity element $1 = (1, 1, \ldots, 1)$ and $(h_1, h_2, \ldots, h_k)^{-1} = (h_1^{-1}, h_2^{-1}, \ldots, h_k^{-1})$. This group is called the *external direct product* of the groups $H_1, H_2, \ldots, H_k$.

Because the operations in the direct product are done separately in each coordinate with no interaction, most of the basic properties of the direct product follow immediately from the properties of the individual groups. For example, if all H_i are finite then $|G| = |H_1||H_2| \ldots |H_k|$, since this is true of the cartesian product of sets. If $(h_1, \ldots, h_k) \in H_1 \times \cdots \times H_k$, then $(h_1, \ldots, h_k)^n = (h_1^n, \ldots, h_k^n)$, which immediately implies that $|(h_1, \ldots, h_k)| = \operatorname{lcm}(|h_1|, \ldots, |h_k|)$ if all the $|h_i|$ are finite.

For each i, the group $G = H_1 \times \cdots \times H_k$ has a subgroup

$$\overline{H_i} = \{(1, 1, \ldots, 1, \overset{i}{h}, 1, \ldots, 1) | h \in H_i\}$$

which is clearly isomorphic to H_i as a group. A quick calculation shows that $\overline{H_i} \trianglelefteq G$ for all i. Note that we have

$$\overline{H_1}\,\overline{H_2} \ldots \overline{H_{i-1}}\,\overline{H_{i+1}} \ldots \overline{H_k} = \{(h_1, h_2, \ldots, h_{i-1}, 1, h_{i+1}, \ldots, h_k) | h_i \in H_i\}$$

and so $\overline{H_i} \cap \overline{H_1}\,\overline{H_2} \ldots \overline{H_{i-1}}\,\overline{H_{i+1}} \ldots \overline{H_k} = \{1\}$. A similar calculation shows that $\overline{H_1}\,\overline{H_2} \ldots \overline{H_k} = G$.

We abstract the properties that the subgroups $\overline{H_i}$ satisfy in the following definition.

Definition 6.1. Let G be a group with subgroups $H_1, H_2, \ldots, H_k$. We say that G is the *internal direct product* of the subgroups $H_1, H_2, \ldots, H_k$ if

(i) $H_i \trianglelefteq G$ for all $1 \leq i \leq k$;

(ii) $H_1 H_2 \ldots H_k = G$; and

(iii) $H_i \cap H_2 \ldots H_{i-1} H_{i+1} \ldots H_k = \{1\}$ for all $1 \leq i \leq k$.

The comments made before the definition show that the external direct product $H_1 \times \cdots \times H_k$ is the internal direct product of the subgroups $\overline{H_1}, \ldots, \overline{H_k}$. We now prove a kind of converse.

Theorem 6.2. *Suppose that G is the internal direct product of the subgroups $H_1, H_2, \dots H_k$. Then $G \cong H_1 \times H_2 \times \cdots \times H_k$.*

Proof. Define a function $\phi : H_1 \times H_2 \times \cdots \times H_k \to G$ by $\phi((h_1, h_2, \dots, h_k)) = h_1 h_2 \dots h_k$. Since $H_1 H_2 \dots H_k = G$ by property (ii), the function ϕ is surjective.

Property (iii) implies in particular that $H_i \cap H_j = \{1\}$ for any $i \neq j$. Now for $h_i \in H_i, h_j \in H_j$, we have $(h_j^{-1} h_i^{-1} h_j) h_i = h_j^{-1}(h_i^{-1} h_j h_i) \in H_i \cap H_j = \{1\}$, and so $h_i h_j = h_j h_i$. Using this, we get

$$\phi((g_1, \dots, g_k)(h_1, \dots, h_k)) = \phi((g_1 h_1, \dots, g_k h_k)) = g_1 h_1 g_2 h_2 \dots g_k h_k = g_1 g_2 \dots g_k h_1 h_2 \dots h_k$$
$$= \phi((g_1, \dots, g_k))\phi((h_1, \dots h_k))$$

because $h_i g_j = g_j h_i$ whenever $i \neq j$. Thus ϕ is a homomorphism of groups. Finally, suppose that $(h_1, \dots, h_k) \in \ker \phi$, so $h_1 h_2 \dots h_k = 1$. Since h_i commutes with h_j for all $i \neq j$, we have $h_i h_1 h_2 \dots h_{i-1} h_{i+1} \dots h_k = 1$ and thus by property (iii),

$$h_i^{-1} = h_1 h_2 \dots h_{i-1} h_{i+1} \dots h_k \in H_i \cap H_2 \dots H_{i-1} H_{i+1} \dots H_k = \{1\}.$$

This implies $h_i = 1$. Since i was arbitrary, $h_i = 1$ for all i and so $(h_1, h_2, \dots, h_k) = 1$. Hence ϕ is injective and ϕ is the desired isomorphism of groups. □

From now on, when we have an external direct product $H_1 \times \cdots \times H_k$ of groups, we identify H_i with the subgroup $\overline{H_i}$ defined earlier, and so we can think of $H_1 \times \cdots \times H_k$ as the internal direct product of the subgroups H_i. Conversely, we just showed that an internal direct product is isomorphic to an external direct product in a canonical way. This shows that the difference between internal and external direct products is mostly a point of view, and mathematicians tend not to distinguish carefully between them.

Let us give some applications.

Proposition 6.3. *Let G be a finite group with normal subgroups $H_1, \dots, H_k$ such that $|G| = |H_1||H_2| \dots |H_k|$ and $\gcd(|H_i|, |H_j) = 1$ for all $i \neq j$. Then G is an internal direct product of the subgroups $H_1, \dots, H_k$ and so $G \cong H_1 \times H_2 \times \cdots \times H_k$.*

Proof. We have $H_i \trianglelefteq G$ by assumption. We know that if H and K are normal subgroups of G, then HK is a subgroup of G with $|HK| = |H||K|/|(H \cap K)|$. In particular $|HK|$ divides $|H||K|$. This result extends by induction to any finite number of normal subgroups, so we get $|H_1 H_2 \dots H_{i-1} H_{i+1} \dots H_k|$ divides $|H_1||H_2| \dots |H_{i-1}||H_{i+1}| \dots |H_k|$ for any i. Now since $|H_i|$ and $|H_j|$ are relatively prime for all $j \neq i$, we get that $|H_i|$ is also relatively prime to the product

$|H_1||H_2|\dots|H_{i-1}||H_{i+1}|\dots|H_k|$. It follows that the order $|H_i \cap H_1H_2\dots H_{i-1}H_{i+1}\dots H_k|$ divides $\gcd(|H_i|, |H_1||H_2|\dots|H_{i-1}||H_{i+1}|\dots|H_k|) = 1$, so $H_i \cap H_1H_2\dots H_{i-1}H_{i+1}\dots H_k = \{1\}$.

Now let $K = H_1H_2\dots H_k$. Since $H_i \trianglelefteq K$ for all i, we have checked all of the conditions needed to conclude that K is an internal direct product of $H_1, H_2, \dots, H_k$. In particular, we have $K \cong H_1 \times H_2 \times \cdots \times H_k$. But this means that $|K| = |H_1||H_2|\dots|H_k| = |G|$, so necessarily $K = G$. □

Corollary 6.4. *Let G be a finite group of order $p_1^{e_1}\dots p_k^{e_k}$ for some distinct primes p_i and $e_i \geq 1$. Suppose that for each i, G has a normal Sylow p-subgroup P_i. Then G is the internal direct product of $P_1, \dots, P_k$, and so $G \cong P_1 \times \cdots \times P_k$.*

Proof. This is immediate from the proposition, using that $|P_i| = p_i^{e_i}$ and that $\gcd(p_i^{e_i}, p_j^{e_j}) = 1$ for $i \neq j$. □

Example 6.5. Let $n = p_1^{e_1}\dots p_k^{e_k}$ for distinct primes p_i and integers $e_i \geq 1$. Consider $G = \mathbb{Z}_n$ under addition, a which is cyclic of order n, and write $\overline{a} = a + n\mathbb{Z} \in G$. For each i define $q_i = n/(p_i^{e_i})$. Then $H_i = \langle \overline{q_i} \rangle$ is the unique subgroup of $\mathbb{Z}_n$ with order $p_i^{e_i}$. We know that H_i is also cyclic, so $H_i \cong \mathbb{Z}_{p_i^{e_i}}$. By Proposition 6.3 (or Corollary 6.4), G is the internal direct product of the H_i and so

$$G = \mathbb{Z}_n \cong H_1 \times \cdots \times H_k \cong \mathbb{Z}_{p_1^{e_1}} \times \cdots \times \mathbb{Z}_{p_k^{e_k}}.$$

Example 6.6. Suppose that $|G| = pq$ for distinct primes p and q with $p < q$ Let P be a sylow p-subgroup and Q a Sylow q-subgroup. We saw earlier that $Q \trianglelefteq G$. If $P \trianglelefteq G$ also (which is always the case if p does not divide $q - 1$), then by Corollary 6.4 and Example 6.5, we immediately get $G \cong P \times Q \cong \mathbb{Z}_p \times \mathbb{Z}_q \cong \mathbb{Z}_{pq}$ is cyclic, receovering the claims in Example 4.9.

There is also no particular reason to restrict the definition of a direct product to finitely many groups; we focused on that case above because our main interest in this course in finite groups. Here is the general definition.

Definition 6.7. Let $\{H_\alpha\}_{\alpha \in I}$ be any indexed collection of groups. The direct product of these groups is defined to be the cartesian product of sets,

$$\prod_{\alpha \in I} H_\alpha = \{(h_\alpha) | h_\alpha \in H_\alpha\},$$

with the coordinatewise operation $(g_\alpha)(h_\alpha) = (g_\alpha h_\alpha)$.

Of course the direct product. Note that an element of $\Pi_{\alpha \in I} H_\alpha$ is an I-tuple: a list of elements indexed by $\alpha \in I$, where the element in the α-coordinate belongs to H_α. We usually just write an I-tuple as (h_α), though $(h_\alpha)_{\alpha \in I}$ would be more formally correct.

We can use infinite direct products to construct some interesting examples.

Example 6.8. Let H_i be a cyclic group of order n_i for all $i \geq 1$. Consider the direct product $G = \prod_{i \geq 1} H_i$. Clearly G is an infinite group.

If $n_i = m$ for some fixed m and all $i \geq 1$, then G is an infinite group such that every $g \in G$ has finite order dividing m.

If $n_i = i$ for all $i \geq 1$, then G is an infinite group with elements of all possible finite orders. If $H_i = \langle a_i \rangle$ then $(a_1, a_2, a_3, \dots) \in G$ has infinite order, so G has infinite order elements as well.

There is another way to join a collection of groups together which is different when the collection is infinite.

Definition 6.9. Let $\{H_\alpha\}_{\alpha \in I}$ be any indexed collection of groups. The *restricted product* of these groups is the subset of the direct product $\prod_{\alpha \in I} H_\alpha$ consisting of those elements which are the identity element in all but finitely many coordinates:

$$\prod_{\alpha \in I}^{\text{restr}} H_\alpha = \{(h_\alpha) | h_\alpha \in H_\alpha, h_\alpha = 1 \text{ for all } \alpha \in I - X, \text{ for some finite subset } X.\}$$

We have chosen an ad-hoc notation, as there does not seem to be any standard notation for the restricted product in this generality. It is easy to check that $\prod_{\alpha \in I}^{\text{restr}} H_\alpha \trianglelefteq \prod_{\alpha \in I} H_\alpha$.

Example 6.10. Again let H_i be cyclic of order n_i for $i \geq 1$. Let $G = \prod_{i \geq 1}^{\text{restr}} H_i$.

Let p be prime and let $n_i = p^i$ for all i. Then for each $i \geq 0$, G has an element of order p^i. Moreover, G is an infinite group which is a p-group, i.e. every element of G has finite order equal to a power of p.

If $n_i = i$ for all $i \geq 1$, then G is an infinite group with elements of all possible finite orders. Unlike the case of the full direct product, however, in this case all elements of G have finite order.

The restricted product comes up primarily in the context of abelian groups. If $\{H_\alpha\}_{\alpha \in I}$ is a collection of abelian groups, the restricted product of the H_α is usually called the *direct sum* and is notated $\bigoplus_{\alpha \in I} H_\alpha$. This is a special case of the notion of a direct sum of modules which we will define later.

6.2. **Semidirect products.** Suppose we have a group G with normal subgroups H and K. In this case G is an internal direct product of H and K if and only if $HK = G$ and $H \cap K = \{1\}$. Thus

under these conditions we get $G \cong H \times K$ by Theorem 6.2. As part of the proof of that theorem, we showed (using that H and K are normal and $H \cap K = \{1\}$) that $hk = kh$ for all $h \in H$, $k \in K$.

It is much more common for a group to have a pair of subgroups intersecting trivially in which only *one* of them is normal. In this section we aim to analyze how we can understand the structure of the group in that case. We will see that we will be able to show that G is isomorphic to a kind of "twisted" version of a direct product.

So we now consider the setup where $H \trianglelefteq G$, $K \leq G$, $HK = G$, and $H \cap K = \{1\}$. We think about the proof of Theorem 6.2 and what goes wrong with the proof in this case. We can still define a function $\psi : H \times K \to HK$ by the formula $\psi((h,k)) = hk$. Because $HK = G$, ψ is still surjective as a function. However, ψ will no longer be a homomorphism of groups in general, because H and K will not necessarily commute with each other. Injectivity, though, is fine: if $\psi((h_1,k_1)) = \psi((h_2,k_2))$, then $h_1k_1 = h_2k_2$ and so $h_2^{-1}h_1 = k_2k_1^{-1} \in H \cap K = \{1\}$, so that $h_1 = h_2$ and $k_1 = k_2$. (Note that since we don't know that ψ is a homomorphism, we couldn't check injectivity just by looking at which elements map to 1.)

We can understand the failure of elements of H and K to commute, and the failure of ψ to be a homomorphism, quite specifically. Let $h \in H$ and $k \in K$. Since H is normal, ${}^kh = khk^{-1} \in H$. This means if we have the product kh, we can "move the k to the right of the h" at the expense of applying a conjugation to h:

$$kh = khk^{-1}k = ({}^kh)k.$$

In this process k stays the same, but we think of it acting on h (by conjugation) as it moves past to the right. Then if we have $(h_1,k_1) \in H \times K$ and $(h_2,k_2) \in H \times K$,

$$\begin{aligned}(6.11)\quad \psi((h_1,k_1))\psi((h_2,k_2)) &= (h_1k_1)(h_2k_2) = h_1(k_1h_2)k_2 = h_1({}^{k_1}h_2k_1)k_2 \\ &= (h_1({}^{k_1}h_2))(k_1k_2) = \psi((h_1({}^{k_1}h_2), k_1k_2)).\end{aligned}$$

This shows how we could fix things so that ψ is a homomorphism of groups. We put a new product $*$ on the cartesian product of sets $H \times K$, where $(h_1,k_1) * (h_2,k_2) = (h_1({}^{k_1}h_2), k_1k_2)$. Then (6.11) shows that ψ satifies the homomorphism property from $(H \times K, *)$ to G. One can now check that $(H \times K)$ is a group under the operation $*$, and that ψ gives an isomorphism between this group and G. We don't check this here because it will follow from the next results.

We now abstract what we saw in the previous example to define an "external" version of this construction, which takes two groups and joins them together in a new way with a product defined by one acting on the other.

Definition 6.12. Let H and K be two groups and let $\phi : K \to \mathrm{Aut}(H)$ be a homomorphism of groups. Write $k \cdot h = \phi(k)(h)$, for $k \in K$ and $h \in H$. The *semidirect product* $H \rtimes_\phi K$ is defined to be the cartesian product $H \times K$ as a set, with operation $*$ defined by $(h_1, k_1) * (h_2, k_2) = (h_1(k_1 \cdot h_2), k_1 k_2)$.

We will check momentarily that the semidirect product is a group under $*$, but let us first explain the meaning of the extra piece of data we use to construct it, the homomorphism $\phi : K \to \mathrm{Aut}(H)$, and the notation $k \cdot h$. First of all, $\mathrm{Aut}(H)$ is a subgroup of $\mathrm{Sym}(H)$, so we can think of ϕ as a homomorphism $K \to \mathrm{Sym}(H)$. We know that such homomorphisms correspond to actions of K on H. Specifically, setting $k \cdot h = \phi(k)(h)$ as we have done, then this is the corresponding action of K on H. However, the fact that ϕ lands in $\mathrm{Aut}(H)$ gives us additional information—this means that $\phi(k)(h_1 h_2) = \phi(k)(h_1)\phi(k)(h_2)$, or equivalently $k \cdot (h_1 h_2) = (k \cdot h_1)(k \cdot h_2)$, for all $k \in K$, $h_1, h_2 \in H$. We say that K *acts on* H *by automorphisms*. Note that since acting by k is an automorphism of H, it must preserve the identity element, and so $k \cdot 1 = 1$ for all $k \in K$.

Proposition 6.13. *Let H and K be groups and let $\phi : K \to \mathrm{Aut}(H)$ be a homomorphism. Then the semidirect product $H \rtimes_\phi K$ is a group.*

Proof. This is a straightforward proof, but it is useful to go through the details to get a better feel for the construction. The associativity of the multiplication $*$ is not at all obvious, since it treats the two coordinates asymmetrically. First we calculate

$$((h_1, k_1) * (h_2, k_2)) * (h_3, k_3) = (h_1(k_1 \cdot h_2), k_1 k_2) * (h_3, k_3) = (h_1(k_1 \cdot h_2)((k_1 k_2) \cdot h_3), k_1 k_2 k_3)$$

and

$$(h_1, k_1) * ((h_2, k_2) * (h_3, k_3)) = (h_1, k_1) * (h_2(k_2 \cdot h_3), k_2 k_3) = (h_1 k_1 \cdot (h_2(k_2 \cdot h_3)), k_1 k_2 k_3).$$

From this we see there is no issue in the second coordinate, which is simply the multiplication in K. Now using that K is acting on H by automorphisms, we have

$$k_1 \cdot (h_2(k_2 \cdot h_3)) = (k_1 \cdot h_2)(k_1 \cdot (k_2 \cdot h_3)) = (k_1 \cdot h_2)((k_1 k_2) \cdot h_3)$$

which shows that the first coordinates of the expressions are also the same. This verifies associativity of $*$.

We claim that $(1, 1)$ is an identity element for $H \rtimes_\phi K$ under $*$. For this we check that $(1, 1) * (h, k) = (1(1 \cdot h), 1k) = (1h, 1k) = (h, k)$ and $(h, k) * (1, 1) = (h(k \cdot 1), k1) = (h1, k1) = (h, k)$, verifying the claim.

Finally, given $(h,k) \in H \rtimes_\phi K$, we claim that $(k^{-1} \cdot h^{-1}, k^{-1})$ is an inverse of (h,k) under $*$. First,

$$(h,k) * (k^{-1} \cdot h^{-1}, k^{-1}) = (h(k \cdot (k^{-1} \cdot h^{-1})), kk^{-1}) = (h(1 \cdot h^{-1}), kk^{-1}) = (hh^{-1}, kk^{-1}) = (1,1).$$

On the other side we calculate

$$(k^{-1} \cdot h^{-1}, k^{-1}) * (h,k) = ((k^{-1} \cdot h^{-1})(k^{-1} \cdot h), k^{-1}k) = (k^{-1} \cdot (h^{-1}h), k^{-1}k) = (k^{-1} \cdot 1, 1) = (1,1).$$

This verifies that every element has an inverse, and so $H \rtimes_\phi K$ is a group under $*$. □

Now that we have defined the semidirect product, we can complete the analysis of groups which are a product of two subgroups intersecting trivially, with only one of them required to be normal.

Theorem 6.14. *Let G be a group with subgroups H, K such that $H \trianglelefteq G$, $HK = G$, and $H \cap K = \{1\}$. Then $G \cong H \rtimes_\phi K$ for the homomorphism $\phi : K \to \mathrm{Aut}(H)$ defined by $\phi(k) = \rho_k$, where ρ_k is the automorphism $\rho_k(h) = {}^k h = khk^{-1}$ of H.*

Proof. For each $k \in G$ we have the inner automorphism θ_k of G defined by $\theta_k(g) = kgk^{-1}$ for $g \in G$. Since H is normal, its restriction $\rho_k = \theta_k|_H : H \to H$ is an automorphism of H (note that ρ_k need not be an inner automorphism of H, though). We have the formula $\theta_k \circ \theta_l = \theta_{kl}$ for inner automorphisms. Restricting to H we get $\rho_k \circ \rho_l = \rho_{kl}$ and thus $\phi : K \to \mathrm{Aut}(H)$ is a homomorphism of groups. So the the semidirect product $H \rtimes_\phi K$ is a well-defined group.

Now define a map $\psi : H \rtimes_\phi K \to G$ by $\psi((h,k)) = hk$. In the analysis at the beginning of this section we showed that ψ is a bijection of sets, and (6.11) showed that ψ is a homomorphism of groups. So ψ is an isomorphism of groups. □

We could call any group G with two subgroups H, K with $H \trianglelefteq G$, $HK = G$ and $H \cap K = \{1\}$ an "internal semidirect product". Theorem 6.14 then shows that the group is isomorphic to an "external semidirect product" of H and K, meaning a group defined by definition 6.12. The needed extra data ϕ comes from the internal relationship between H and K (the action of K on H by conjugation) that exists because they are two subgroups of a larger group G.

On the other hand we can show that an "external semidirect product" can always be thought of as an "internal semidirect product" of two of its subgroups. This is the content of the next proposition. (We are referring informally to internal and external semidirect products only to make an analogy with direct products. This is not standard terminology, which is why we have put the terms in quotes and will not use them from now on.)

Proposition 6.15. *Let H and K be groups, and let $\phi : K \to \mathrm{Aut}(H)$ be a homomorphism. Write $k \cdot h = \phi(k)(h)$ for all $k \in K, h \in H$. Let $G = H \rtimes_\phi K$.*

(1) *$\overline{K} = \{(1,k) | k \in K\}$ is a subgroup of G isomorphic to K.*

(2) *$\overline{H} = \{(h,1) | h \in H\}$ is a normal subgroup of G isomorphic to H.*

(3) *$\overline{H}\,\overline{K} = G$ and $\overline{H} \cap \overline{K} = \{1\}$.*

(4) *$(1,k)(h,1)(1,k)^{-1} = (k \cdot h, 1)$ for $k \in K$, $h \in H$.*

Proof. (1) Since $(1,k_1) * (1,k_2) = (1(k_1 \cdot 1), k_1 k_2) = (1, k_1 k_2)$ for $k_1, k_2 \in K$, it is immediate that $\overline{K}$ is a subgroup and that $\psi : K \to \overline{K}$ defined by $\psi(k) = (1,k)$ is an isomorphism. In particular, $(1,k)^{-1} = (1, k^{-1})$.

(2) Note that $(h_1,1) * (h_2,1) = (h_1(1 \cdot h_2), 1) = (h_1 h_2, 1)$. Thus it is also immediate that $\overline{H}$ is a subgroup of G and that $\psi : H \to \overline{H}$ defined by $\psi(h) = (h,1)$ is an isomorphism. We will prove that $\overline{H}$ is normal below.

(3) It is obvious that $\overline{H} \cap \overline{K} = \{1\}$ by definition. Also, note that $(h,1) * (1,k) = (h(1 \cdot 1), 1k) = (h,k)$ for any $h \in H$, $k \in K$. This shows that $\overline{H}\,\overline{K} = G$.

(4) We calculate

$$(1,k)(h,1)(1,k)^{-1} = (1,k)(h,1)(1,k^{-1}) = (1,k)(h,k^{-1}) = (k \cdot h, kk^{-1}) = (k \cdot h, 1).$$

We can now finish the proof of (2). Obviously $\overline{H} \subseteq N_G(\overline{H})$ since any subgroup normalizes itself. The formula in (4) shows that $\overline{K} \subseteq N_G(\overline{H})$. Thus $G = \overline{H}\,\overline{K} \subseteq N_G(\overline{H})$ and hence $\overline{H} \trianglelefteq G$. □

The proposition shows that any semidirect product $G = H \rtimes_\phi K$ has coordinate subgroups $\overline{H}$ and $\overline{K}$ such that $\overline{H}\,\overline{K} = G$, $\overline{H} \cap \overline{K} = \{1\}$, and $\overline{H} \trianglelefteq G$. Just as the case for direct products, we tend to identify $\overline{H}$ with H and $\overline{K}$ with K and think of H and K as subgroups of G. Moreover, although the homomorphism $\phi : K \to \mathrm{Aut}(H)$ starts out as "external data" which is needed to join H and K together into a semidirect product, once G is constructed the corresponding action of K on H can be recovered "internally" from the conjugation action of K on H inside G. This is exactly what Proposition 6.15(4) says.

We summarize the results so far as follows. Given any groups H and K and an action of K on H by automorphisms, we can use that action to construct a new group $G = H \rtimes K$, which contains copies of H and K as subgroups such that $HK = G$, $H \cap K = 1$, H is normal, and where the conjugation action of K on H inside G is equal to the original given action. Conversely, if G is a group with subgroups H and K such that H is normal, $HK = G$, and $H \cap K = 1$, then using the conjugation action of K on H to define a semidirect product $H \rtimes K$, that semidirect product is isomorphic to G.

It is worth noting that semidirect products of two groups contain direct products as a special case.

Lemma 6.16. *Let H and K be two groups, and let $\phi : K \to \mathrm{Aut}(H)$ be a homomorphism. Let $G = H \rtimes_\phi K$ and identify H and K with the coordinate subgroups of G. The following are equivalent:*

(1) *ϕ is the trivial homomorphism, that is $\phi(k) = 1_H$ for all k.*

(2) *$K \trianglelefteq G$.*

(3) *G is the internal direct product of H and K.*

Proof. We know that the subgroups H and K of the semidirect product always satisfy $HK = G$, $H \cap K = \{1\}$, and $H \trianglelefteq G$. Thus by definition G is the internal direct product of H and K if and only if $K \trianglelefteq G$ also, so (2) and (3) are equivalent.

Now one calculates $(h, 1) * (1, k) * (h, 1)^{-1} = (h, k)(h^{-1}, 1) = (h(k \cdot h^{-1}), k)$. Thus $K \trianglelefteq G$ if and only if $h(k \cdot h^{-1}) = 1$ for all $h \in H, k \in K$. But this is equivalent to $k \cdot h^{-1} = h^{-1}$, which clearly holds for all $h \in H$ and $k \in K$ if and only if ϕ is trivial. So (1) and (2) are equivalent as well. □

The lemma above says that $H \times_\phi K$ cannot be an internal direct product of the two special coordinate subgroups H and K unless ϕ is trivial. One warning: it is does not say that $H \times K$ and $H \times_\phi K$ cannot be isomorphic as groups without ϕ being trivial. It is possible that $H \times_\phi K$ could be an internal direct product of two different subgroups H' and K' which satisfy $H' \cong H$ and $K' \cong K$.

6.3. **Some automorphism groups.** Since a semidirect product depends on a homomorphism $\phi : K \to \mathrm{Aut}(H)$, to analyze the possibilities for specific K and H first requires one to understand the automorphism group of H, and then the possible homomorphisms from K to that group. Two examples that we will want to understand in detail are when H is cyclic and when H is an elementary abelian p-group for a prime p.

The automorphism group of a cyclic group $\mathbb{Z}_n$ can be calculated quite exactly.

Lemma 6.17. *Let $\mathbb{Z}_n$ be the additive group of integers modulo n. Let $\mathbb{Z}_n^\times = \{\bar{i} | \gcd(i, n) = 1\}$ be the group of units modulo n under multiplication. (This group was called U_n earlier in the notes.) In other words, $\mathbb{Z}_n^\times$ is the set of invertible elements in the monoid $\mathbb{Z}_n$ of congruence classes modulo n under multiplication.*

There is an isomorphism $\theta : \mathbb{Z}_n^\times \to \mathrm{Aut}(\mathbb{Z}_n)$, where $\theta(\bar{i}) = \sigma_i$, with $\sigma_i(\bar{j}) = i\,\bar{j} = \overline{ij}$.

We omit the proof of this lemma, leaving it as an exercise. In words, the automorphisms σ_i can be described as the maps "take the ith multiple", for any i which is relatively prime to n.

The structure of $\mathbb{Z}_n^\times$ is also understood. Note that this is a group of order $\varphi(n)$, where φ is the Euler φ-function, since $\mathbb{Z}_n^\times$ consists of those congruence classes modulo n that are relatively prime to n. We state the following theorem without proof at the moment.

Theorem 6.18. *Let $n \geq 1$ have prime factorization $n = p_1^{e_1} \dots p_k^{e_k}$, where the p_i are distinct primes and $e_i \geq 1$.*

(1) $\mathbb{Z}_n^\times \cong \mathbb{Z}_{p_1^{e_1}}^\times \times \cdots \times \mathbb{Z}_{p_k^{e_k}}^\times$.

(2) *if p is an odd prime and $e \geq 1$ then $\mathbb{Z}_{p^e}^\times \cong \mathbb{Z}_{p^e - p^{e-1}}$ is cyclic of order $p^e - p^{e-1} = p^{e-1}(p-1)$.*

(3) *$\mathbb{Z}_2^\times$ is trivial and $\mathbb{Z}_4^\times \cong \mathbb{Z}_2$ is cyclic. For $e \geq 3$, $\mathbb{Z}_{2^e}^\times \cong (\mathbb{Z}_2 \times \mathbb{Z}_{2^{e-2}})$, which is not cyclic.*

Part (1) of this theorem will be easily proved later when we study rings. We will also prove using ring theory the special case of part (2) where $e = 1$, namely that the group $\mathbb{Z}_p^\times$ is cyclic for any prime p. We will not prove the more general statement in part (2), or part (3); the proofs are not particularly difficult, though, and can be found in a text on number theory.

While it is straightforward to show abstractly that the group $\mathbb{Z}_n^\times$ decomposes as a certain product of cyclic groups, as described in the theorem above, actually finding an explicit isomorphism between $\mathbb{Z}_n^\times$ and that product of cyclic groups is another matter. For example, part (2) in the case $e = 1$ says that $\mathbb{Z}_p^\times$ is a cyclic group of order $p-1$ under multiplication. A number i such that $\overline{i}$ is a generator of $\mathbb{Z}_p^\times$ is called a *primitive root (modulo p)*. From the structure of cyclic groups, one can see that a cyclic group of order d has $\varphi(d)$ generators. Thus $\varphi(p-1)$ is the number of primitive roots. There is no formula that will produce primitive roots, and finding a primitive root for a large prime p is a computationally difficult task that depends on being able to find the prime factorization of $p-1$. We will only consider small primes in our examples, where it is easy to find a primitive root by trial and error.

Example 6.19. Let $G = \mathbb{Z}_{17}$. We know by Theorem 6.18 that $\mathbb{Z}_{17}^\times$ is a cyclic group of order $\varphi(17) = 16$, since 17 is prime. Now the number of generators of a cyclic group of order 16 is $\varphi(16) = 8$. So half of the classes in $\mathbb{Z}_{17}^\times$ are primitive roots modulo 17, that is, have order 16 in this group. We first try $\overline{2}$. We calculate $\overline{2}^4 = \overline{16} = \overline{-1}$, so $\overline{2}^8 = \overline{-1}^2 = \overline{1}$. Thus $\overline{2}$ has order 8 and is not a primitive root. So we try $\overline{3}$. $\overline{3}^2 = \overline{9}$, $\overline{3}^4 = \overline{9}^2 = \overline{81} = \overline{-4}$, so $\overline{3}^8 = \overline{-4}^2 = \overline{16} = \overline{-1} \neq \overline{1}$. Since all elements in this group must have order dividing 16, the only possibility is $|\overline{3}| = 16$ and so $\overline{3}$ is a primitive root. This allows us to find an explicit isomorphism $\theta : \mathbb{Z}_{16} \to \mathbb{Z}_{17}^\times$, by putting $\theta(\overline{i}) = \overline{3}^i$.

Recalling that by Lemma 6.17 we have $\mathbb{Z}_{17}^\times \cong \mathrm{Aut}(\mathbb{Z}_{17})$, we also see that $\mathrm{Aut}(\mathbb{Z}_{17})$ is cyclic of order 16, and that we can take $\sigma_3 : \overline{i} \mapsto \overline{3i}$ as a generator of this automorphism group.

Now we consider another example where we can calculate the automorphism group. Fix a prime p. An *elementary abelian p-group* is a group of the form $G = \prod_{i=1}^{m} \mathbb{Z}_p = \mathbb{Z}_p \times \mathbb{Z}_p \times \cdots \times \mathbb{Z}_p$ for some $m \geq 1$. The order of such a G is p^m so it is a p-group; moreover, it is easy to see that every non-identity element of G has order p. We know that $\mathbb{Z}_p$ also has a multiplication operation on congruence classes. Together with its addition operation, $\mathbb{Z}_p$ is a ring. In fact $\mathbb{Z}_p$ is a *field* which means that every nonidentity element of $\mathbb{Z}_p$ is invertible under multiplication, because $\mathbb{Z}_p^\times = \mathbb{Z}_p - \{0\}$. When thinking of $\mathbb{Z}_p$ as a field we write it as $\mathbb{F}_p$.

We can define a vector space over any field F: this is an abelian group V together with an action of F on V (scalar multiplication) satisfying the usual axioms. We can identify G with the set of column vectors

$$\mathbb{F}_p^m = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{pmatrix} \middle| a_i \in \mathbb{F}_p \right\},$$

and then define a scalar multiplication of $\mathbb{F}_p$ on elements of $\mathbb{F}_p^m$ in the obvious way. Then $G = \mathbb{F}_p^m$ becomes a vector space over the field $\mathbb{F}_p$. Write (a_i) for the vector with coordinates $a_1, a_2, \ldots, a_m$.

Now consider the group $\mathrm{Aut}(G)$. Since G is additive, an automorphism of G is a map $\sigma : G \to G$ which satisfies $\sigma(v + w) = \sigma(v) + \sigma(w)$ for all $v, w \in G$, that is, a map preserving vector addition. If $\lambda \in \mathbb{F}_p$, say $\lambda = \overline{j}$ for some $0 \leq j < p$, we have

$$\sigma(\lambda(a_i)) = \sigma((ja_i)) = \sigma(\overbrace{(a_i) + (a_i) + \cdots + (a_i)}^{j}) = \overbrace{\sigma((a_i)) + \sigma((a_i)) + \cdots + \sigma((a_i))}^{j} = \lambda\sigma((a_i))$$

for any $(a_i) \in G$. In other words, because σ preserves addition, it automatically preserves scalar multiplication. Thus σ is a linear transformation of the vector space $G = \mathbb{F}_p^m$. As such, it corresponds to an $m \times m$ matrix A with $\mathbb{F}_p$-coefficients, such that for $v \in \mathbb{F}_p^m$, $\sigma(v)$ is the same as the matrix product Av. Because σ is bijective, it is an invertible linear transformation and so $A \in \mathrm{GL}_m(\mathbb{F}_p)$, the group of invertible $m \times m$ matrices with coefficients in $\mathbb{F}_p$. Conversely, if $A \in \mathrm{GL}_m(\mathbb{F}_p)$, then left multiplication by A defines an invertible linear transformation of $\mathbb{F}_p^m$ and hence an automorphism of G as a group.

Proposition 6.20. *Let p be a prime and let $G = \overbrace{\mathbb{Z}_p \times \mathbb{Z}_p \times \cdots \times \mathbb{Z}_p}^{m}$ be an elementary abelian p-group.*

(1) $\mathrm{Aut}(G) \cong \mathrm{GL}_m(\mathbb{F}_p)$ *as groups.*

(2) $|\mathrm{Aut}(G)| = (p^m - 1)(p^m - p)\ldots(p^m - p^{m-1})$.

Proof. (1) It was shown in the discussion above that there is a natural bijection $\mathrm{Aut}(G) \to \mathrm{GL}_m(\mathbb{F}_p)$, where $\sigma \in \mathrm{Aut}(G)$ corresponds to the invertible matrix $A \in \mathrm{GL}_m(\mathbb{F}_p)$ such that $\sigma(v) = Av$ for

all $v \in G = \mathbb{F}_p^m$. This is an isomorphism of groups because, as shown in a linear algebra course, composition of linear transformations corresponds to multiplication of matrices.

(2) By (1), it suffices to calculate the size of $|\operatorname{GL}_m(\mathbb{F}_p)|$. An $m \times m$ matrix is invertible if and only if it has rank m, or in other words, its m columns form a basis of $\mathbb{F}_p^m$. So to count the number of invertible matrices we count the number of ordered bases $\{v_1, \ldots, v_m\}$ of $\mathbb{F}_p^m$. Any nonzero vector v_1 can be the start of a basis, so there are $(p^m - 1)$ choices for v_1. Once v_1 is chosen, v_2 can be any vector outside the span $\mathbb{F}_p v_1$ of v_1, which has p vectors, so there are $p^m - p$ choices for v_2. Similarly, the span of v_1, v_2 has p^2 elements and so there are $p^m - p^2$ choices for v_3. Continuning inductively, there are ultimately $p^m - p^{m-1}$ choices for v_m. This leads to the formula $(p^m - 1)(p^m - p) \ldots (p^m - p^{m-1})$ for the number of ordered bases of $\mathbb{F}_p^m$, and hence this is the size of $|\operatorname{GL}_m(\mathbb{F}_p)|$. □

Example 6.21. Consider $G = \operatorname{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2)$. We know that $G \cong \operatorname{GL}_2(\mathbb{Z}_2)$, and also $|G| = 6$ from Proposition 6.20 above. Since $\mathbb{Z}_2 \times \mathbb{Z}_2$ has 4 elements, 1 identity element and 3 elements of order 2, any automorphism of this group is determined by its permutation of the 3 non-identity elements. Since there are $|S_3| = 6$ such permutations, they all occur, and so we also have $G \cong S_3$ in this case.

6.4. **Examples and applications of semidirect products.** We can now return to groups of order pq and fully analyze them.

Example 6.22. Let G be a group with $|G| = pq$ where $p < q$ and p and q are primes. Let P and Q be a Sylow p-subgroup and a Sylow q-subgroup, respectively. We have seen that $Q \trianglelefteq G$, $PQ = G$, and $P \cap Q = \{1\}$ in Example 4.9. This is exactly the information we need to conclude that $G \cong Q \rtimes_\phi P$ is a semidirect product, where $\phi : P \to \operatorname{Aut}(Q)$ is a homomorphism, by Theorem 6.14.

We know that all groups of order p are cyclic, and so $P \cong \mathbb{Z}_p$. Similarly, $Q \cong \mathbb{Z}_q$. Additive notation can be confusing when used for the groups in a semidirect product $H \rtimes_\phi K$, particularly if one of H and K is written additively and the other is not. We often also want to find a presentation for our semidirect product, and free groups and presentations are written multiplicatively. So we prefer here to choose a generator a of P, so $P = \langle a \rangle = \{1, a, a^2, \ldots, a^{p-1}\}$, with $a^p = 1$, and we use multiplicative notation for P. In order words, we are thinking of P as the presented group $F(a)/(a^p)$. Similarly, we write $Q = \langle b \rangle = \{1, b, b^2, \ldots, b^{q-1}\}$, with $b^q = 1$.

To describe the possible semidirect products $G = Q \rtimes_\phi P$ we need to understand homomorphisms of groups $\phi : P \to \operatorname{Aut}(Q)$. Since Q is cyclic, by Lemma 6.17 there is an isomorphism $\theta : \mathbb{Z}_q^\times \to \operatorname{Aut}(Q)$. Transferring the isomorphism exhibited in that lemma to the multiplicative notation we

are using for Q, we see that $\theta(\bar{i}) = \sigma_i$, where $\sigma_i(b^j) = b^{ij} = (b^j)^i$ is the ith power map. Since q is prime, $\mathbb{Z}_q^\times = \mathbb{Z}_q - \{0\}$ is a cyclic group of order $q-1$, by Theorem 6.18.

Suppose that p does not divide $q-1$. Then any homomorphism $\phi : P \to \mathrm{Aut}(Q)$ is trivial, since the domain and target have relatively prime orders. In this case $Q \rtimes_\phi P \cong Q \times P \cong \mathbb{Z}_q \times \mathbb{Z}_p \cong \mathbb{Z}_{pq}$, and P must be normal in G as well. We already saw this in Example 4.9, where the fact that p does not divide $q-1$ was used to prove that $P \trianglelefteq G$ using the Sylow theorems instead, and hence G can be recognized as an internal direct product of P and Q.

If instead p does divide $q-1$, then since $\mathrm{Aut}(Q)$ is cyclic of order $q-1$, it has a unique subgroup of order p. If $\sigma \in \mathrm{Aut}(Q)$ is any element of order p, then there is a unique homomorphism $\phi : P \to \mathrm{Aut}(Q)$ such that $\phi(a) = \sigma$. This determines a semidirect product $G = Q \rtimes_\phi P$ for which P is not a normal subgroup, according to Lemma 6.16. In particular, G is not abelian.

The subgroup of order p in $\mathrm{Aut}(Q)$ has $p-1$ possible generators, i.e. every nonidentity element in this group. So there are actually $p-1$ different possible homomorphisms ϕ we could have chosen above, depending on which order p element the generator a of P gets sent to. Each one gives a nonabelian semidirect product $Q \rtimes_\phi P$. However, there is nothing that really distinguishes one generator of a cyclic group from another, and so it turns out that all of these semidirect products are isomorphic. We leave the details to Exercise 6.23(b).

Of course when p divides $q-1$ there is still also the possibility of taking $\phi : P \to \mathrm{Aut}(Q)$ to be the trivial homomorphism, and so $G \cong Q \times P$, which is abelian. Thus up to isomorphism there are two possible groups of order pq when p divides $q-1$: $Q \times P \cong \mathbb{Z}_q \times \mathbb{Z}_p \cong \mathbb{Z}_{pq}$, and $Q \rtimes_\phi P$ for any homomorphism $\phi : P \to \mathrm{Aut}(Q)$ mapping the generator of P to an element of order p.

The following exercise gives two common situations in which semidirect products $H \rtimes_{\phi_1} K$ and $H \times_{\phi_2} K$ for different homomorphisms $\phi_1, \phi_2 : K \to \mathrm{Aut}(H)$ can be proved to be isomorphic as groups.

Exercise 6.23. Let H and K be groups. Let $\phi : K \to \mathrm{Aut}(H)$ be a homomorphism of groups.

(a) Suppose that $\sigma \in \mathrm{Aut}(H)$ and let $\theta_\sigma : \mathrm{Aut}(H) \to \mathrm{Aut}(H)$ be the inner automorphism of $\mathrm{Aut}(H)$ given by $\rho \mapsto \sigma \circ \rho \circ \sigma^{-1}$. Let $\phi_2 = \theta_\sigma \circ \phi : K \to \mathrm{Aut}(H)$. Prove that $H \rtimes_\phi K$ and $H \rtimes_{\phi_2} K$ are isomorphic groups.

(b) Suppose that $\rho : K \to K$ is an automorphism of K and define $\phi_2 = \phi \circ \rho : K \to \mathrm{Aut}(H)$. Prove that $H \rtimes_\phi K$ and $H \rtimes_{\phi_2} K$ are isomorphic groups.

Let us demonstrate how one would find presentations for the groups of order pq. Rather than giving a general statement, let us just do this for a specific example.

Example 6.24. Consider groups of order $39 = (3)(13)$. Here $p = 3 < q = 13$, so we have p divides $q - 1$. We want to find an explicit primitive root modulo 13, in other words a generator of the order 12 group $\mathbb{Z}_{13}^{\times}$. Trying $\overline{2}$, we have $\overline{2}^4 = \overline{16} = \overline{3}$ and $\overline{2}^6 = \overline{64} = \overline{-1}$. Since every proper divisor of 12 divides 4 or 6, we must have $|\overline{2}| = 12$ and so 2 is a primitive root. Let $Q = \{1, b, b^2, \ldots, b^{12}\}$ be a cyclic group of order 13, where $b^{13} = 1$. Because $\overline{2}$ is a generator for $\mathbb{Z}_{13}^{\times}$, $\sigma \in \mathrm{Aut}(Q)$ given by "taking to the power 2", $\sigma(b^i) = b^{2i}$, generates the cyclic group $\mathrm{Aut}(Q)$, i.e. $|\sigma| = 12$. Then $H = \{1, \sigma^4, \sigma^8\}$ is the unique order 3 subgroup of $\mathrm{Aut}(Q)$. If $P = \{1, a, a^2\}$ is cyclic of order 3, we can define a homomorphism $\phi : P \to \mathrm{Aut}(Q)$ by sending a to any element of H. So we have three possible semidirect products $Q \rtimes_{\phi_i} P$, where $\phi_i(a) = \sigma^{4i}$, for $i \in \{0, 1, 2\}$.

Consider any of these groups $G = Q \rtimes_{\phi_i} P$. Since $(b^i, a^j) = (b^i, 1)(1, a^j) = (b, 1)^i(1, a)^j$ in G, clearly G is generated by the two elements $(b, 1)$ and $(1, a)$. Moreover, $(b, 1)^{13} = (b^{13}, 1) = (1, 1)$ and $(1, a)^3 = (1, a^3) = (1, 1)$. The key relation comes from looking at conjugation in G by the generator $(1, a)$: using Proposition 6.15(4), we have

$$(1, a)(b, 1)(1, a)^{-1} = (\phi_i(a)(b), 1) = (\sigma^{4i}(b), 1) = (b^{2^{4i}}, 1).$$

Note that $\overline{2}^{4i} = \overline{16}^i = \overline{3}^i$ in $\mathbb{Z}_{13}^{\times}$, so $b^{2^{4i}} = b^{3^i}$.

We claim now that $F(x, y)/(x^3 = 1, y^{13} = 1, xy = y^{3^i}x)$ is a presentation of G; the argument for this is similar to other examples we saw in the study of presentations earlier. There is clearly a homomorphism $\theta : F(x, y)/(x^3 = 1, y^{13} = 1, xy = y^{3^i}x) \to G$ sending $x \mapsto (1, b)$, $y \mapsto (a, 1)$, which is surjective since $(1, b)$ and $(a, 1)$ generate G. From the form of the relations we easily deduce that any element in $F(x, y)/(x^3 = 1, y^{13} = 1, xy = y^{3^i}x)$ is equal modulo relations to a word of the form $\{y^i x^j | 0 \leq i \leq 12, 0 \leq j \leq 2\}$. From this the presented group has order at most 13, and since it surjects onto a group of order 13, it must have exactly 13 elements and θ must be an isomorphism.

When $i = 0$, the presentation we get is $F(x, y)/(x^3 = 1, y^{13} = 1, xy = yx)$. This is the case where ϕ is trivial, and we know the group we get is $Q \times P$.

When $i = 1$ we get $F(x, y)/(x^3 = 1, y^{13} = 1, xy = y^3x)$ and when $i = 2$ we have $F(x, y)/(x^3 = 1, y^{13} = 1, xy = y^9x)$. It is claimed in Example 6.22 above that these two groups are isomorphic. Here one can easily demonstrate the isomorphism explicitly, by checking that there is an isomorphism $F(x, y)/(x^{13} = 1, y^3 = 1, yx = x^3y) \to F(x, y)/(x^{13} = 1, y^3 = 1, yx = x^9y)$ defined by $x \mapsto x$ and $y \mapsto y^2$.

Example 6.25. Consider groups G of order $2q$ for an odd prime q. This is a special case of the classification of groups of order pq. We have noted that there is one abelian such group and one

nonabelian group up to isomorphism. Since we know one nonabelian group of order $2q$ already, namely D_{2q}, the two possible groups must be $\mathbb{Z}_{2q}$ and D_{2q}.

To be more explicit, if $P = \langle b \rangle$ is cyclic of order 2 and $Q = \langle a \rangle$ is cyclic of order q, then there is a unique nontrivial homomorphism $\phi : P \to \mathrm{Aut}(Q)$, which maps b to the unique element σ of order 2 in the cyclic group $\mathrm{Aut}(Q)$. That element must be the "inversion map" $\sigma : Q \to Q$ given by $a^k \mapsto a^{-k}$ for all k, which obviously has order 2. Finding the corresponding presentation, similarly as in Example 6.24, leads to $F(a,b)/(a^q = 1, b^2 = 1, ba = a^{-1}b)$, the standard presentation for D_{2q}.

Next, let us consider an example where the structure of the automorphism group of an elementary abelian group comes into play.

Example 6.26. Consider a group G with $|G| = 18 = 2 \cdot 3^2$. The number n_3 of Sylow 3-subgroups divides 2 and is congruent to 1 modulo 3, so $n_3 = 1$ and a Sylow 3-subgroup Q is normal. Let P be a Sylow 2-subgroup. Then clearly $P \cap Q = \{1\}$, so $|PQ| = 18$ and $PQ = G$. We conclude that $G \cong Q \rtimes_\phi P$ for some homomorphism $\phi : P \to \mathrm{Aut}(Q)$. Since $|Q| = 3^2$, from our classification of groups of order p^2, either $Q \cong \mathbb{Z}_9$ or else $Q \cong \mathbb{Z}_3 \times \mathbb{Z}_3$.

Let us first consider the case $Q \cong \mathbb{Z}_9$. Then $\mathrm{Aut}(Q) \cong \mathbb{Z}_9^\times$, which is cyclic of order $\varphi(9) = 6$, by Theorem 6.18. It could be that $\phi : P \to \mathrm{Aut}(Q)$ is trivial. In this case we get $G \cong P \times Q \cong \mathbb{Z}_2 \times \mathbb{Z}_9 \cong \mathbb{Z}_{18}$, so G is cyclic. Since $\mathrm{Aut}(Q)$ is cyclic, it has a unique element of order 2. Thus the there is a unique nontrivial homomorphism $\phi : P \to \mathrm{Aut}(Q)$ which sends the generator of P to that element $\sigma \in \mathrm{Aut}(Q)$ with $|\sigma| = 2$. Similarly as in Example 6.25, this element σ must be the inversion map $a^i \mapsto a^{-i}$, where a is a generator of Q, and $Q \rtimes_\phi P$ will be isomorphic to the dihedral group D_{18}.

Otherwise, we have $Q \cong \mathbb{Z}_3 \times \mathbb{Z}_3$. In this case, we know that $\mathrm{Aut}(Q) \cong \mathrm{GL}_2(\mathbb{F}_3)$, by Proposition 6.20. Also, $|\mathrm{GL}_2(\mathbb{F}_3)| = (9-1)(9-6) = 48$. A map $\phi : P \to \mathrm{GL}_2(\mathbb{F}_3)$ is determined by sending the generator of P to an element $A \in \mathrm{GL}_2(\mathbb{F}_3)$ of order dividing 2. if $A = I$ is the identity matrix, then ϕ is trivial and so $Q \times_\phi P \cong Q \times P \cong \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_2 \cong \mathbb{Z}_3 \times \mathbb{Z}_6$. This is a non-cyclic abelian group.

We are left with the case where $|A| = 2$. Here, A is an invertible 2×2 matrix with entries in the field $\mathbb{F}_3$ with three elements. Suppose that BAB^{-1} is a conjugate of A in $\mathrm{GL}_2(\mathbb{F}_3)$. Then $|BAB^{-1}| = 2$ also, and if $\phi' : P \to \mathrm{GL}_2(\mathbb{F}_3)$ sends a generator to BAB^{-1} instead, then $Q \rtimes_\phi P \cong Q \rtimes_{\phi'} P$ follows from Exercise 6.23(a), since conjugation by B is an inner automorphism of $\mathrm{GL}_2(\mathbb{F}_3) \cong \mathrm{Aut}(Q)$. Because of this we only need to consider one matrix A from each conjugacy class in $\mathrm{GL}_2(\mathbb{F}_3)$ consisting of elements of order 2.

We will study conjugacy classes of matrices over fields in detail later when we develop the theory of canonical forms. Here we just state the end result; it will easily be justified by the reader later using canonical forms, or can be proved through brute force here. It turns out that every matrix A of order 2 is conjugate to one of the following matrices:

$$A_1 = \begin{pmatrix} \overline{1} & 0 \\ 0 & \overline{-1} \end{pmatrix} \qquad \text{or} \qquad A_2 = \begin{pmatrix} \overline{-1} & 0 \\ 0 & \overline{-1} \end{pmatrix}.$$

If $\phi_1 : P \to \mathrm{GL}_2(\mathbb{F}_3)$ sends the generator to A_1, note that A_1 is the automorphism of $\mathbb{Z}_3 \times \mathbb{Z}_3$ such that $(i, j) \mapsto (i, -j)$. In order to more easily find presentations, let us think of Q as the presented group $Q = F(a, b)/(a^3 = b^3 = 1, ba = ab)$. So the elements in Q are $\{a^i b^j | 0 \leq i \leq 2, 0 \leq j \leq 2\}$. Then in multiplicative notation, the matrix A_1 corresponds to the automorphism σ of Q with $\sigma(a^i b^j) = a^i b^{-j}$. Now consider $G = Q \rtimes_{\phi_1} P$ and identify P and Q with subgroups of G; this will make for simpler notation than we used when finding presentations in Example 6.24. If we write $P = \langle c \rangle$, then in G we will have a relation $c(a^i b^j)c^{-1} = \sigma(a^i b^j) = a^i b^{-j}$, by Proposition 6.15(4). A presentation of this group is given by $F(a, b, c)/(a^3 = b^3 = 1, ba = ab, c^2 = 1, ca = ac, cb = b^{-1}c)$, as the reader may easily check. This group is also isomorphic to $\mathbb{Z}_3 \times D_6$.

Finally, if $\phi_2 : P \to \mathrm{GL}_2(\mathbb{F}_3)$ sends the generator to A_2, this corresponds to the automorphism σ of Q with $\sigma(a^i b^j) = a^{-i} b^{-j}$. In other words, σ is the inversion map which is an order 2 automorphism of any abelian group. In this case $F(a, b, c)/(a^3 = b^3 = 1, ba = ab, c^2 = 1, ca = a^{-1}c, cb = b^{-1}c)$ is a presentation of the group $Q \rtimes_{\phi_2} P$. We call this group D'_{18} because it is a bit similar to the dihedral group, in that the generator of P is acting by the inversion automorphism on on the abelian group Q.

The analysis we have done shows that every group of order 18 is isomorphic to one of the following groups: $\mathbb{Z}_{18}$, $\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_2$, D_{18}, $\mathbb{Z}_3 \times D_6$, or D'_{18}. To complete the classification of groups of order 18, we ought to show that no two of these 5 groups are isomorphic. The first two are the only abelian ones, and they are not isomorphic since $\mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_2$ is not cyclic—all of its elements have order at most 6. Among the three remaining groups, D_{18} is the only one whose Sylow 3-subgroup is cyclic. Finally, $\mathbb{Z}_3 \times D_6$ and D'_{18} are not isomorphic because you can check that D'_{18} has a trivial center, while $\mathbb{Z}_3 \times D_6$ has center $\mathbb{Z}_3 \times \{1\}$.

6.5. **Groups of low order.** We now have enough techniques to fully classify groups of order less than or equal to 15 up to isomorphism.

First, groups of prime orders $p = 2, 3, 5, 7, 11$, or 13 are cyclic and isomorphic to $\mathbb{Z}_p$. Groups of of order a square of a prime, $p^2 = 2^2 = 4, 3^2 = 9$ are isomorphic to $\mathbb{Z}_{p^2}$ or $\mathbb{Z}_p \times \mathbb{Z}_p$. Groups of order pq

for primes $p < q$ are now classified by Example 6.22; there are two such groups when p divides $q-1$, and one group otherwise. In particular, groups of order $n = 6 = (2)(3), 10 = (2)(5)$ or $14 = (2)(7)$ are either the cyclic group $\mathbb{Z}_n$ or the dihedral group D_n; and groups of order $15 = (3)(5)$ are cyclic. Note that $|S_3| = 6$, so as an nonabelian group of order 6 we must have $S_3 \cong D_6$ (which is also easy to check directly). The only orders left which do not fall under any of our general classification results are 8 and 12, and so we will classify those next.

We should first mention here the classification of finite abelian groups. We will prove it later in these notes in the context of module theory, so have chosen not to emphasize it here.

Theorem 6.27. *Let G be a finite abelian group of order n. Then $G \cong \mathbb{Z}_{p_1^{e_1}} \times \mathbb{Z}_{p_2^{e_2}} \cdots \times \mathbb{Z}_{p_m^{e_m}}$, where each p_i is prime and $e_i \geq 1$ (the p_i need not be distinct). The list of prime powers $p_1^{e_1}, \ldots, p_m^{e_m}$ is uniquely determined by G up to reaarrangement, and two abelian groups of order n are isomorphic if and only if they have the same list of prime powers up to rearrangement.*

The theorem makes finding the abelian groups of a given order a triviality.

Example 6.28. Consider abelian groups of order 54. Each one corresponds to a sequence of prime powers whose product is $54 = (2)(3^3)$. Clearly then 2 is one of the prime powers, and for the others the possibilities are 3^3; 3^2 and 3; or 3, 3, and 3. So up to isomorphism, the abelian groups of order 54 are

$$\mathbb{Z}_2 \times \mathbb{Z}_{27}; \qquad \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_9; \qquad \text{and} \quad \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_3.$$

Theorem 6.27 also implies that these three groups are distinct up to isomorphism.

Now let us classify groups of order 8. Actually, groups of order p^3 for a prime p can be fully classified without too much work; but the case $p = 2$ behaves differently and has to be separately handled anyway.

Theorem 6.29. *There are precisely 5 distinct groups of order 8 up to isomorphism. The abelian ones are $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, and $\mathbb{Z}_8$. The nonabelian ones are D_8 and the quaternion group Q_8.*

Proof. The abelian part of the classification follows immediately from Theorem 6.27. So now let us assume that G is a nonabelian group of order 8, and show that either $G \cong D_8$ or $G \cong Q_8$.

If G has an element of order 8, then G is cyclic and we are back to the abelian case $\mathbb{Z}_8$. Similarly, if all nonidentity elements of G have order 2, then by an easy exercise, G again has to be abelian and in fact isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. So G has an element of order 4.

Let $a \in G$ have order 4, and let $H = \langle a \rangle = \{1, a, a^2, a^3\}$. Suppose that there is $b \notin H$ with $|b| = 2$. Then $K = \langle b \rangle = \{1, b\}$ satisfies $H \cap K = \{1\}$, and this clearly forces $|HK| = 8$ and thus

$HK = G$. Moreover, $H \trianglelefteq G$ because $|G : H| = 2$. We now recognize that G is isomorphic to a semidirect product $H \rtimes_\phi K$ for some homomorphism $\phi : K \mapsto \mathrm{Aut}(H)$. Since we are assuming G is not abelian, ϕ should be nontrivial. The only nontrivial automorphism of a cyclic group of order 4 such as H is the inversion map $\sigma : a \mapsto a^{-1}$, so we must have $\phi(b) = \sigma$. This means that a and b are related by $bab^{-1} = a^{-1}$. Thus in this case $G \cong D_8$, similarly as in Example 6.25.

Otherwise, every element outside of H has order 4. Since $|a| = |a^3| = 4$, a^2 is the only element of order 2 in the group. Let us name the element a^2 as -1. If x is another element of order 4 in G, then $|x^2| = 2$ and again $x^2 = -1$. Thus -1 commutes with x. Hence -1 commutes with all elements of the group and $-1 \in Z(G)$. For any $x \in G$, write $a^2x = xa^2$ as $-x$. Then this minus sign satisfies the obvious rules: $-(-x) = x$, and $-(x)(y) = (-x)(y) = x(-y)$. Also, if x has order 4, then $x(-x) = -x^2 = (-1)(-1) = 1$, so $-x = x^{-1}$.

Now choose $b \notin H$, so $|b| = 4$. Let $K = \langle b \rangle$. Let $c = ab$. Note that $c \notin H$ and $c \notin K$, as otherwise we would get the contradiction $H = K$. Since $|c| = 4$, $c^2 = -1$ as well. Now $c^{-1} = (ab)^{-1} = b^{-1}a^{-1} = (-b)(-a) = -(-ba) = ba$, so $ba = -ab = -c$. Multiplying $c = ab$ by a on the left gives $ac = a^2b = -b$, and multiplying $c = ab$ by b on the right gives $cb = ab^2 = -a$. Also, $ca = aba = a(-ab) = -a^2b = -(-b) = b$ and $bc = bab = (-ab)b = -a(b^2) = -(-a) = a$.

We now have elements $a, b, c, -1$ in G satisfying the relations $a^2 = b^2 = c^2 = -1$, $ab = c = -ba$; $bc = a = -cb$, and $ca = b = -ac$. It also easy to see that the 8 distinct elements of G are $\{\pm 1, \pm a, \pm b, \pm c\}$. Thus G has exactly the multiplication table of Q_8. □

Next we attack groups of order 12.

Theorem 6.30. *There are precisely* 5 *groups of order* 12 *up to isomorphism. The abelian ones are* $\mathbb{Z}_4 \times \mathbb{Z}_3 \cong \mathbb{Z}_{12}$ *and* $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3$. *The nonabelian ones are* A_4, D_{12}, *and a group* $T = \mathbb{Z}_3 \rtimes_\phi \mathbb{Z}_4$, *where* $\phi : \mathbb{Z}_4 \to \mathrm{Aut}(\mathbb{Z}_3)$ *is the unique nontrivial homomorphism.*

Proof. The classification of the abelian groups is immediate from Theorem 6.27. So let G be nonabelian of order 12. Let P be a Sylow 2-subgroup and Q a Sylow 3-subgroup of G. Consider the number n_3 of Sylow 3-subgroups. Since $n_3 \equiv 1(\mod 3)$ and $n_3|4$, the possibilities are $n_3 = 1$ or $n_3 = 4$. If $n_3 = 4$, counting elements gives $(4)(3-1) = 8$ elements of order 3 in G. Thus the remaining 4 elements are forced to form a Sylow 2-subgroup, and necessarily $P \trianglelefteq G$. It is easy to see that $P \cap Q = \{1\}$ and thus $PQ = G$. In this case we can proceed by noting that $G \cong P \rtimes_\phi Q$ and classifying the possible maps $\phi : Q \to \mathrm{Aut}(P)$. If $P \cong \mathbb{Z}_4$, then $\mathrm{Aut}(P) \cong \mathbb{Z}_2$ and there are no maps ϕ. So $P \cong \mathbb{Z}_2 \times \mathbb{Z}_2$ and $\phi : Q \to \mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z})$, where $\mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2) \cong \mathrm{GL}_2(\mathbb{F}_2)$. We saw in Example 6.21 that $\mathrm{GL}_2(\mathbb{F}_2) \cong S_3$. There is in fact a homomorphism $\phi : Q \to S_3$ (two of them,

depending on which element of order 3 a generator of Q maps to, but these lead to isomorphic semidirect products using Exercise 6.23). This leads to a unique nonabelian group $\mathbb{Z}_3 \rtimes_\phi (\mathbb{Z}_2 \times \mathbb{Z}_2)$ which has 4 Sylow 3-subgroups.

Actually, there is an easier way to see that there is a unique group up to isomorphism in the case there are 4 Sylow 3-subgroups, which shows that this semidirect product is something more familiar. If we have G act on Sylow 3-subgroups by conjugation, it gives a homomorphism $\psi : G \to S_4$. The kernel of ψ is $\{g \in G | gQg^{-1} = Q \text{ for all Sylow 3-subgroups } Q\}$. Since $n_3 = 4$, $N_G(Q) = Q$ for any Sylow 3-subgroup, so the kernel is contained in the intersection of all the Sylow 3-subgroups, which is clearly trivial. So ψ is injective, and hence $G \cong \psi(G)$. Now $\psi(G)$ is a subgroup of S_4 of order 12. We claim that if $H \leq S_4$ with $|S_4 : H| = 2$ then $H = A_4$. Because $|S_4 : H| = 2$, $H \trianglelefteq S_4$. Then if $\sigma \in S_4$, $(\sigma H)^2 = 1H$ in S_4/H since this group has order 2. This says $\sigma^2 \in H$. However, any 3-cycle is a square in S_4, since $(123) = (132)^2$. So H contains all 3-cycles. Now the 3-cycles generate A_4, so $A_4 = H$, proving the claim. Thus we see that any group of order 12 with 4 Sylow 3-subgroups is isomorphic to A_4. It follows that the nonabelian semidirect product $\mathbb{Z}_3 \rtimes_\phi (\mathbb{Z}_2 \times \mathbb{Z}_2)$ found above is isomorphic to A_4. This is not hard to see directly.

The other case is where $n_3 = 1$ and hence a Sylow 3-subgroup Q is normal. In this case we get $G \cong Q \rtimes_\phi P$ for a homomorphism $\phi : P \to \mathrm{Aut}(Q)$, where $\mathrm{Aut}(Q)$ is cyclic of order 2. If $P \cong \mathbb{Z}_4$, then there is a unique nontrivial homomorphism ϕ, sending a generator of P to the generator of $\mathrm{Aut}(Q)$. This leads to the group T described in the proposition.

If instead $P \cong \mathbb{Z}_2 \times \mathbb{Z}_2$, then there are multiple nontrivial homomorphisms $\phi : \mathbb{Z}_2 \times \mathbb{Z}_2 \to \mathrm{Aut}(Q) \cong \mathbb{Z}_2$, but one can see that they all differ by an automorphism ρ of $\mathbb{Z}_2 \times \mathbb{Z}_2$ and hence lead to isomorphic semidirect products by Exercise 6.23. Such a semidirect product $\mathbb{Z}_3 \rtimes_\phi (\mathbb{Z}_2 \times \mathbb{Z}_2)$ is easily shown to be isomorphic to D_{12}. This group is also isomorphic to $\mathbb{Z}_2 \times D_6$.

We leave the argument that D_{12}, T, and A_4 are all different up to isomorphism to the reader. □

7. Series in groups

7.1. **Commutators and the commutator subgroup.**

Definition 7.1. Let G be a group. For $x, y \in G$, we define the *commutator* of x and y to be $[x, y] = x^{-1}y^{-1}xy$. If X and Y are subsets of G, we define $[X, Y]$ to be the subgroup of G generated by all commutators $[x, y]$ with $x \in X$ and $y \in Y$.

It is easy to see that $[x, y] = 1$ if and only if $xy = yx$. Clearly, $[X, Y] = 1$ if and only if $xy = yx$ for all $x \in X$, $y \in Y$. Thus commutators give a way of expressing when every element

of one subset commutes with every element of another. We most often use this when X and Y are subgroups of G. It is important to note, however, that even if H and K are subgroups of G, then $S = \{[h,k]|h \in H, k \in k\}$ might not be a subgroup of G. We will give various constructions below in which it is crucial that $[H,K]$ be a subgroup, so one must take $[H,K]$ to be the subgroup generated by the set of commutators S, and not S itself.

Definition 7.2. Let G be a group. The *commutator subgroup* or *derived subgroup* of G is $G' = [G,G]$.

Since G' is the subgroup generated by all commutators, more explicitly it can be described as the set of all finite products of commutators of elements in G and the inverses of these commutators. Note that $[x,y]^{-1} = (x^{-1}y^{-1}xy)^{-1} = y^{-1}x^{-1}yx = [y,x]$. Thus in this case we can describe G' more compactly as the set of all finite products of commutators of elements in G.

Commutators interact with homomorphisms in the expected way.

Lemma 7.3. *Let $\phi : G \to H$ be a homomorphism of groups.*

(1) *Let K, L be subgroups of G. Then $\phi([K,L]) = [\phi(K), \phi(L)]$.*

(2) *$\phi(G') \subseteq H'$, with equality if ϕ is surjective.*

Proof. (1) Let $S = \{[x,y]|x \in K, y \in L\}$ and $T = \{[w,z]|w \in \phi(K), z \in \phi(L)\}$. Note that if $[x,y] \in S$ then $\phi([x,y]) = \phi(x^{-1}y^{-1}xy) = \phi(x)^{-1}\phi(y)^{-1}\phi(x)\phi(y) = [\phi(x),\phi(y)] \in T$. Similarly, if $[w,z] \in T$ then choosing $x \in K$ and $y \in L$ such that $\phi(x) = w$ and $\phi(y) = z$, we have $\phi([x,y]) = [w,z]$. Thus $\phi(S) = T$. Now taking the groups these generate we get

$$\phi([K,L]) = \phi(\langle S \rangle) = \langle \phi(S) \rangle = \langle T \rangle = [\phi(K), \phi(L)].$$

(2) Take $K = L = G$ in (1). □

We now give an important alternative characterization of the commutator subgroup.

Proposition 7.4. *Let G be a group, and G' its commutator subgroup.*

(1) *G' char G.*

(2) *If $H \trianglelefteq G$, then G/H is abelian if and only if $G' \subseteq H$.*

Proof. (1) This is immediate from applying Lemma 7.3(2) to an automorphism $\theta : G \to G$.

(2) We have that G/H is abelian if and only if $xHyH = yHxH$ for all $x, y \in G$, in other words if $xyH = yxH$ or $x^{-1}y^{-1}xy = [x,y] \in H$ for all $x, y \in G$. Since H is a subgroup this occurs if and only if $G' \subseteq H$. □

Note that since G' is normal in G (even characteristic), the proposition says that G' is the unique smallest normal subgroup H of G for which G/H is abelian. Equivalently, we can say that G/G' is the uniquely largest abelian factor group of G. This interpretation is the key to the applications of the commutator subgroup.

Example 7.5. Let $G = S_n$ for $n \geq 5$. Since A_n is a simple group, it is straightforward to see that $\{1\}$, A_n and S_n are the only normal subgroups of S_n. We cannot have $G' = 1$, since $G/G' = S_n$ is not abelian. On the other hand S_n/A_n has order 2 and is certainly abelian, so $G' \subseteq A_n$. It follows that $G' = A_n$. We could continue and ask what the commutator subgroup of A_n is. Again we cannot have $(A_n)' = 1$. Since A_n is simple, we must have $(A_n)' = A_n$.

For $n = 4$ the situation is different. We know that S_4 has proper normal subgroups A_4 and $V = \{1, (12)(34), (13)(24), (14)(23)\}$. S_4/V is not abelian, but rather isomorphic to S_3. On the other hand, $(S_4)' \subseteq A_4$ just as above. It follows that $(S_4)' = A_4$. One can also check that $(A_4)' = V$, and of course $V' = 1$, as V is abelian.

7.2. **Solvable groups.**

Definition 7.6. Let G be a group. A *subnormal series* in G is a chain of subgroups

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq H_2 \trianglelefteq \ldots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$$

where, as indicated, each H_i is normal in H_{i+1}. It is a *normal series* if each $H_i \trianglelefteq G$.

The n groups $H_1/H_0 \cong H_1$, H_2/H_1, $\ldots$, H_n/H_{n-1} are called the *factors* of the series.

Unfortunately there is not a consensus in the literature about the terminology for series. Some authors call what we have called a subnormal series a normal series. Some authors avoid giving names to these concepts at all, presumably because the existing terminology is confusing.

Definition 7.7. A group G is *solvable* if it has a subnormal series whose factors are abelian.

Example 7.8. Consider again $G = S_n$ for $n \geq 5$. Then the only possible subnormal series for G are $1 \trianglelefteq A_n \trianglelefteq S_n$ or $1 \trianglelefteq S_n$, which do not have abelian factors. So S_n is not solvable.

On the other hand, S_4 is solvable: the subnormal series $1 \trianglelefteq V \trianglelefteq A_4 \trianglelefteq S_4$ has abelian factors $V \cong Z_2 \times Z_2$, $A_4/V \cong \mathbb{Z}_3$, and $S_4/A_4 \cong \mathbb{Z}_2$, respectively.

The term solvable arises from Galois theory, where finite solvable groups are the ones that correspond to polynomial equations whose roots are solvable by radicals. We will see the connection when we study the theory of fields. While the original motivation came from Galois theory, solvable

groups are now an important object of study in group theory itself, and the definition is interesting for infinite groups as well as finite ones.

Definition 7.9. For any group G, let $G^{(0)} = G$, $G^{(1)} = G'$, and define inductively $G^{(n+1)} = (G^{(n)})'$ for all $n \geq 1$. Then $G \geq G^{(1)} \geq G^{(2)} \geq \cdots \geq G^{(n)} \geq \ldots$ is called the *derived series* of G.

Note that we have $G^{(n+1)} \operatorname{char} G^{(n)}$ for all n, by Proposition 7.4. Then $G^{(n)} \operatorname{char} G$ for all n by Proposition 1.60.

The derived series gives us a useful test for solvability of a group.

Theorem 7.10. *A group G is solvable if and only if $G^{(n)} = \{1\}$ for some $n \geq 0$.*

Proof. First let G be solvable, where $\{1\} = H_0 \trianglelefteq H_1 \trianglelefteq \ldots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$ is a subnormal series whose factors H_{i+1}/H_i are all abelian. It is actually more convenient to index in the other direction here, so let $K_i = H_{n-i}$. Then $\{1\} = K_n \trianglelefteq K_{n-1} \trianglelefteq \ldots \trianglelefteq K_1 \trianglelefteq K_0 = G$, with the factors K_i/K_{i+1} abelian.

Now we claim that $G^{(i)} \leq K_i$ for all $i \geq 0$. This is trivial when $i = 0$. Assume that $G^{(i)} \leq K_i$. Now $K_{i+1} \trianglelefteq K_i$ and K_i/K_{i+1} is abelian. By Proposition 7.4, this means that $(K_i)' \subseteq K_{i+1}$. But also $G^{(i)} \leq K_i$ clearly implies that $(G^{(i)})' \leq (K_i)'$, either by definition or by applying Lemma 7.3 to the inclusion map. Thus $G^{i+1} = (G^{(i)})' \leq (K_i)' \leq K_{i+1}$, completing the induction step. Thus $G^{(i)} \leq K_i$ holds for all $i \geq 0$ as claimed. In particular we have $G^{(n)} \leq K_n = \{1\}$.

Conversely, if $G^{(n)} = \{1\}$ for some n, then $\{1\} = G^{(n)} \trianglelefteq G^{(n-1)} \trianglelefteq \ldots \trianglelefteq G^{(1)} \trianglelefteq G^{(0)} = G$ is a subnormal series. The factors $G^{(i)}/G^{(i+1)} = G^{(i)}/(G^{(i)})'$ are abelian by Lemma 7.3. Thus G is solvable. □

Suppose that G is solvable. The theorem shows that the derived series reaches the bottom of the group G in a finite number of steps, but we have actually shown a bit more. The proof shows that given any subnormal series for G with abelian factors, then the terms of the derived series are descending from the top at least as fast. Thus the derived series descends fastest among subnormal series whose factors are abelian. Another conclusion from the result is that if G is solvable, then it has a *normal* series in which the factors are abelian, namely the derived series.

The next result could be proved directly from the definition of solvability by working with an arbitrary subnormal series with abelian factors. But our criterion for solvability using the derived series allows for a more elegant proof.

Proposition 7.11. *Let G be a group.*

(1) *If G is solvable, then any subgroup H of G is solvable.*

(2) *If G is solvable and $H \trianglelefteq G$, then G/H is solvable.*

(3) *If $H \trianglelefteq G$ and both H and G/H are solvable, then G is solvable.*

Proof. (1) We have $G^{(n)} = 1$ for some n, by Theorem 7.10. But applying Lemma 7.3 and induction, we have $H^{(i)} \subseteq G^{(i)}$ for all i. Thus $H^{(n)} = 1$ and H is solvable by Theorem 7.10 again.

(2) Again $G^{(n)} = 1$ for some n. Now apply Lemma 7.3 to the natural surjection $\pi : G \to G/H$ to obtain $\pi(G') = (G/H)'$. In particular, π restricts to a surjection from G' to $(G/H)'$. By induction we obtain $\pi(G^{(i)}) = (G/H)^{(i)}$ for all $i \geq 0$. Thus $(G/H)^{(n)} = \pi(G^{(n)}) = \pi(\{1\}) = \{1\}$ and so G/H is solvable by Theorem 7.10.

(3) As we just saw, $\pi(G^{(m)}) = (G/H)^{(m)}$, where $\pi : G \to G/H$ is the natural surjection. Since G/H is solvable, we have $(G/H)^{(m)} = \{1\}$ for some $m \geq 0$, by Theorem 7.10, and so $\pi(G^{(m)}) = \{1\}$. Hence $G^{(m)} \subseteq \ker \pi = H$. Now since H is solvable, we have $H^{(p)} = \{1\}$ for some $p \geq 0$. Then $(G^{(m)})^{(p)} \subseteq H^{(p)} = \{1\}$. But clearly $(G^{(m)})^{(p)} = G^{m+p}$. So $G^{(m+p)} = \{1\}$ and G is solvable by Theorem 7.10. □

Let us make some additional comments about the theorem. Given a solvable group G, its *derived length* is the smallest integer $n \geq 0$, if any, such that $G^{(n)} = \{1\}$. The derived length is a rough measure of how far a solvable group is from being abelian, since a nontrivial abelian group has derived length 1. Note that the proposition above implies relationships among the derived lengths. Namely, we actually proved that if G has derived length n, then the derived length of any subgroup $H \leq G$ or any factor group G/H is at most n. Also, if $H \trianglelefteq G$ where G/H has derived length m and H has derived length p, then G has derived length at most $m + p$.

Suppose that $H \trianglelefteq G$, and let $K = G/H$. In some sense G is "built up" out of the subgroup H and the factor group K. In this setting we say that G is an *extension* of K by H. Calling G an extension of H by K might seem more natural, because we are enlarging H to the group G, and $K = G/H$ is what is "added on". However, the given terminology is standard for historical reasons.

If one starts with groups H and K, one can ask what the ways are that one can put them together to form a group G which is an extension of K by H. This is called the *extension problem*, which is closely related to the theory of cohomology of groups. The reader can see Chapter 7 of Rotman's book "An introduction to the theory of groups" for an introduction to this theory. In this language, Proposition 7.11(3) says that any group which is an extension of a solvable group by another solvable one is itself solvable. We can express this by saying that the property of being solvable is "closed under extensions".

Of course, all abelian groups are solvable. We saw above that S_4 is solvable, while S_n is not for $n \geq 5$. It is easy to see that finite p-groups are solvable, as will become clear in the next section. More generally, Burnside proved that if $|G| = p^i q^j$ for primes p and q, then G is solvable. The proof is considerably more difficult and requires the methods of representation theory. One of the biggest acheivements in this direction is a famous theorem of Feit and Thompson. They proved that if G is finite of odd order, then G is solvable. Their theorem was a major stepping stone toward the classification of finite simple groups, since it ruled out the possibility of nonabelian simple groups of odd order.

7.3. **Nilpotent groups.** Nilpotent groups are a class of groups more special than solvable groups. We will see that finite nilpotent groups can be characterized in a nice way in terms of their Sylow subgroups. The reader is more likely to encounter the notion of nilpotence in the case of infinite groups, for example in the theory of Lie groups.

Definition 7.12. A group G is nilpotent if it has a normal series

$$\{1\} = H_0 \leq H_1 \leq \cdots \leq H_{n-1} \leq H_n = G$$

(so $H_i \trianglelefteq G$ for all i) such that $H_{i+1}/H_i \subseteq Z(G/H_i)$ for all $0 \leq i \leq n-1$. Such a normal series is called a *central series* for G.

Recall that by definition in a normal series each term H_i is normal in G, as opposed to a subnormal series where each H_i is only required to be normal in the next term H_{i+1}. This is necessary since the definition refers to the factor group G/H_i. Of course this implies that $H_i \trianglelefteq H_{i+1}$ for all i as well, but we avoided writing that in the notation for the series so as to not suggest that the series is only subnormal.

The condition that each factor H_{i+1}/H_i be inside the center of the factor group G/H_i takes some time to process. We will see a number of examples shortly. Actually, it is convenient to recast this condition using the notation of commutators, which allows one to avoid the explicit use of cosets.

Lemma 7.13. *Let $H \leq K \leq G$ where $H \trianglelefteq G$. Then $K/H \subseteq Z(G/H)$ if and only if $[G, K] \subseteq H$.*

Proof. An arbitrary element of K/H is xH with $x \in K$, and an arbitrary element of G/H is gH with $g \in G$. For K/H to be contained in the center of G/H means that $xHgH = gHxH$ for all $x \in K$ and all $g \in G$. This is equivalent to $xgH = gxH$ or $[g, x] = g^{-1}x^{-1}gx \in H$ for all $g \in G, x \in K$. Since H is a subgroup, this is equiavlent to $[G, K] \subseteq H$. □

Using the lemma, we see that a normal series $\{1\} = H_0 \leq H_1 \leq \cdots \leq H_{n-1} \leq H_n = G$ is a central series if and only if $[G, H_{i+1}] \subseteq H_i$ for all $0 \leq i \leq n-1$. We can think of $[G, -]$ as an operation on subgroups of G, and a central series is one where hitting each term of the series by this operation pushes you down into the next lowest term.

Example 7.14. Any nilpotent group is solvable. If G has a central series $\{1\} = H_0 \leq H_1 \leq \cdots \leq H_{n-1} \leq H_n = G$, then it is also a subnormal series, and since each H_{i+1}/H_i is in the center of a group G/H_i, in particular H_{i+1}/H_i is abelian.

Obviously any abelian group is nilpotent. We will show in a bit that any finite p-group for a prime p is nilpotent.

Example 7.15. Any nontrivial nilpotent group has a nontrivial center. If G has a central series $\{1\} = H_0 \leq H_1 \leq \cdots \leq H_{n-1} \leq H_n = G$, we can certainly assume that $H_i \subsetneq H_{i+1}$ for all i, otherwise some of the terms of the series can just be removed to get a shorter central series. Then since G is nontrivial, H_1 is a nontrivial subgroup of G, and by definition H_1/H_0 is in the center of G/H_0, i.e. $\{1\} \neq H_1 \subseteq Z(G)$.

For example, S_3 is not nilpotent, since $Z(S_3) = \{1\}$. This is the smallest example of a non-nilpotent group. On the other hand, S_3 is solvable.

Above, we defined one particularly special series of subgroups, the derived series, which can be investigated to tell if a group is solvable: namely, G is solvable if its derived series reaches the identity subgroup in finitely many steps. We can define a special series of subgroups which serves the same purpose for detecting whether a group is nilpotent. But actually in this case there are two different choices, both of which can be useful.

Definition 7.16. Let G be a group. The *upper central series* of G is defined as follows. Put $Z_0 = \{1\}$ and $Z_1 = Z(G)$. Then $Z_1 \trianglelefteq G$, so we can consider the factor group G/Z_1. The center $Z(G/Z_1)$ of G/Z_1 has the form $Z(G/Z_1) = Z_2/Z_1$ for some subgroup Z_2 with $Z_1 \leq Z_2 \leq G$, and since $Z(G/Z_1) \trianglelefteq G/Z_1$ we have $Z_2 \trianglelefteq G$. Continuing in this way, we construct a sequence of subgroups $Z_0 \leq Z_1 \leq Z_2 \ldots$ of G which we call the upper central series.

Proposition 7.17. *Let G be a group and let $Z_0 \leq Z_1 \leq Z_2 \leq \ldots$ be the upper central series of G.*

(1) *$Z_i \operatorname{char} G$ for all $i \geq 0$.*

(2) *G is nilpotent if and only if $Z_n = G$ for some $n \geq 0$.*

Proof. (1) $Z_0 \operatorname{char} G$ is obvious. Assume that $Z_i \operatorname{char} G$ for some i. If $\sigma \in \operatorname{Aut}(G)$, then $\sigma(Z_i) = Z_i$ and it follows that there is an induced automorphism $\overline{\sigma} : G/Z_i \to G/Z_i$ given by $\overline{\sigma}(gZ_i) = \sigma(g)Z_i$. Since the center of a group is characteristic, $\overline{\sigma}(Z(G/Z_i)) = Z(G/Z_i)$. But since $Z(G/Z_i) = Z_{i+1}/Z_i$ this is equivalent to $\sigma(Z_{i+1}) = Z_{i+1}$. So $Z_{i+1} \operatorname{char} G$ and the result is proved by induction.

(2) Suppose first that $Z_n = G$. Then $Z_0 \leq Z_1 \leq Z_2 \leq \cdots \leq Z_n = G$ is a normal series for G, by (1). By definition, for all i we have $Z_{i+1}/Z_i \subseteq Z(G/Z_i)$ (in fact this is an equality) and so we have a central series for G, and G is nilpotent.

Conversely, if G is nilpotent, let $H_0 = \{1\} \leq H_1 \leq \cdots \leq H_n = G$ be some central series of G. Then we claim that $H_i \subseteq Z_i$ for all i. This is trivial when $i = 0$. Assume that $H_i \subseteq Z_i$. Since $H_{i+1}/H_i \subseteq Z(G/H_i)$, this means that $[G, H_{i+1}] \subseteq H_i \subseteq Z_i$. This translates back to $(H_{i+1}Z_i)/Z_i \leq Z(G/Z_i) = Z_{i+1}/Z_i$, which implies $H_{i+1} \leq Z_{i+1}$. The claim that $H_i \subseteq Z_i$ for all i now holds by induction.

In particular, $H_n = G \subseteq Z_n$ and so $Z_n = G$. □

This proof showed that the terms Z_i of the upper central series are "above" the terms H_i of an arbitary central series. This is why it is called the upper central series; it is the central series ascending most quickly from the bottom of the group.

Example 7.18. Let G be a finite p-group for a prime p. Then we claim that G is nilpotent. This is easiest to prove using the upper central series. We may assume that G is nontrivial. Let $Z_0 = \{1\}$ and $Z_1 = Z(G)$. We know that nontrivial p-groups have a non-trivial center, so $Z_0 \subsetneq Z_1$. If $Z_1 = G$, we are done. Otherwise the group G/Z_1 is again a nontrivial p-group, so it has a nontrivial center, which is by definition Z_2/Z_1. So $Z_1 \subsetneq Z_2$. In this way we prove that as long as $Z_i < G$, that $Z_i \subsetneq Z_{i+1}$. Since G is finite this process must terminate with $Z_n = G$ for some n. Hence by Proposition 7.17, G is nilpotent as claimed.

We briefly discuss the other canonical series of groups that can be used to check nilpotence.

Definition 7.19. Let G a group. We define the *lower central series* of G as follows. Let $G^1 = G$. For each $n \geq 1$, define by induction $G^{i+1} = [G, G^i]$. The lower central series for G is $G^1 = G \geq G^2 \geq G^3 \geq \ldots$

Note that $G^2 = [G, G^1] = [G, G] = G'$ is the same as the derived subgroup of G. But $G^3 = [G, G^2]$ is in general bigger than the next term in the derived series, which is $G'' = [G', G']$. Also, notice that the lower central series is traditionally indexed differently, starting at the top with G^1 rather than G^0.

Similarly as for the derived series, we can check if a group is nilpotent by seeing if the lower central series reaches the identity subgroup in finitely many steps.

Proposition 7.20. *Let G be a group.*

(1) *G^i $\operatorname{char} G$ for all $i \geq 1$.*

(2) *G is nilpotent if and only if $G^n = \{1\}$ for some $n \geq 1$.*

Proof. (1) This is proved by induction on i. Assuming $G^i \operatorname{char} G$, by Lemma 7.3 if $\sigma \in \operatorname{Aut}(G)$ then $\sigma([G, G^i]) = [\sigma(G), \sigma(G^i)] = [G, G^i]$, so $[G, G^i] = G^{i+1} \operatorname{char} G$ as well, completing the induction step.

(2) Suppose that $G^n = \{1\}$. Consider the series $\{1\} = G^n \leq G^{n-1} \leq \cdots \leq G^1 = G$, which is a normal series by (1). By definition, $[G, G^i] = G^{i+1}$ for all $i \geq 1$. We saw in Lemma 7.13 that this implies $G^i/G^{i+1} \leq Z(G/G_{i+1})$ for all $i \geq 1$. So we have a central series and G is nilpotent.

Conversely, suppose $\{1\} = H_n \leq H_{n-1} \leq \cdots \leq H_2 \leq H_1 = G$ is some central series for G (we choose an indexing that is most convenient for comparison to the lower central series). We claim that $G^i \leq H_i$ for all $i \geq 1$. This is trivial when $i = 1$. Assume now that $G^i \leq H_i$ for some i. Then since the H_i form a central series, $[G, H_i] \subseteq H_{i+1}$, using Lemma 7.13. So $G^{i+1} = [G, G^i] \subseteq [G, H_i] \subseteq H_{i+1}$, proving the induction step and the claim. In particular, $G^n \leq H_n = \{1\}$. □

The proof of the proposition actually shows that the terms of the lower central series G^n are contained in the terms H_n of an arbitrary central series. That is, the central series G^n is the "lowest" possible central series, the one that descends most quickly from the top.

Corollary 7.21. *Let G be nilpotent.*

(1) *If $H \leq G$, then H is nilpotent.*

(2) *If $H \trianglelefteq G$, then G/H is nilpotent.*

(3) *If G and K are nilpotent groups, then $G \times K$ is nilpotent.*

Proof. (1) It is easy to prove by induction that $H^i \leq G^i$ for all i. Since G is nilpotent, $G^n = \{1\}$ for some $n \geq 1$ by Proposition 7.20. Then $H^n = \{1\}$ and so H is also nilpotent by Proposition 7.20 again.

(2) Let $\pi : G \to G/H$ be the natural quotient homomorphism. We claim that $\pi(G^i) = (G/H)^i$ for all $i \geq 1$. This is trivial when $i = 1$. If it is true for some i, then $\pi(G^{i+1}) = \pi([G, G^i]) = [\pi(G), \pi(G^i)] = [G/H, (G/H)^i] = (G/H)^{i+1}$ by Lemma 7.3, proving the induction step and the claim. Now Since $G^n = \{1\}$ for some n, we also have $(G/H)^n = \{1\}$ and so G/H is nilpotent by Proposition 7.20.

(3) It is easy to prove by induction that $(H \times K)^i = H^i \times K^i$. Since $H^m = \{1\}$ and $K^p = \{1\}$ for some m and p, then $(H \times K)^n = \{(1,1)\}$ for $n = \max(m,p)$. □

Note that Corollary 7.21(3) is weaker than the corresponding property of solvable groups; only products of nilpotent groups are nilpotent, not arbitrary extensions of nilpotent groups. We have already seen that S_3 is not nilpotent since it has a trivial center; on the other hand S_3 is certainly an extension of two nilpotent groups, since it has a normal subgroup $H = \{(123)\}$ such that $S_3/H \cong \mathbb{Z}_2$ and $H \cong \mathbb{Z}_3$.

Example 7.22. If $G = P_1 \times P_2 \times \ldots P_n$, where each P_i is a p_i-group for some prime p_i, then G is nilpotent. This follows since each P_i is nilpotent, by Example 7.18, and nilpotent groups are closed under taking products, by Corollary 7.21.

We will see later that all finite nilpotent groups look like the ones in Example 7.22.

7.4. **The Frattini argument and more on nilpotent groups.** We have seen examples of groups G with subgroups H that are "self-normalizing", that is $N_G(H) = H$. For example, if P is a Sylow p-subgroup and $n_p = |G : P|$ is as large as possible, then since $n_p = |G : N_G(P)|$ by the Sylow theorems, we must have $P = N_G(P)$. For a more specific example, this happens if $|G| = pq$ with p dividing $q - 1$, where the nonabelian such example has q Sylow p-subgroups, so $P = N_G(P)$ for a Sylow p-subgroup P.

We see next that, in contrast, a nilpotent group cannot have any proper self-normalizing subgroups. One summarizes this by saying that "normalizers grow in nilpotent groups".

Proposition 7.23. *Let G be a nilpotent group. If H is a proper subgroup of G, then $H \subsetneq N_G(H)$.*

Proof. Consider any central series for G, say $\{1\} = G_0 \leq G_1 \leq \cdots \leq G_{n-1} \leq G_n = G$. Let H be a proper subgroup of G. Note that $G_0 = \{1\} \subseteq H$. Let $i \geq 0$ be maximum such that $G_i \subseteq H$. Since H is proper, $i < n$, so $G_i \subseteq H$ and $G_{i+1} \nsubseteq H$.

Now by the definition of a central series and Lemma 7.13, $[G, G_{i+1}] \subseteq G_i$. In particular, $[H, G_{i+1}] \subseteq G_i$. If $g \in H$ and $x \in G_{i+1}$ this says that $[g,x] = g^{-1}(x^{-1}gx) \in G_i$. Thus $x^{-1}gx \in gG_i \subseteq H$ since $g \in H$ and $G_i \subseteq H$. This shows that $x^{-1}Hx \subseteq H$, so $x^{-1}Hx = H$ since H is finite. This implies that $G_{i+1} \subseteq N_G(H)$. But since $G_{i+1} \nsubseteq H$, we obtain $H \subsetneq N_G(H)$. □

There is a nice technique called "Frattini's argument" that sometimes comes in handy in the analysis of normalizers.

Lemma 7.24 (Frattini's argument)**.** *Let G be a group with $N \trianglelefteq G$. Suppose that N is finite and P is a Sylow p-subgroup of N for some prime p. Then $N_G(P)N = G$.*

The statement of the result is not very intuitive, as it suggests the normalizers of Sylow p-subgroups should be "big", i.e. big enough to generate G along with N. After all, we gave examples above of Sylow p-subgroups that are self-normalizing. But one must remember that P is a Sylow p-subgroup of N, not of G, so its normalizer may well be bigger than that of a Sylow p-subgroup of G. And the fact that N is itself normal plays a key role in ensuring that $N_G(P)$ is large. This may be an example of a theorem that only makes sense once one sees the rather simple and elegant proof.

Proof. Let $x \in G$. Note that $xPx^{-1} \subseteq xNx^{-1} = N$, since $N \trianglelefteq G$. Since xPx^{-1} is a conjugate of P, $|xPx^{-1}| = |P|$ and so xPx^{-1} must be another Sylow p-subgroup of N. Now we use the Sylow conjugacy theorem in the group N: all Sylow p-subgroups of N are conjugate *in* N, that is, by an element of N. So there is $y \in N$ with $y(xPx^{-1})y^{-1} = P$. Now $(yx)P(yx)^{-1} = P$, which means that $yx \in N_G(P)$. Setting $z = yx \in N_G(P)$, we have $x = y^{-1}z \in NN_G(P)$. Since $x \in G$ was arbitrary, $G = NN_G(P) = N_G(P)N$ (since $N \trianglelefteq G$). □

We now have all of the ingredients for some very nice characterizations of finite nilpotent groups.

Theorem 7.25. *Let G be a finite group. The following are equivalent:*

(1) *G is nilpotent.*
(2) *All maximal subgroups of G are normal in G.*
(3) *All Sylow p-subgroups of G are normal in G.*
(4) *G is a finite direct product of groups of prime power order.*

Proof. (1) $\implies$ (2): Let G be nilpotent and let $M \subsetneq G$ be a maximal subgroup of G. By definition, there is no subgroup H with $M \subsetneq H \subsetneq G$. However, we know that normalizers grow in nilpotent groups, so $M \subsetneq N_G(M)$, by Proposition 7.23. This forces $N_G(M) = G$, so $M \trianglelefteq G$.

(2) $\implies$ (3): Let P be a Sylow p-subgroup of G for some prime p. Suppose that P is not normal in G, so $N_G(P) \subsetneq G$. Since G is finite and $N_G(P)$ is proper, we can choose some maximal subgroup M of G with $N_G(P) \subseteq M \subsetneq G$. Now by assumption (2), M is normal. Apply Frattini's argument to M, noting that because P is a Sylow p-subgroup of G, it must also be a Sylow p-subgroup of M. Lemma 7.24 gives $G = MN_G(P)$. But $N_G(P) \subseteq M$ so $MN_G(P) = M \subsetneq G$, a contradiction. So P is normal in G after all.

(3) $\implies$ (4): Let $p_1, \ldots, p_k$ be the distinct prime factors of $|G|$ and let P_i be a Sylow p_i-subgroup for each i. We saw earlier that when $P_i \trianglelefteq G$ for all i, that G is an internal direct product of $P_1, P_2, \ldots, P_k$ and so $G \cong P_1 \times P_2 \times \cdots \times P_k$ (Corollary 6.4).

(4) $\implies$ (1): this is the content of Example 7.22. □

The theorem shows that finite nilpotent groups are just the groups in which all of their Sylow p-subgroups are normal. They are also just mild generalizations of finite p-groups (finite products of p-groups). Given that, the reader might wonder we we bother with the rather more complicated definition of nilpotent group. The point is that this concept is also important in the theory of infinite groups, where nilpotent groups don't admit such a simple alternative description.

7.5. **Composition series.** In this optional section, we review some of the basic properties of composition series, another type of series that is useful in describing finite groups.

Definition 7.26. A *composition series* for a group G is a subnormal series

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \ldots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$$

such that every factor H_{i+1}/H_i is a simple group. The factors of the composition series are called *composition factors*. The *length* of the composition series is the number n of simple factors; A group G has *finite length* if it has a composition series. In this case the *length of* G, written $\ell(G)$, is the smallest $n \geq 0$ such that G has a composition series of length n. By convention, the trivial group $G = \{1\}$ is considered to have the composition series $\{1\} = H_0 = G$ of length 0 with no factors.

Notice that a composition series is a subnormal series with nontrivial factors which is maximal in the sense that we cannot insert any more terms. If, say in between H_i and H_{i+1} we tried to add another subgroup K with $H_i \trianglelefteq K \trianglelefteq H_{i+1}$, then by subgroup correspondence we would have $K/H_i \trianglelefteq H_{i+1}/H_i$. Since H_{i+1}/H_i is simple, that would force $K = H_i$ or $K = H_{i+1}$, so inserting K would lead to a subnormal series with a trivial factor. (Recall that by convention the trivial group is not simple.)

We claim that every finite group G has a composition series. If G is trivial, we agree by the above convention that G has a composition series with no factors. If G is nontrivial, first note that among the proper normal subgroups of G, since G is finite we can choose one, say H_1, which is maximal in the sense that there are no normal subgroups K of G with $H_1 \subsetneq K \subsetneq G$. Then G/H_1 must be a simple group by subgroup correspondence. Now in a similar way we can choose a maximal proper normal subgroup H_2 of H_1, and so on. Because each time we choose a proper subgroup, this

process must end at some point with $H_n = \{1\}$, and then $\{1\} = H_n \trianglelefteq H_{n-1} \trianglelefteq \ldots \trianglelefteq H_1 \trianglelefteq H_0 = G$ is a composition series for G.

Thus all finite groups have finite length. An infinite group might or might not have finite length.

Example 7.27. Given a cylic group of order n, say $G = \langle a \rangle$, then choosing any sequence of (not necessarily distinct) prime numbers $p_1, p_2, \ldots p_k$ whose product is n, we get a sequence of subgroups

$$H_0 = \{1\} \trianglelefteq H_1 = \langle a^{p_2 p_3 \cdots p_k} \rangle \trianglelefteq H_2 = \langle a^{p_3 \cdots p_k} \rangle \trianglelefteq \cdots \trianglelefteq H_{k-1} = \langle a^{p_k} \rangle \trianglelefteq H_k = G = \langle a \rangle$$

where H_{i+1}/H_i has prime order p_i for each i, and hence $H_{i+1}/H_i \cong \mathbb{Z}_p$ is simple. So this is a composition series for G.

We see from the previous example that a group may have many different composition series; in that example one can take the primes whose product is n and put them in any desired order. For example, if $n = p_1 p_2 \ldots p_k$ happened to be a product of distinct primes $p_1, p_2, \ldots, p_k$ then there would be $k!$ choices.

Since a given group might have many different composition series, an obvious question is how different they can actually be. The Jordan-Hölder Theorem, which we prove next, shows that for most purposes the differences are not substantial. Namely, the number of terms in a composition series of a group is always the same, and the same list of simple composition factors must occur up to isomorphism after rearranging the lists. The result is important to know, but the proof is rather technical and the reader may safely skip the proof on a first reading.

Theorem 7.28 (Jordan-Hölder)**.** *Let G be a group of finite length $n = \ell(G) < \infty$. Choose a composition series $G_0 = \{1\} \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{n-1} \trianglelefteq G_n = G$ that achieves this minimal length, with simple factors $T_i = G_i/G_{i-1}$ for $1 \leq i \leq n$. Let $H_0 = \{1\} \trianglelefteq H_1 \cdots \trianglelefteq H_{m-1} \trianglelefteq H_m = G$ be another composition series for G, with simple factors $U_i = H_i/H_{i-1}$ for $1 \leq i \leq m$.*

Then $m = n$ and there is a permutation π of $\{1, \ldots, n\}$ such that $U_i \cong T_{\pi(i)}$ for all i.

Proof. We induct on the length of G. We say finite lists of groups $T_1, \ldots, T_m$ and $U_1, \ldots, U_n$ are *equivalent* if $m = n$ and there is a permutation π of $\{1, \ldots, n\}$ such that $U_i \cong T_{\pi(i)}$ for all $1 \leq i \leq n$. In other words, the goal is precisely to prove that the lists of simple factors associated to the two given composition series are equivalent.

If $\ell(G) = 0$ then G is trivial and there is nothing to show. So assume that $\ell(G) = n \geq 1$ and that the theorem holds for all groups H with $\ell(H) < n$.

Suppose first that $H_{m-1} = G_{n-1}$, i.e. that both given composition series of G have the same next to last term. Both $\{1\} \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{n-2} \trianglelefteq G_{n-1}$ and $\{1\} \trianglelefteq H_1 \cdots \trianglelefteq H_{m-2} \trianglelefteq H_{m-1} = G_{n-1}$ are

composition series of G_{n-1}, with $n-1$ and $m-1$ factors, respectively. In particular, $\ell(G_{n-1}) \leq n-1$ and so the induction hypothesis applies, giving $m-1 = n-1$ and hence $m = n$. Moreover, the lists $T_1, \ldots, T_{n-1}$ and $U_1, \ldots, U_{n-1}$ are equivalent. Then since $T_n = G/G_{n-1} = G/H_{n-1} = U_n$ also, we see that $T_1, \ldots, T_n$ and $U_1, \ldots, U_n$ are equivalent lists as well, as desired.

The other case is where $K = H_{m-1} \neq L = G_{n-1}$. Since $K \trianglelefteq G$ and $L \trianglelefteq G$, $KL \trianglelefteq G$. Because G/L is simple and $L \leq KL \trianglelefteq G$, by subgroup correspondence either $KL = L$ or $KL = G$. But if $KL = L$ then $K \subseteq L$. Since G/K is simple and L/K is a proper normal subgroup, this gives $L = K$, a contradiction. Thus $KL = G$. By the second isomorphism theorem, $T_n = G/L = KL/L \cong K/(K \cap L)$ and $U_m = G/K = LK/K \cong L/(K \cap L)$.

Choose any composition series of $K \cap L$, say $\{1\} = N_0 \trianglelefteq N_1 \trianglelefteq N_2 \trianglelefteq \cdots \trianglelefteq N_p = K \cap L$, with simple factors $V_i = N_i/N_{i-1}$ for $1 \leq i \leq p$. Then $\{1\} = N_0 \trianglelefteq N_1 \trianglelefteq N_2 \trianglelefteq \cdots \trianglelefteq N_p = K \cap L \trianglelefteq L$ is a composition series of L with $p+1$ simple factors, $V_1, V_2, \ldots, V_p, L/(K \cap L) \cong U_m$. As in the previous step, $L = G_{n-1}$ also has a composition series $\{1\} \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{n-2} \trianglelefteq G_{n-1}$, so $\ell(G_{n-1}) \leq n-1$, and the induction hypothesis applies. So $p+1 = n-1$ and $p = n-2$. Moreover, $V_1, V_2, \ldots, V_{n-2}, U_m$ is equivalent to $T_1, \ldots, T_{n-1}$. Similarly, $\{1\} = N_0 \trianglelefteq N_1 \trianglelefteq N_2 \trianglelefteq \cdots \trianglelefteq N_{n-2} = K \cap L \trianglelefteq K$ is a composition series of K with the $n-1$ factors $V_1, \ldots, V_{n-2}, T_n$. This shows that $\ell(K) \leq n-1$ and so the induction hypothesis applies to K. Since $\{1\} \trianglelefteq H_1 \cdots \trianglelefteq H_{m-2} \trianglelefteq H_{m-1} = K$ is also a composition series of K, $m-1 = n-1$ and $m = n$. Moreover, $U_1, \ldots, U_{n-1}$ is equivalent to $V_1, V_2, \ldots V_{n-2}, T_n$.

Finally, since $T_1, \ldots, T_{n-1}$ is equivalent to $V_1, V_2, \ldots, V_{n-2}, U_n$, then $T_1, \ldots, T_{n-1}, T_n$ is equivalent to $V_1, V_2, \ldots, V_{n-2}, U_n, T_n$. Similarly, since $U_1, \ldots, U_{n-1}$ is equivalent to $V_1, V_2, \ldots V_{n-2}, T_n$, we have $U_1, \ldots, U_n$ is equivalent to $V_1, V_2, \ldots, V_{n-2}, T_n, U_n$. But obviously $V_1, V_2, \ldots, V_{n-2}, U_n, T_n$ is equivalent to $V_1, V_2, \ldots, V_{n-2}, T_n, U_n$. So $T_1, \ldots, T_n$ and $U_1, \ldots, U_n$ are equivalent as required. □

Example 7.29. In Example 7.27, we saw that $\mathbb{Z}_n$ has many different composition series. As the Jordan-Hölder Theorem predicts, the composition factors are always the groups $\mathbb{Z}_{p_1}, \mathbb{Z}_{p_2}, \ldots, \mathbb{Z}_{p_k}$ in some order, where $p_1, p_2, \ldots, p_k$ are primes whose product is n. In turn this can be used to show that any composition series of $\mathbb{Z}_n$ must be of the form given in Example 7.27, since a cyclic group has a unique subgroup of each order dividing the order of the group.

Example 7.30. A composition series for S_4 is $1 \trianglelefteq \langle (12)(34) \rangle \trianglelefteq V \trianglelefteq A_4 \trianglelefteq S_4$. The group $\langle (12)(34) \rangle$ can be replaced by any of the other order 2 subgroups of V, obtaining a different composition series, but one with the same composition factors $\mathbb{Z}_2, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_2$ (in fact they always occur in this order in this case).

A composition series for a finite group G exhibits the simple groups which are "building blocks" for G. If G has a composition series of length two, for example, then $\{1\} \trianglelefteq G_1 \trianglelefteq G$ where G_1 is simple and G/G_1 is simple. If we could understand all simple finite groups and also understand all extensions of one by another, then we could classify all such groups. Then a group with composition series length 3 is an extension of a simple group by a group of composition series length 2, so if we understand such extensions we could classify such groups as well. In this way via composition series the classification of finite groups reduces to the classification of simple groups and the extension problem.

In fact, as has already been mentioned in these notes, the classification of finite simple groups has been completed, with several well-understood infinite families of examples and a number of "sporadic" simple groups which do not naturally occur in families. The extension problem is still very difficult, and one should not expect to be able to completely classify all groups with a given set of composition factors up to isomorphism, except in special cases. But there are many problems about groups that reduce to showing something holds for the composition factors of a group. Since we know now what the finite simple groups are, this has allowed for new results to be proved about finite groups by checking each of the simple groups in the classification.

Let us also discuss the relationship between composition series and solvable groups. Composition series are subnormal series where the factors are simple, and a solvable group has a subnormal series where the factors are abelian. What if a subnormal series has both properties, i.e. the factors are simple and abelian? In fact simple abelian groups are very special.

Lemma 7.31. *The following are equivalent:*

(1) *G is solvable and simple.*
(2) *G is abelian and simple.*
(3) *G is finite of prime order p.*

Proof. (1) $\implies$ (2): Recall that simple groups are nontrivial. If $G' = G$, then $G^{(i)} = G$ for all i by induction. But by Theorem 7.10, since G is solvable we have $G^{(n)} = \{1\}$ for some n. So G is trivial, a contradiction. Thus G' must be a proper subgroup of G, and we know G' is normal in G. Since G is simple, $G' = \{1\}$. This means that $G = G/G'$, which is abelian by Proposition 7.4.

(2) $\implies$ (3): Since G is abelian, all of its subgroups are normal. Since G is simple, its only normal subgroups are the trivial subgroup and G. So G has only two subgroups, $\{1\}$ and G. Given $g \in G$, either $g = 1$ or else $G = \langle g \rangle$. So G is cyclic, and every nonidentity element of G is a

generator. Since the trivial group is not simple by definition, this is true only when G is finite cyclic of prime order.

(3) $\implies$ (1): A group of prime order p is isomorphic to $\mathbb{Z}_p$, which is obviously solvable and simple. □

This leads to another useful characterization of finite solvable groups.

Theorem 7.32. *If G is a group of finite length, then G is solvable if and only if all composition factors of G have prime order.*

Proof. Note that by the Jordan-Hölder Theorem, whether the composition factors of G have prime order is independent of the choice of composition series.

Suppose that G is solvable. Let $1 = H_0 \trianglelefteq H_1 \trianglelefteq \ldots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$ be a composition series for G. By Proposition 7.11, solvability passes to subgroups and factor groups, so each subgroup H_i is solvable, and then each factor group H_{i+1}/H_i is solvable, as well as simple. Hence each factor is finite of prime order p, by Lemma 7.31.

Conversely, if G has a composition series $1 = H_0 \trianglelefteq H_1 \trianglelefteq \ldots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$ where each factor H_{i+1}/H_i has prime order, then each factor is cyclic and so abelian. Thus this subnormal series shows that G is solvable. □

In particular, the theorem applies to all finite groups, and characterizes which are solvable in terms of their composition factors: only the abelian simple groups $\mathbb{Z}_p$ can occur, no non-abelian simple groups. Also, the theorem implies that a solvable group of finite length must actually be finite.

Example 7.33. In Example 7.30, we see that a composition series for S_4 has factors of prime orders $2, 2, 3, 2$, confirming that this group is solvable. On the other hand, the only possible composition series of S_n for $n \geq 5$ is $\{1\} \trianglelefteq A_n \trianglelefteq S_n$, which has a factor A_n which is not of prime order, confirming that S_n is not solvable.

Example 7.34. Let G be a finite nontrivial p-group for a prime p, so $|G| = p^n$ for some $n \geq 1$. In any composition series for G, the simple factors must be p-groups also. We saw earlier that any p-group has a nontrivial center, so a simple p-group must be abelian and therefore isomorphic to $\mathbb{Z}_p$. Thus every composition factor of G is isomorphic to $\mathbb{Z}_p$ and so G is solvable by Theorem 7.32. In fact we showed earlier that a p-group is even nilpotent, which is stronger than solvable.

Example 7.35. Using the techniques coming from the Sylow theorems, it is straightforward to show that there are no nonabelian simple groups G with $|G| < 60$. In other words, A_5 is the

smallest nonabelian simple group. But then if $|G| < 60$, every simple factor in a composition series for G must be an abelian simple group, so G is solvable. Thus A_5 is also the smallest nonsolvable group.

8. Crash course on rings

In these notes, we also assume the reader has some familiarity with rings from an undergraduate course, so as with groups we review the basic facts quickly. Also, some concepts, such as the isomorphism theorems for rings, are very similar to their group-theoretic counterparts and are easier to digest the second time you see them.

A ring is an object that captures the properties familiar to us from common systems of numbers, such as the integers and real numbers. In particular, a ring has both an addition and multiplication operation which satisfy some basic compatibilities. As we will see, however, this definition is general enough to apply to systems of "numbers" far removed from the original examples.

8.1. **Basic definitions and examples.**

Definition 8.1. A *ring* is a set R with two binary operations $+$ and $\cdot$ (called addition and multiplication, respectively) with the following properties:

(1) R is an abelian group under $+$. The identity element is called 0 and the additive inverse of a is written $-a$.
(2) R is a monoid under $\cdot$; that is, $\cdot$ is an associative operation with identity element called 1, where $a \cdot 1 = a = 1 \cdot a$ for all $a \in R$. The element 1 is also called the *unit* of the ring.
(3) The addition and multiplication are related by the two *distributive laws*:
   (a) $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in R$
   (b) $(b + c) \cdot a = b \cdot a + c \cdot a$ for all $a, b, c \in R$.

If $a \cdot b = b \cdot a$ for all $a, b \in R$, the ring R is called *commutative*; otherwise it is *noncommutative*.

Usually when the context is clear one simply writes the product $a \cdot b$ as ab. Historically, rings were often defined without the assumption of an identity element 1 for multiplication, that is, R with its operation $\cdot$ was only assumed to be a semigroup. However, the more modern convention is to include the existence of 1 as part of the main definition, as we have done. An object that satisfies all of the axioms except for the existence of 1 is called a *ring without identity* or *ring without unit*. (Nathan Jacobson introduced the amusing term "rng" for a ring without identity in his well-known algebra text, but it didn't catch on.) Occasionally it is useful to work with a ring without unit but we will seldom encounter such rings in this course.

Because of the distributive laws, the identity element 0 for addition also has special properties with regard to multiplication. If $a \in R$ for a ring R, then $0a = (0+0)a = 0a + 0a$. Since $0a$ has an additive inverse $-(0a)$, adding it to both sides gives $0 = 0a$. Similarly, $0 = a0$. Other easy consequences of the definition are in the following exercise.

Exercise 8.2. Show the following for any a, b in a ring R:

(1) (-a) b = -(ab) = a(-b).

(2) a(-1) = -a = (-1) a.

(3) (-a)(-b) = ab.

Some simple examples of rings are given as follows. We generally will leave the routine verifications of the ring axioms to the reader.

Example 8.3. The familiar number systems of $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ are all rings under the usual operations. Note that the natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$ do not form a ring, as additive inverses do not exist for the positive numbers in $\mathbb{N}$.

Example 8.4. The subset $2\mathbb{Z}$ of even integers in $\mathbb{Z}$, under the usual addition and multiplication, is a ring without identity.

Example 8.5. The one-element set $R = \{0\}$, with the only possible operations $0 + 0 = 0$ and $00 = 0$, is a ring, called the *trivial* or *zero* ring. Obviously 0 must serve as both the additive and multiplicative identity, so $0 = 1$.

Conversely, suppose that R is a ring whose multiplicative and additive identities coincide. Then for any $r \in R$ we have $r = 1r = 0r = 0$, so that $R = \{0\}$ is the zero ring.

The zero ring is obviously uninteresting. It sometimes needs to be excluded from theorem statements to make them strictly true, but hopefully the reader will forgive the author if he forgets to do that.

Example 8.6. For any integer $n \geq 1$, the set $\mathbb{Z}_n$ of congruence classes modulo n, with the usual addition and multiplication of congruence classes, is a ring. Usually we take $n \geq 2$, since when $n = 1$ we obtain the zero ring. We can think of $\mathbb{Z}_n$ as the factor group $\mathbb{Z}/n\mathbb{Z}$ under addition, and we write the coset $a + n\mathbb{Z}$ as $\overline{a}$. Then of course $\overline{a} + \overline{b} = \overline{a+b}$, and the multiplication in $\mathbb{Z}_n$ is given by $\overline{a}\,\overline{b} = \overline{ab}$.

All of the examples so far are commutative rings. One learns in a first course in linear algebra that matrix multiplication is not commutative, and in fact rings of matrices are among the simplest examples of noncommutative rings.

Example 8.7. Let R be a ring, for example any of the familiar number systems in Example 8.3, and let $n \geq 1$. We form a new ring $S = M_n(R)$ whose elements are formal $n \times n$ matrices with entries in the ring R. Write an element of S as (r_{ij}) where $r_{ij} \in R$ is in the (i, j)-position of the matrix (that is, row i and column j). We define an addition and multiplication on S in the usual way for matrices. More specifically, addition is done coordinatewise, so $(r_{ij}) + (s_{ij}) = (r_{ij} + s_{ij})$, and the product $(r_{ij})(s_{ij})$ is the matrix (t_{ij}) with $t_{ij} = \sum_{k=1}^{n} r_{ik}s_{kj}$. The identity matrix with 1's along the main diagonal and 0's elsewhere is a unit element for S. Since R is a ring, it is routine to see that S is again a ring.

As long as $n \geq 2$, it is easy to find matrices $A, B \in M_n(R)$ such that $AB \neq BA$, so $M_n(R)$ is a noncommutative ring. (Here you must exclude the case where R is the zero ring, for which $M_n(R)$ is also the zero ring. We will not keep mentioning it.)

There are various other constructions which, like matrix rings, produce new rings from a given ring or rings. Here are some further examples.

Example 8.8. Let $\{R_\alpha | \alpha \in A\}$ be an indexed collection of rings. The *direct product* is the ring $\prod_{\alpha \in A} R_\alpha$, that is, the Cartesian product of these sets, is a ring with coordinatewise operations. In other words, if we write an element of this ring as (r_α), where $r_\alpha \in R_\alpha$ is the element in the α-coordinate, then $(r_\alpha) + (s_\alpha) = (r_\alpha + s_\alpha)$ and $(r_\alpha)(s_\alpha) = (r_\alpha s_\alpha)$. Note that as groups under $+$, this is just the direct product of the abelian groups $(R_\alpha, +)$. If R_α has additive identity 0_α and multiplicative identity 1_α, then the elements (0_α) and (1_α) are the additive identity and multiplicative identity of the product.

Example 8.9. Let R be any ring. We define the *ring of power series* $R[[x]]$ in an indeterminate x to be the set of all formal sums $\{a_0 + a_1x + a_2x^2 + \cdots + a_mx^m + \ldots \mid a_i \in R\}$. Note that no convergence is expected or implied, and we don't try to think of these as functions in the variable x; an element of $R[[x]]$ is simply determined by the countable sequence of coefficients $(a_0, a_1, a_2, a_3, \ldots)$, and the powers of x can be viewed as placeholders to help explain the multiplication rule. Formally as an abelian group we can identify R with $\prod_{i=0}^{\infty} R$, the product of a countable number of copies of R.

We write an element of $R[[x]]$ as $\sum_{n=0}^{\infty} a_n x^n$. The addition and multiplication are as expected for power series; namely, $(\sum a_n x^n) + (\sum b_n x^n) = \sum (a_n + b_n)x^n$, and

$$(\sum a_n x^n)(\sum b_n x^n) = \sum_{n=0}^{\infty} [\sum_{i=0}^{n} a_i b_{n-i}] x^n$$

(note that only finite sums of elements in R are needed to define each coefficient of the product).

Example 8.10. Actually more important for us than the ring of power series is the *polynomial ring* $R[x]$, which is the subset of $R[[x]]$ consisting of elements $\sum a_n x^n$ such that $a_n = 0$ for all $n > m$, some m. Thus a typical element is a formal polynomial $a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$ with $a_i \in R$. As an abelian group, we can identify $R[x]$ with the direct sum $\bigoplus_{n=0}^{\infty} R$ of a countable number of copies of R. (the direct sum of a set of abelian groups was also called the *restricted product* earlier). $R[x]$ is is a ring under the same operations as for the power series ring restricted to this subset, in other words $R[x]$ is a subring of $R[[x]]$ in the sense to be defined soon.

The next example gives an interesting link between group theory and ring theory.

Example 8.11. Let G be a group and let R be a ring. The *group ring* RG consists of finite formal sums of elements in G with coefficients in R. We can write any such formal sum as $\sum_{g \in G} r_g\, g$, where $r_g \in R$ and $r_g = 0$ for all but finitely many g; in other words $RG \cong \bigoplus_{g \in G} R$ as Abelian groups.

The addition operation simply adds like coefficients: $\sum r_g\, g + \sum s_g\, g = \sum (r_g + s_g)\, g$. The multiplication operation is defined on elements with one term using the group structure of G, so $(rg)(sh) = (rs)(gh)$, where rs is the product in R and gh is the product in G. This is then extended linearly to define a product on finite sums, so

$$(\sum r_g\, g)(\sum s_g\, g) = \sum_{g \in G} [\sum_{h \in G} r_h s_{h^{-1}g}]\, g.$$

The identity element of RG is $1_R 1_G$.

For a finite group G, studying the group ring FG over a field F gives a surprisingly powerful tool for understanding better the properties of G; in particular, the structure of this group ring is directly related to the *representation theory* of the group G over F. For simplicity consider the case of group rings over $\mathbb{C}$. If G is a finite group, then it turns out the $\mathbb{C}G$ is isomorphic as a ring to a direct product of finitely many matrix rings over $\mathbb{C}$ (we will review isomorphism of rings in the next section). More specifically, $\mathbb{C}G \cong M_{n_1}(\mathbb{C}) \times \cdots \times M_{n_s}(\mathbb{C})$, where the number of factors s is equal to the number of conjugacy classes of G, and the numbers $n_1, \ldots, n_s$ are the dimensions of the distinct irreducible representations of G. You can find more information in Chapter 18 of Dummit and Foote.

8.2. **Zero-divisors and units.** The standard rings of numbers such as $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ which one uses in calculus have some special properties which are not satisfied by arbitrary rings. First, in a general ring one can have $ab = 0$ even if a and b are not 0.

Definition 8.12. Let R be a ring. If $a, b \in R$ are elements with $a \neq 0$ and $b \neq 0$ but $ab = 0$, then a and b are called *zero-divisors*. Notice that by definition a zero-divisor is nonzero. A ring R with no zero-divisors is called a *domain*. A commutative domain is often called an *integral domain* for historical reasons, since among the rings studied extensively were certain (commutative) rings important in number theory which are so-called "rings of integers" in a number field.

Note that the rings of numbers in Example 8.3 are all integral domains. We can ask what the zero-divisors are in some of our other examples so far.

Example 8.13. The ring $\mathbb{Z}_n$ of integers mod n is an integral domain if and only if n is prime. For if n is not prime, then $n = mk$ with $1 < m < n$ and $1 < k < n$; thus $\overline{m} \neq \overline{0}$ and $\overline{k} \neq \overline{0}$; however $\overline{m}\overline{k} = \overline{n} = \overline{0}$.

Conversely, if n is a prime p, then if $\overline{a}\overline{b} = \overline{0}$ we get that p divides ab, and so either p divides a or p divides b by Euclid's Lemma. Thus $\overline{a} = \overline{0}$ or $\overline{b} = \overline{0}$.

The other special property that number systems like $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ have is the ability to divide a number by any other nonzero number. Formally, this is the property that all nonzero numbers have multiplicative inverses, as in the following definition.

Definition 8.14. Let R be a ring. An element $a \in R$ is a *unit* if there is $b \in R$ such that $ab = 1 = ba$; there is clearly a unique such b if it exists. The element b is called the *inverse* of a and one writes $b = a^{-1}$.

Note that a unit in a ring cannot be a zero-divisor; for if $ac = 0$ and also a is a unit, then $c = a^{-1}ac = a^{-1}0 = 0$; similarly, $ca = 0$ forces $c = 0$. The set $R^\times$ of all units in a ring is easily seen to be a group under the multiplication operation of the ring. (This is a special case of Lemma 1.5, which showed that the set of invertible elements in any monoid is a group.) $R^\times$ is called the *units group* of R. Another common notation for this group is $U(R)$.

Definition 8.15. A ring R is a *division ring* if $R^\times = R - \{0\}$, that is, every nonzero element is a unit. A commutative division ring is called a *field*. (An older term for division ring is *skew field*.) By convention the zero ring is not considered a field.

Example 8.16. $\mathbb{Z}^\times = \{-1, 1\}$, while $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are fields.

Example 8.17. Let F be any field, so we can apply results in linear algebra to the matrix ring $M_n(F)$. It is easy to see that a nonzero matrix A is a zero-divisor if and only if it is singular, i.e.

has a nonzero nullspace. (If $Av = 0$ for some nonzero column vector v, let B be any nonzero matrix whose columns are all multiples of v; then $AB = 0$.) By theorems in linear algebra, A is singular if and only if $\det A = 0$.

Example 8.18. The units in $\mathbb{Z}_n$ are $\mathbb{Z}_n^\times = \{\overline{a} | \gcd(a, n) = 1\}$. This was shown earlier in Example 1.8.

In particular, when $n = p$ is a prime number, then $\mathbb{Z}_p$ is a field, since $\mathbb{Z}_p^\times = \mathbb{Z}_p - \{\overline{0}\}$. This field is also written as $\mathbb{F}_p$.

Division rings which are not fields exist in abundance, but it is less obvious how to construct examples. The ring of quaternions $\mathbb{H}$, discovered by William Rowan Hamilton in 1843, was the first such example.

Example 8.19. Let $\mathbb{H}$ be a 4-dimensional vector space over $\mathbb{R}$ with basis $1, i, j, k$. We define a product on these 4 symbols, where $1x = x = x1$ for $x \in \{i, j, k\}$; $ij = k = -ji$; $jk = i = -kj$, $ki = j = -ik$, and $i^2 = j^2 = k^2 = -1$. This product is extended $\mathbb{R}$-linearly to give a product on all of $\mathbb{H}$; an easy calculation shows that the product is associative on the basis $\{1, i, j, k\}$, which implies that the product is associative on all of $\mathbb{H}$. We leave the verification that $\mathbb{H}$ is a division ring to Exercise 8.31.

Note that $\mathbb{H}$ contains the subset $\{\pm 1, \pm i, \pm j, \pm k\}$ which is isomorphic to the quaternion group Q_8 under multiplication; this is how the quaternion group got its name.

Example 8.20. If F is a field, then the units in $M_n(F)$ are exactly the invertible matrices by definition. In other words, the units group $(M_n(F))^\times$ is the general linear group $\mathrm{GL}_n(F)$. By results in linear algebra one knows that any matrix is either invertible A (if $\det A \neq 0$) or singular (if $\det A = 0$). Since we noted above that the singular nonzero matrices are zero-divisors, every nonzero element in $M_n(F)$ is either a zero-divisor or a unit.

Figuring out which elements are zero-divisors, and which are units, can be surprisingly complicated even for rings which are easy to define. Let us give some more examples.

Example 8.21. Let $S = \prod_\alpha R_\alpha$. The units in S are the (r_α) such that r_α is a unit in R_α for all α. An element (r_α) of S is a zero-divisor if and only if at least one of the coordinates r_α is either 0 or a zero-divisor in R_α, but not all of the coordinates are 0. Thus as long as S is a product of at least 2 nonzero rings, then S is not a domain.

An element $r \in R$ of a ring is *nilpotent* if there exists $n \geq 1$ such that $r^n = 0$.

Example 8.22. Let R be a commutative ring and let $S = R[x]$. An element $\sum_{i=0}^{m} a_i x^i$ is a unit in S if and only if a_0 is a unit in R and $a_1, \ldots, a_m$ are nilpotent in R. This is most easily proved after we have seen a bit more theory (see Exercise 9.10). *McCoy's Theorem* states that $\sum_{i=0}^{m} a_i x^i$ is a zero-divisor in R if and only if there is $b \neq 0$ in R such that $a_i b = 0$ for $0 \leq i \leq m$ (Exercise 8.30).

Example 8.23. Let R be a commutative ring and let $S = R[[x]]$ be a power series ring over R. An element $\sum_{i=0}^{\infty} a_i x^i$ is a unit in S if and only if a_0 is a unit in R (see Exercise 8.27). The classification of zero-divisors is apparently not known in complete generality, though if R is a *Noetherian* ring (as we will define later), the analog of McCoy's Theorem holds here (i.e. if $\sum_{i=0}^{\infty} a_i x^i$ is a zerodivisor, then there exists $b \neq 0$ in R such that $a_i b = 0$ for all $i \geq 0$.)

Example 8.24. As mentioned earlier, if G is a finite group, then there is an isomorphism $\phi : \mathbb{C}G \to M_{n_1}(\mathbb{C}) \times \ldots M_{n_s}(\mathbb{C})$ for some integers $n_1, \ldots, n_s$. If one finds this isomorphism explicitly, one could then determine the units and zerodivisors of $\mathbb{C}G$ explicitly because this problem is solved in the ring $M_{n_1}(\mathbb{C}) \times \ldots M_{n_s}(\mathbb{C})$. Namely, if $(A_1, \ldots, A_s)$ is an element of the latter ring, it is a unit if and only if each A_i is an invertible matrix in $M_{n_i}(\mathbb{C})$, and it is a zerodivisor if at least one A_i is singular (but not all A_i are 0). Exercise 8.59 shows how to find the isomorphism ϕ explicitly when G is finite cyclic.

On the other hand, for an arbitrary group G and an arbitrary ring R, the structure of the units and zerodivisors of the group ring RG is a very complicated subject about which there are still many open questions. This is true even if F is a field. For example, Kaplansky's unit conjecture asks if F is a field and G is a (necessarily infinite) group in which all nonidentity elements have infinite order, is every unit of FG of the form ag for some $0 \neq a \in F$ and $g \in G$? A counterexample to this long-standing conjecture was apparently found by Giles Gardam and announced just in 2021.

One thing that is elementary to see here is the fact that if R is a domain, so are $R[x]$ and $R[[x]]$. Thus the formation of polynomial or power series rings does not "create" zero-divisors. Let us concentrate on $R[x]$; we leave the case of $R[[x]]$ as an exercise. For any $0 \neq f \in R[x]$, we can write f as $a_0 + a_1 x + \cdots + a_m x^m$, where $a_m \neq 0$; thus x^m is the largest power of x to occur with nonzero coefficient. Then we call m the *degree* of f and write $\deg(f) = m$. This definition doesn't make sense for the zero-polynomial (where $a_i = 0$ for all i) and by convention we set $\deg(0) = -\infty$.

Lemma 8.25. *Let R be a domain.*

(1) *If $f, g \in R[x]$ then* $\deg(fg) = \deg(f) + \deg(g)$.

(2) *$R[x]$ is a domain.*

Proof. (1) Suppose first that f and g are both nonzero. If $f = \sum_{i=0}^{m} a_i x^i$ and $g = \sum_{i=0}^{n} b_i x^i$ with $a_m \neq 0$, $b_n \neq 0$, then by the definition of multiplication we have $fg = \sum_{i=0}^{m+n} (\sum_{j=0}^{i} a_j b_{i-j}) x^i$ which clearly has degree at most $m+n$; the coefficient of x^{n+m} is $a_m b_n$, which is nonzero since R is a domain. Thus $\deg(fg) = \deg(f) + \deg(g)$. If either f or g is 0, then $fg = 0$, and in this case the result holds with the conventions that $-\infty + n = -\infty$ for any number n, and $-\infty + -\infty = -\infty$.

(2) If $f, g \in R[x]$ with $f \neq 0$, $g \neq 0$, and therefore $\deg(f) \geq 0$ and $\deg(g) \geq 0$, by (1) we have $\deg(fg) \geq 0$. In particular $\deg(fg) \neq -\infty$ and so $fg \neq 0$. □

8.2.1. *Exercises.*

Exercise 8.26. Let R be a commutative ring, and consider the ring $R[[x]]$ of formal power series in one variable. Prove that if R is a domain then $R[[x]]$ is a domain.

Exercise 8.27. Let R be a commutative ring. Prove that $\sum_{n=0}^{\infty} a_n x^n$ is a unit in the ring $R[[x]]$ if and only if a_0 is a unit in R.

Exercise 8.28. Recall that the *center* of a ring R is

$$Z(R) = \{r \in R | rs = sr \text{ for all } s \in R\}.$$

Now let R be any commutative ring, and G any finite group. Consider the group ring RG.

(a). Suppose that $\mathcal{K} = \{k_1, \ldots, k_m\}$ is a conjugacy class in the group G. Prove that the element $K = k_1 + k_2 + \cdots + k_m \in RG$ is an element of $Z(RG)$.

(b). Let $\mathcal{K}_1, \ldots, \mathcal{K}_r$ be the distinct conjugacy classes in G and for each i let K_i be the sum of the elements in $\mathcal{K}_i$, as in part (a). Prove that $Z(RG) = \{a_1 K_1 + \cdots + a_r K_r | a_i \in R \text{ for all } 1 \leq i \leq r\}$. In other words, the center consists of all R-linear combinations of the K_i.

Exercise 8.29. Let R be a commutative ring. Suppose that x is nilpotent and u is a unit in R. Show that $u - x$ is a unit in R.

(Hint: reduce to the case that $u = 1$. Note that $(1-x)(1 + x + x^2 + \cdots + x^{m-1}) = 1 - x^m$.)

Exercise 8.30. Prove *McCoy's Theorem*: If $f = a_0 + a_1 x + \cdots + a_m x^m \in R[x]$ for a commutative ring R and f is a zero-divisor in $R[x]$, then there exists $0 \neq b \in R$ such that $ba_i = 0$ for all $0 \leq i \leq m$. (Hint: assume that $a_m \neq 0$ and let $0 \neq g \in R[x]$ be of minimal degree such that $fg = 0$. Write $g = b_0 + b_1 x + \cdots + b_n x^n$ with $b_n \neq 0$. Suppose that $a_i g = 0$ for all i; then $a_i b_j = 0$ for all i, j and so $b_n f = 0$ and we are done. Thus some $a_i g \neq 0$ and we can take j maximal such that $a_j g \neq 0$. Then $f(a_j g) = 0$ but $\deg(a_j g) < \deg g$.)

Exercise 8.31. Let $\mathbb{H}$ be the ring of Hamilton's quaternions as in Example 8.19.

(a). Define the *conjugate* of $x = a + bi + cj + dk$ to be $\overline{x} = a - bi - cj - dk$. Define $N(x) = x\overline{x}$. Show that $N(x) = a^2 + b^2 + c^2 + d^2 \in \mathbb{R}$.

(b). Use part (a) to show that any nonzero element of $\mathbb{H}$ is a unit; thus $\mathbb{H}$ is a division ring.

(c). Show that for $x, y \in \mathbb{H}$ we have $\overline{xy} = \overline{y}\ \overline{x}$. Using this, show that $N(xy) = N(x)N(y)$.

(d). An element of the form $x = bi + cj + dk$ is called a *pure quaternion*. Show that such an x satisfies $x^2 = -1$ if and only if $N(x) = 1$. Conclude that -1 has uncountably many square roots in $\mathbb{H}$.

8.3. **Subrings, ideals, factor rings, and homomorphisms.** Similarly as in group theory (and as for many other algebraic structures) we have notions of homomorphisms of rings, subrings, factor rings, isomorphism theorems, and so on. We now review the definitions of these basic concepts.

Definition 8.32. Let S be a ring. A subset R of S is a *subring* if R is itself a ring under the same operations as S, and with the same unit element. Explicitly, this is the same as requiring that R is closed under subtraction and multiplication in S, and $1_S \in R$.

Example 8.33. $\mathbb{Z}$ is a subring of $\mathbb{Q}$; similarly, $\mathbb{Q}$ is a subring of $\mathbb{R}$ and $\mathbb{R}$ is a subring of $\mathbb{C}$.

Example 8.34. If R is a ring and G is a group, then for any subgroup H of G the group ring RH is a subring of the group ring RG. If R is a subring of a ring S, then the group ring RG is a subring of the group ring SG.

Example 8.35. In the polynomial ring $R[x]$, the set of constant polynomials is a subring. A similar comment holds for the power series ring $R[[x]]$. In each case we can identify this subring with R and think of $R \subseteq R[x]$ and $R \subseteq R[[x]]$.

Example 8.36. In $M_n(R)$, the subsets of diagonal matrices, upper triangular matrices, and lower triangular matrices are all subrings of $M_n(R)$.

It is possible to have a subset R of a ring S such that R is a ring under the same operations as S, but with a different unit element. In this case we say that R is a *non-unital* subring of S.

Example 8.37. Let $S = M_2(R)$ be the ring of 2 by 2 matrices over a ring R. The subset $T = \{\left(\begin{smallmatrix} r & 0 \\ 0 & 0 \end{smallmatrix}\right) | r \in R\}$ is closed under subtraction and multiplication in S, and has a unit element $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ different from the unit element $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ of S (the identity matrix).

Non-unital subrings are occasionally useful, but it is good to point it out whenever one is allowing this weaker definition of subring.

Definition 8.38. If R and S are rings, a function $\phi : R \to S$ is a *homomomorphism* (of rings) if

(1) ϕ is a homomorphism of additive groups; that is, $\phi(a+b) = \phi(a) + \phi(b)$ for all $a, b \in R$;
(2) $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in R$; and
(3) $\phi(1_R) = 1_S$.

As usual, a bijective homomorphism is called an *isomorphism*, and an isomorphism from a ring R to itself is called an *automorphism*. If there exists an isomorphism from R to S we write $R \cong S$ and say that R and S are isomorphic.

Note that a homomorphism of groups always sends the identity to the identity, and this does not have to be made part of the definition—thus, for example, $\phi(0_R) = 0_S$ holds for a homomorphism of rings as above, without being specified. On the other hand, a ring is not a group under multiplication, so preserving the product, as in condition (2), does not imply condition (3). A function which satisfies conditions (1) and (2) but not necessarily (3) is called a *non-unital* homomorphism. The inclusion map of a non-unital subring R into a ring S is an example of a non-unital homomorphism. Similarly as for non-unital surbrings, the modern consensus seems to be that it is easiest to include unitality in the definition of homomorphism, and explicitly point out whenever a homomorphism is non-unital.

Example 8.39. The natural inclusion $\phi : \mathbb{Z} \to \mathbb{Q}$ is a ring homomorphism; similarly for the inclusions $\mathbb{Q} \to \mathbb{R}$ and $\mathbb{R} \to \mathbb{C}$.

Example 8.40. If R is a ring and G is a group, there is a surjective homomorphism $\rho : RG \to R$ given by $\rho(\sum_{g \in G} a_g g) = \sum_{g \in G} a_g$.

Example 8.41. Let R be a commutative ring which is a subring of a commutative ring S. For any $s \in S$, there is a homomorphism $\phi : R[x] \to S$ defined by *evaluation at* s: $\phi(\sum_{i=0}^{m} a_i x^m) = \sum_{i=0}^{m} a_i s^m$. To see why we might want to evaluate at an element in a bigger ring than R, we might, for example, want to evaluate a polynomial with real coefficients at a complex number.

Example 8.42. If R and S are rings, let $T = R \times S$ be the direct product. There are two surjective ring homomorphisms $\pi_1 : R \times S \to R$ with $\pi_1(r, s) = r$ and $\pi_2 : R \times S \to S$ with $\pi_2(r, s) = s$, called the projection maps. We also have the obvious inclusion maps $i_1 : R \to R \times S$ with $i_1(r) = (r, 0)$ and $i_2 : S \to R \times S$ with $i_2(s) = (0, s)$. Note, however, that i_1 and i_2 are only non-unital ring homomorphisms, as the identity of $R \times S$ is $(1, 1)$, which is not equal to $i_1(1) = (1, 0)$ or $i_2(1) = (0, 1)$.

Example 8.43. Consider a cyclic group $G = \{1, a\}$ of order 2. We claim that $\mathbb{C}G \cong \mathbb{C} \otimes \mathbb{C}$, that is, that we have a direct product of two 1×1 matrix rings. This is a (very) special case of the fact mentioned earlier, that $\mathbb{C}G$ is isomorphic to a direct product of matrix rings over $\mathbb{C}$ for any finite group G.

Note that the ring $\mathbb{C}\otimes\mathbb{C}$ has two special elements $e_1 = (1, 0)$ and $e_2 = (0, 1)$ which are *idempotent* in the sense that $e_1^2 = e_1$ and $e_2^2 = e_2$. They are the unit elements of the non-unital subrings which are the images of the maps i_1 and i_2 as in the previous example. Moreover $e_1 + e_2 = (1, 1)$, the multiplicative identity element. Thus if we seek a ring isomorphism $\phi : \mathbb{C} \otimes \mathbb{C} \to \mathbb{C}G$, Then $\phi(e_1)$ and $\phi(e_2)$ should be idempotents in $\mathbb{C}G$ whose sum is 1. A short calculation shows that $f_1 = \frac{1}{2}(1+a)$ and $f_2 = \frac{1}{2}(1-a)$ are the only idempotents in $\mathbb{C}G$ besides 0 and 1. It is easy to check that defining ϕ on a $\mathbb{C}$-basis by $\phi(e_i) = f_i$ for $i = 1, 2$ and extending linearly gives an isomorphism of rings.

The definitions of kernel, image, and factor ring, are built on the definitions for the underlying abelian groups.

Definition 8.44. Let $\phi : R \to S$ be a homomorphism of rings. The *image* of ϕ is $\phi(R)$, and the *kernel* of ϕ is $\ker \phi = \{r \in R | \phi(r) = 0\}$.

Definition 8.45. If R is a ring, a *left ideal* of R is a subset $I \subseteq R$ such that

(1) I is a subgroup of R under $+$.
(2) For all $r \in R$, $x \in I$, $rx \in I$.

A *right ideal* of R is defined similarly, replacing condition (2) by the condition that for all $r \in R$ and $x \in I$, $xr \in I$. Finally I, is an *ideal* of R if it is both a left and right ideal, or equivalently if for all $r, s \in R$ and $x \in I$, $rxs \in I$.

Condition (2) in the definition of left ideal does not look similar to anything we saw in group theory; the reason is that R is only a monoid under multiplication, not a group. Note that in a commutative ring, there is no distinction between left ideals, right ideals, and ideals, so one only refers to ideals.

Example 8.46. Let R be a ring and let $S = M_2(R)$. The subset $J = \{\left(\begin{smallmatrix} r & s \\ 0 & 0 \end{smallmatrix}\right) | r, s \in R\}$ is a right ideal of S, but not a left ideal. Similarly, $K = \{\left(\begin{smallmatrix} r & 0 \\ s & 0 \end{smallmatrix}\right) | r, s \in R\}$ is a left but not right ideal.

Example 8.47. If I and J are ideals of a ring R, then so is $I + J = \{x + y | x \in I, y \in J\}$. It is the smallest ideal containing I and J. Similarly, for any set of ideals $\{I_\alpha | \alpha \in A\}$ we can define its sum

as

$$\sum_{\alpha \in A} I_\alpha = \{\sum_\alpha x_\alpha | x_\alpha \in I_\alpha \text{ and only finitely many of the } x_\alpha \text{ are nonzero}\},$$

which is also an ideal. Note here that while only finite sums are defined in a ring, the convention is often used that an infinite sum of elements may be written if all but finitely many of the elements are 0; the sum is defined to be the sum of the finitely many nonzero elements.

The intersection $I \cap J$ is also an ideal, and is the largest ideal contained in I and J. Similarly, the intersection of any set of ideals in R is again an ideal.

Example 8.48. In any ring R, $\{0\}$ is an ideal, called the *zero ideal* for obvious reasons. We usually just write it as 0. Similarly, R itself is an ideal, often called the *unit ideal* because any ideal I which contains a unit is equal to R. (check!)

Example 8.49. We have seen that the additive subgroups of $\mathbb{Z}$ are all of the form $m\mathbb{Z}$ for $m \geq 0$; in fact these are ideals of $\mathbb{Z}$ as a ring, also. Since any ideal must be an additive subgroup, these are all of the ideals of the ring $\mathbb{Z}$.

Ideals of a ring can be seen as analogous to *normal* subgroups of a group, in the sense that they are exactly the structures we can mod out by to get a factor ring. We will see why left and right ideals are useful when we study module theory later.

Lemma 8.50. *Let R be a ring with ideal I. Let R/I be the factor group of $(R,+)$ by its subgroup $(I,+)$. Thus $R/I = \{r + I | r \in R\}$ is the set of additive cosets of I, with addition operation $(r+I)+(s+I) = (r+s)+I$. Then R/I is also a ring, with multiplication $(r+I)(s+I) = rs+I$ and unit element $1+I$. The surjective map $\phi : R \to R/I$ given by $\phi(r) = r+I$ is a homomorphism of rings.*

Proof. The main issue is to make sure the claimed multiplication rule is well defined. Let $r + I = r' + I$ and $s + I = s' + I$, so $r - r' \in I$ and $s - s' \in I$. Then $rs - r's' = r(s - s') + (r - r')s' \in I$ (note that we use that I is closed under both left and right multiplication by elements in R) and so $rs + I = r's' + I$. Having shown the multiplication is well defined, the ring axioms for R/I follow immediately from the axioms for R, and the fact that ϕ is a homomorphism follows directly from the definition. □

Example 8.51. For any $m \geq 1$, the factor ring $\mathbb{Z}/m\mathbb{Z}$ can be identified with the ring $\mathbb{Z}_m$ of congruence classes modulo m, with the usual addition and multiplication.

The isomorphism theorems for rings are very similar to their group-theoretic counterparts. Here is the 1st isomorphism theorem.

Theorem 8.52. *Let $\phi : R \to S$ be a homomorphism of rings. Then $I = \ker \phi$ is an ideal of R, $\phi(R)$ is a subring of S, and there is an isomorphism of rings $\overline{\phi} : R/I \to \phi(S)$ defined by $\overline{\phi}(r + I) = \phi(r)$.*

Proof. Since ϕ is a homomorphism of additive groups, the 1st isomorphism theorem for groups gives that I is a subgroup of R under $+$, $\phi(R)$ is a subgroup of S under $+$, and $\overline{\phi}$ is a well-defined isomorphism of additive groups. To check that I is an ideal, simply note that for $r, s \in R$, $x \in I$, we have $\phi(rxs) = \phi(r)\phi(x)\phi(s) = \phi(r)0\phi(s) = 0$, so $rxs \in I$. It is trivial to see that $\phi(R)$ is closed under multiplication in S and contains 1_S, and that $\overline{\phi}$ is a homomorphism of rings. □

Example 8.53. If I is an ideal of R, there is a homomorphism $\phi : M_n(R) \to M_n(R/I)$ given by $\phi((r_{ij})) = (r_{ij} + I)$. It is easy to see that the kernel is $M_n(I) = \{(r_{ij}) | r_{ij} \in I \text{ for all } i, j\}$ and that ϕ is surjective, so that the first isomorphism theorem gives $M_n(R)/M_n(I) \cong M_n(R/I)$.

Example 8.54. Let R be a ring with ideal I. Similarly as in the previous example, $I[x] = \{a_0 + a_1x + \cdots + a_mx^m | a_i \in I \text{ for all } i\}$ is an ideal of $R[x]$, and $R[x]/I[x] \cong (R/I)[x]$.

Example 8.55. Let R be commutative and let $\phi : R[x] \to R$ be evaluation at 0, so that we have $\phi(a_0 + a_1x + \cdots + a_mx^m) = a_0$. Then $I = \ker \phi$ consists of all polynomials with 0 constant term, and this is an ideal of $R[x]$. It is easy to see that ϕ is surjective, so that $R[x]/I \cong R$. Note that the polynomials with 0 constant term are exactly those that can have an x factored out, so $I = \{xf(x) | f(x) \in R[x]\}$, which we also write as $xR[x]$.

Recall that since a ring R is an abelian group under addition, using additive notation we write $nr = \overbrace{r + r + \cdots + r}^{n}$ for the sum of n copies of r in R, when $n \geq 1$; we also set $0r = 0$, and let $(-n)r = -nr$ for $n \geq 1$, so nr is defined for all $n \in \mathbb{Z}$. These multiples of r are the additive versions of the powers of an element, and instead of rules for exponents we have the rules for multiples: $m(nr) = (mn)r$, $(m + n)r = mr + nr$, for $m, n \in \mathbb{Z}$ and $r \in R$.

Let R be a ring. Let $\phi : \mathbb{Z} \to R$ be defined by $\phi(n) = n(1)$, i.e. the nth multiple of the unit $1 \in R$. It is easy to check that ϕ is a homomorphism of rings using the rules for multiples. Let $I = \ker \phi$; since this is an ideal of $\mathbb{Z}$, it has the form $I = m\mathbb{Z}$ for a unique $m \geq 0$. We call m the *characteristic* of the ring R and write $\operatorname{char} R = m$. Thus if $m > 0$, then m is the least positive integer such that $m(1) = 0$, in other words the additive order of 1 in the group $(R, +)$. Note that the case $m = 1$ occurs if and only if R is the zero ring. When $m = 0$, then $I = 0$ and this is the only

case in which ϕ is injective. The 1st isomorphism theorem implies that $\mathbb{Z}/m\mathbb{Z} \cong \phi(\mathbb{Z})$. Thus when $m \geq 1$ then R contains a canonical copy of $\mathbb{Z}_m$ as a subring, where $m = \operatorname{char} R$. When $m = 0$, R contains a copy of $\mathbb{Z}$.

The characteristic of a ring is an important notion. In general, rings with positive characteristic may behave in quite different ways than rings with characteristic 0—we will see this especially when we study fields later on. Note that all of the traditional rings of numbers such as $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ have characteristic 0. Here is another basic fact about the characteristic.

Lemma 8.56. *Let R be a nonzero domain. Then* $\operatorname{char} R = 0$ *or* $\operatorname{char} R = p$ *is a prime number.*

Proof. Supose that $p = \operatorname{char} R > 0$. Then R contains a subring isomorphic to $\mathbb{Z}_p$, namely the additive subgroup generated by 1, by the above discussion. Since R is a domain, so is $\mathbb{Z}_p$. We have seen this forces p to be prime in Example 8.13. □

Remark 8.57. There is sometimes confusion between ideals and subrings of a ring. In group theory, subgroups are the substructures that are themselves groups, while the substructures that one can factor out by are the normal subgroups— subgroups with an additional property. In ring theory, subrings are the substructures that are themselves rings, while the substructures that one can factor out by are the ideals. Ideals are usually not subrings as we have defined them, because they will generally not contain 1, but one can think of an ideal as a subring without identity. Then ideals are subrings (without 1) which satisfy an additional property (closure by multiplication by arbitrary elements of the ring on either side). In this sense the analogy with group theory is not far off.

There is also a important version for rings of the 3rd and 4th isomorphism theorems; we leave the proof to the reader.

Theorem 8.58. *Let R be a ring with ideal I. There is a bijective correspondence*

$$\Phi : \{\textit{ideals } J \textit{ with } I \subseteq J \subseteq R\} \longrightarrow \{\textit{ideals of } R/I\}$$

given by $\Phi(J) = J/I$. Moreover, for any such J as on the left hand side, we have $(R/I)/(J/I) \cong R/J$ as rings.

The ring-theoretic version of the 2nd isomorphism theorem exists, though it is not used very often, so we omit it here.

8.3.1. *Exercises.*

Exercise 8.59. This problem generalizes Example 8.43. Consider a cyclic group G of order n and let R be the group ring $\mathbb{C}G$. Let $\zeta = e^{2\pi i/n}$ be a primitive nth root of 1, so the order of ζ in $\mathbb{C}^\times$ is n. let $G = \langle a \rangle = \{1, a, a^2, \ldots, a^{n-1}\}$. For each $0 \leq j \leq n-1$ define $e_j = \frac{1}{n}\sum_{i=0}^{n-1} \zeta^{ij} a^i$.

(a) Show that $e_0, e_1, \ldots, e_{n-1}$ is a $\mathbb{C}$-basis of $\mathbb{C}G$ using formula for the determinant of a Vandermonde matrix.

(b) Prove that $e_i e_j = 0$ if $i \neq j$, while $e_j e_j = e_j$ for all j.

(c). Show that the map $\mathbb{C}^{\times n} \to \mathbb{C}G$ given by $(a_0, \ldots, a_{n-1}) \mapsto a_0 e_0 + \cdots + a_{n-1} e_{n-1}$ is an isomorphism of rings. So the group algebra $\mathbb{C}G$ is just isomorphic to a direct product of n copies of $\mathbb{C}$, as rings.

Exercise 8.60. Check the claims in Example 8.54 using the 1st isomorphism theorem.

Exercise 8.61. Recall that an element x in a ring R is nilpotent if $x^n = 0$ for some $n \geq 1$.

(a) Show that for $x, y \in R$, where R is commutative, the binomial theorem

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-1}$$

holds.

(b) Show that if x and y are nilpotent elements of a commutative ring, then $x+y$ is nilpotent.

(c) Give an example of a noncommutative ring R and nilpotent elements $x, y \in R$, such that $x+y$ is not nilpotent.

Exercise 8.62. Recall that a division ring is a ring such that every nonzero element of the ring is a unit. Show that D is a division ring if and only if the only left ideals of D are 0 and D.

Exercise 8.63. Let R be a ring, and consider the matrix ring $M_n(R)$ for some $n \geq 1$. Given an ideal I of R, let $M_n(I)$ be the set of matrices (a_{ij}) such that $a_{ij} \in I$ for all i, j.

Show that every ideal of $M_n(R)$ is of the form $M_n(I)$ for some ideal I of R. Conclude that if R is a division ring, then $M_n(R)$ is a *simple ring*, that is, that $\{0\}$ and $M_n(R)$ are the only ideals of $M_n(R)$. Show, however, that $M_n(R)$ is not itself a division ring when $n \geq 2$.

8.4. **Prime and Maximal Ideals.** We begin this section with some important notational concepts for ideals. In this section, all rings R will be assumed commutative unless stated otherwise. Some comments about how the results generalize to noncommutative rings will be given in a remark.

Let R be a commutative ring. If X is a subset, we let (X) be the *ideal generated by* X, that is, the intersection of all ideals of R which contain X. An arbitrary intersection of ideals is an ideal. Thus (X) is the unique smallest ideal of R containing X. We can describe (X) explicitly as

$$(X) = \{r_1x_1 + \cdots + r_nx_n | x_i \in X, r_i \in R \text{ for all } i,\ n \geq 1\}.$$

To see this, first note that any ideal containing X contains all expressions in the set on the right hand side. Then check that the right hand side is an ideal, which is clear from its definition. We can think of (X) as consisting of the R-linear combinations of X, analogous to the span of a set of elements in a vector space. We say that an ideal I of a commutative ring is *principal* if $I = (\{x\})$ is generated by a set with one element. In this case we remove the brackets for simplicity and write $I = (x) = \{rx | r \in R\}$. This ideal is also written as Rx (or xR). Similarly, we can write $(x_1, \ldots x_n)$ as $Rx_1 + \cdots + Rx_n$. An ideal I is called *finitely generated* if it equals $(x_1, \ldots, x_n)$ for some $x_i \in I$; otherwise it is called *infinitely generated*. The zero ideal $\{0\}$ is equal to (0) and we also sometimes just write it as 0.

Next, we review the notion of products of ideals. For arbitrary subsets X, Y of a ring R, one defines XY to be the set of all *sums* of the form $\{x_1y_1 + \cdots + x_ny_n | x_i \in X, y_i \in Y, n \geq 1\}$. For example, $RX = (X)$ is the ideal generated by X. This is a different use of the product notation than we saw in groups; closure under sums is necessary because we want a product of ideals to be an ideal. The reader may check that if I and J are ideals of a ring R, then the product IJ is a again an ideal.

We call an ideal I of a ring R *proper* if $I \neq R$.

Definition 8.64. Let R be a commutative ring with proper ideal I. The ideal I is *prime* if whenever $x, y \in R$ such that $xy \in I$, then either $x \in I$ or $y \in I$. The ideal I is *maximal* if there does not exist any ideal J such that $I \subsetneq J \subsetneq R$.

It is important to note the convention that R is not considered a prime ideal of itself.

Lemma 8.65. *Let R be a commutative ring. Then R is a field if and only if 0 and R are the only ideals of R, in other words 0 is a maximal ideal of R.*

Proof. Suppose that R is a field. If I is a nonzero ideal of R, we can choose some $0 \neq x \in I$. Then x is a unit in R, and so $1 = x^{-1}x \in I$, and thus $r1 = r \in I$ for all $r \in R$. So $I = R$. Conversely, suppose that every nonzero ideal of R is equal to R. If $0 \neq x \in R$, then the principal ideal Rx is nonzero and so we must have $Rx = R$. In particular, $1 \in Rx$, so there is $y \in R$ with $yx = 1$, and x is a unit. Thus all nonzero elements are units and so R is a field. $\square$

Both prime and maximal ideals have interesting reinterpretations in terms of the properties of the factor rings they determine.

Proposition 8.66. *Let R be a ring with proper ideal I.*

(1) *I is maximal if and only if R/I is a field.*

(2) *I is prime if and only if R/I is a domain.*

Proof. (1) By the correspondence of ideals in Theorem 8.58, ideals J of R with $I \subsetneq J \subsetneq R$ are in one-to-one correspondence with ideals of R/I which are not equal to 0 or R/I. Thus I is maximal if and only if R/I has only 0 and R/I as ideals, if and only if R/I is a field by Lemma 8.65.

(2) Suppose that I is prime. If $(x+I)(y+I) = 0+I$ in R/I, then $xy+I = 0+I$ and so $xy \in I$. Then by definition $x \in I$ or $y \in I$, so $x + I = 0 + I$ or $y + I = 0 + I$. This shows that R/I is a domain. The converse is similar. □

Corollary 8.67. *Any maximal ideal of a ring is prime.*

Proof. Note that any field is a domain, because a unit is always a non-zero-divisor. Thus this result follows immediately from the proposition. □

Example 8.68. Let $R = \mathbb{Z}$. Note that the zero ideal 0 is prime but not maximal, since $R/0 \cong R$ and R is a domain but not a field. If p is a prime number, then $\mathbb{Z}/p\mathbb{Z} \cong \mathbb{Z}_p$ is a field, as we have seen; so $p\mathbb{Z}$ is a maximal (and hence also prime) ideal of $\mathbb{Z}$. If $m = 1$ then $m\mathbb{Z} = \mathbb{Z}$ which is neither prime nor maximal by definition. If $m > 1$ is not prime then $\mathbb{Z}/m\mathbb{Z} \cong \mathbb{Z}_m$ is not a domain, so $m\mathbb{Z}$ is not a prime ideal of $\mathbb{Z}$ in this case. In conclusion, the non-zero prime ideals of $\mathbb{Z}$ are are all maximal ideals, and they are in one-to-one correspondence with the positive prime numbers.

Example 8.69. Let F be a field and let $I = (x) = xF[x] \subseteq F[x]$. We saw in Example 8.55 that I is the kernel of the homomorphism $\phi : F[x] \to F$ which evaluates x at 0, and thus $F[x]/I \cong F$ by the first isomorphism theorem. Since F is a field, the ideal I must be a maximal ideal of $F[x]$.

Example 8.70. Consider the ring $R = \mathbb{Z}[x]$. Similarly as in previous example, $\mathbb{Z}[x]/(x) \cong \mathbb{Z}$; since $\mathbb{Z}$ is a domain but not a field, (x) is prime but not maximal in this case. Given any prime $p \in \mathbb{Z}$, we know that $p\mathbb{Z}$ is maximal as an ideal of $\mathbb{Z}$; then by the ideal correspondence in Theorem 8.58, the corresponding ideal $(x, p) = x\mathbb{Z}[x] + p\mathbb{Z}[x]$ of $\mathbb{Z}[x]$ is maximal in $\mathbb{Z}[x]$, and moreover $\mathbb{Z}[x]/(x, p) \cong \mathbb{Z}/p\mathbb{Z} = \mathbb{Z}_p$. Since the primes (p) give all maximal ideals of $\mathbb{Z}$, the ideals (x, p) give all maximal ideals of $\mathbb{Z}[x]$ which contain (x).

It is sometimes useful to think of prime ideals in the following alternative way, which works with ideals rather than elements.

Lemma 8.71. *Let P be an ideal of a commutative ring R. The following are equivalent:*

(i) *Whenever I and J are ideals with $IJ \subseteq P$, then $I \subseteq P$ or $J \subseteq P$.*

(ii) *Whenever I and J are ideals with $P \subseteq I$, $P \subseteq J$, and $IJ \subseteq P$, then $P = I$ or $P = J$.*

(iii) *P is prime.*

Proof. It is obvious that $(i) \implies (ii)$. Suppose (ii) holds and that $xy \in P$. Let $I = P + (x)$ and $J = P + (y)$. Then $P \subseteq I$ and $P \subseteq J$, and moreover $IJ = (P + (x))(P + (y)) \subseteq P + (x)(y) = P + xyR = P$. Thus either $I = P$ or $J = P$, and thus either $x \in P$ or $y \in P$, implying (iii). Finally, if (iii) holds, let I and J be ideals with $IJ \subseteq P$. If neither $I \subseteq P$ or $J \subseteq P$ holds, then we can choose $x \in I - P$ and $y \in J - P$. Thus $xy \in IJ \subseteq P$ and so $x \in P$ or $y \in P$, a contradiction. Thus in fact $I \subseteq P$ or $J \subseteq P$ and we have (i). □

Remark 8.72. We have focused on commutative rings here. One may develop a theory of maximal and prime ideals in noncommutative rings as well, but they satisfy weaker results. Let R be an arbitrary (not necessarily commutative) ring. If X and Y are subsets of R, the notation $XY = \{x_1y_1 + \cdots + x_ny_n | x_i \in X, y_i \in Y, n \geq 1\}$ is defined in the same way as in the commutative case. A proper ideal P of R is called prime if it satisfies the condition in Lemma 8.71(i): If $IJ \subseteq P$ for ideals I, J, then $I \subseteq P$ or $J \subseteq P$. An ideal P such that $xy \in P$ implies $x \in P$ or $y \in P$ is called *completely prime*; this is a stronger condition than prime and is much more rarely satisfied. An ideal is said to be maximal just as before, if it is maximal under inclusion among proper ideals. Again, maximal ideals must be prime (but need not be completely prime).

A ring is called *prime* if 0 is a prime ideal; similarly as in Proposition 8.66, an ideal P is prime if and only if R/P is a prime ring. However, a prime ring is not necessarily a domain (rather, R/P is a domain if and only if P is completely prime). A ring R is called *simple* if 0 and R are its only ideals; by ideal correspondence, an ideal I of R is maximal if and only if R/I is simple. A simple ring need not be a division ring, however, or even a domain, though it is a prime ring.

The ring of matrices $M_n(D)$ over a division ring D, with $n \geq 2$, is an example of a simple ring which is not a domain (Exercise 8.63).

8.4.1. *Exercises.*

Exercise 8.73. A commutative ring R is called *local* if has a unique maximal ideal M. Show that the following are equivalent for a commutative ring R:

(i) R is local.

(ii) The set of non-units in R is an ideal of R.

Exercise 8.74. Let F be a field and let $R = F[[x]]$ be the ring of formal power series.

(a). Show that every proper nonzero ideal of R is of the form (x^n) for some $n \geq 1$.

(b). Show that the only prime ideals of R are 0 and (x), and so (x) is the only maximal ideal and R is a local ring.

Exercise 8.75. Let F be a field. Define the polynomial ring $R = F[x, y]$ in two variables over F by $F[x, y] = (F[x])[y]$.

Show that 0, (x) and (y) are prime but not maximal ideals of R, and that (x, y) is a maximal ideal.

9. Further fundamental techniques in ring theory

9.1. **Zorn's Lemma and applications.** We continue to assume that R is a commutative ring in this section for convenience, although most of the results extend easily to noncommutative rings. Given a ring R, must it have any maximal ideals at all? Throw away the irritating zero ring. Then a ring R has at least one proper ideal, namely 0, so the set of proper ideals is nonempty. But why must there exist a proper ideal which is maximal under inclusion?

The key to proving this is Zorn's Lemma, a basic result in set theory which has many applications in algebra. We begin with a review of some basic concepts of orderings on sets.

Definition 9.1. A *partially ordered set* or *poset* is a set $\mathcal{P}$ with a binary relation $\leq$ such that

(1) (reflexivity) $x \leq x$ for all $x \in \mathcal{P}$.

(2) (transitivity) If $x \leq y$ and $y \leq z$, then $x \leq z$, for all $x, y, z \in \mathcal{P}$.

(3) (antisymmetry) If $x \leq y$ and $y \leq x$, then $x = y$ for all $x, y \in \mathcal{P}$.

We sometimes write $x < y$ to mean $x \leq y$ and $x \neq y$. We might also write $y \geq x$ as a synonym for $x \leq y$.

Example 9.2. Let S be a set and let $\mathcal{P}(S)$ be the power set of S, i.e. the set of all subsets of S. Then $\mathcal{P}(S)$ is a poset where we define $X \leq Y$ to mean $X \subseteq Y$ for subsets X, Y of S. The axioms of a poset are immediate.

Note that in a general poset we may well have elements x, y such that neither $x \leq y$ nor $y \leq x$ holds. This is already clear in the example above; take $S = \{1, 2, 3\}$ for example, and $X = \{1, 2\}$ and $Y = \{2, 3\}$; neither set contains the other. A poset $\mathcal{P}$ is called *totally* or *linearly* ordered if for

all $x, y \in \mathcal{P}$, either $x \leq y$ or $y \leq x$ holds. Totally ordered sets, even of the same cardinality, can have very different kinds of orders. For example, we have the natural numbers $\mathbb{N}$ with their usual order, where given $a, b \in \mathbb{N}$ there are finitely many $c \in \mathbb{N}$ with $a \leq c \leq b$. On the other hand, one has the rational numbers $\mathbb{Q}$ with their usual order, where for any $a < b$ in $\mathbb{Q}$ there are infinitely many $c \in \mathbb{Q}$ with $a \leq c \leq b$.

Definition 9.3. If $\mathcal{P}$ is a poset, and $B \subseteq \mathcal{P}$, an *upper bound* for B is an $x \in \mathcal{P}$ such that $b \leq x$ for all $b \in B$ (note that x might or might not be contained in B itself). A *maximal* element of $\mathcal{P}$ is an element $y \in \mathcal{P}$ such that there does not exist any $x \in \mathcal{P}$ with $y < x$. Equivalently, $y \in \mathcal{P}$ is maximal if $y \leq x$ implies $x = y$.

Note that a poset might have many distinct maximal elements. A totally ordered poset, on the other hand, either has a single maximal element or no maximal elements at all.

Example 9.4. Let R be a (non-zero) ring and let $\mathcal{P}$ be the set of all proper ideals of R. Then $\mathcal{P}$ is a poset under inclusion, where $I \leq J$ means $I \subseteq J$. Since we have excluded R itself from $\mathcal{P}$, note that a maximal ideal of R is the same thing as a maximal element of the poset $\mathcal{P}$.

Given a poset $\mathcal{P}$, any subset $S \subseteq P$ is also a poset under the inherited order, where $x \leq y$ for $x, y \in S$ if and only if $x \leq y$ in $\mathcal{P}$. A subset S of $\mathcal{P}$ is called a *chain* if S is totally ordered under its inherited order. We are now ready to state Zorn's Lemma.

Lemma 9.5. *Let $\mathcal{P}$ be a nonempty poset. Suppose that every chain B in $\mathcal{P}$ has an upper bound in $\mathcal{P}$. Then $\mathcal{P}$ has a maximal element.*

Zorn's Lemma is actually equivalent to the axiom of choice in set theory; each can be proved from the other. So we also just assume Zorn's Lemma as an axiom.

The intuition behind Zorn's lemma is not hard to understand. If we are looking for a maximal element in $\mathcal{P}$, we can start by picking any $x_1 \in \mathcal{P}$; if it is not maximal, pick $x_1 < x_2$; continuing in this way, if no maximal element is acheived, we get a set $S = \{x_i | i \in \mathbb{N}\}$ which is a chain in $\mathcal{P}$. If every chain has an upper bound, then there is $y_1 \in \mathcal{P}$ which is an upper bound for S; in this case it means that $x_i < y_1$ for all i. Now if y_1 is not maximal we can start the process all over again. The hypothesis of Zorn's lemma that chains have upper bounds allows us to never be "stuck"—if we do not have any maximal element yet in our chain, we can make the chain bigger. Thus at some point this (infinitary) process will stop with a maximal element having been found.

Let us now give our first application of Zorn's lemma.

Proposition 9.6. *Let R be a nonzero commutative ring. Then any proper ideal H of R is contained in a maximal ideal.*

Proof. We consider the poset $\mathcal{P}$ of all proper ideals of R which contain H, which is nonempty because $H \in \mathcal{P}$. The order is the inclusion, as in Example 9.4. Our goal is to show that $\mathcal{P}$ must have a maximal element. This is the conclusion of Zorn's lemma, so we just need to verify the hypothesis. Consider an arbitrary chain in $\mathcal{P}$, which is a collection of ideals of R containing H, say $B = \{I_\alpha | \alpha \in A\}$ for some index set A, such that for any $\alpha, \beta \in A$, either $I_\alpha \subseteq I_\beta$ or $I_\beta \subseteq I_\alpha$. We need to find an upper bound for the chain, in other words a proper ideal J of R such that $I_\alpha \subseteq J$ for all $\alpha \in A$. We simply take $J = \bigcup_{\alpha \in A} I_\alpha$ to be the union of all of the ideals in the chain B. Then certainly $I_\alpha \subseteq J$ for all $\alpha \in A$, so if $J \in \mathcal{P}$ then it is an upper bound for B. For any $x, y \in J$, we have $x \in I_\alpha$ for some α and $y \in I_\beta$ for some β. Since B is a chain, either $I_\alpha \subseteq I_\beta$ or $I_\beta \subseteq I_\alpha$. In the former case, both x and y are in the ideal I_β and thus $x - y \in I_\beta$; so $x - y \in J$. Similarly, if $I_\beta \subseteq I_\alpha$ then $x - y \in I_\alpha \subseteq J$. For any $r \in R$ and $x \in J$, again we have $x \in I_\alpha$ for some α, and so $rx \in I_\alpha \subseteq J$. We see that J is again an ideal.

Suppose that $J = R$. Then $1 \in J$, and so $1 \in I_\alpha$ for some α. But then $I_\alpha = R$ is the unit ideal, contradicting that I_α belongs to the poset P of proper ideals of R. This shows that $J \neq R$ and so J is a proper ideal of R. Thus J is in the poset $\mathcal{P}$. Now J is the required upper bound of the chain B, and the hypothesis of Zorn's Lemma has been verified. Thus $\mathcal{P}$ has a maximal element, in other words, R has a maximal ideal containing H. □

There are a couple of pitfalls in the use of Zorn's Lemma that are worth mentioning now. First, the requirement that the poset be nonempty is serious. It is easy to define a poset by some condition that seems reasonable at first, and then use Zorn's lemma to prove a patently absurd statement, if the poset you defined was actually empty. Another common mistake in checking the hypothesis of Zorn's Lemma is to take a chain that is too special. It is not enough, in general, to check that for chains of the form $I_1 \subseteq I_2 \subseteq I_3 \subseteq \ldots I_n \subseteq \ldots$, that this chain has an upper bound. Technically, one needs to take arbitrary (potentially uncountable, for example) index sets for the chains, and not make any assumption as to what kind of order the chain has.

Let us now give another, slightly trickier, application of Zorn's Lemma. If R is a ring, recall that an element $x \in R$ is *nilpotent* if $x^n = 0$ for some $n \geq 1$. If R is commutative, then the set N of all nilpotent elements of R is an ideal; this easily follows from Exercise 8.61. The ideal N is called the *nilradical* of R, and it has the following interesting alternative characterization.

Proposition 9.7. *Let R be a nonzero commutative ring. The nilradical N of R is equal to the intersection of all prime ideals of R.*

Proof. Let J be the intersection of all prime ideals in the ring. Note that since every nonzero ring has a maximal ideal, R does have at least one prime ideal, so J is proper. Suppose that $x \in N$. Since $x^n = 0$ for some $n \geq 1$, for any ideal I we have $x^n \in I$. Now if I is prime, by the defining property of a prime ideal (and induction) we see that $x^n \in I$ implies $x \in I$. Thus x is in every prime ideal, and so $N \subseteq J$.

Conversely, suppose that $x \notin N$, so x is not nilpotent. Let $S = \{1, x, x^2, x^3, \dots\}$ be the set of powers of x; by assumption S does not contain 0. Consider the set $\mathcal{P}$ of all proper ideals I of R such that $I \cap S = \emptyset$. The ideal 0 is one such ideal, so $\mathcal{P}$ is nonempty. Consider $\mathcal{P}$ as a poset under inclusion of ideals, as usual.

We claim that the hypothesis of Zorn's Lemma is satisfied. For, given a chain $\{I_\alpha | \alpha \in A\}$ of ideals in $\mathcal{P}$, the union J of the chain is again a proper ideal of R, by exactly the same argument as in Proposition 9.6. Moreover, J is still in $\mathcal{P}$, for otherwise $J \cap S$ is nonempty, which means that $I_\alpha \cap S$ is nonempty for some α, a contradiction. Thus every chain in $\mathcal{P}$ has an upper bound, and so $\mathcal{P}$ has a maximal element, say M.

Next, we claim that M is a prime ideal. We use the characterization of prime ideal in Lemma 8.71(2). Let $M \subseteq I$ and $M \subseteq J$ for ideals I and J such that $IJ \subseteq M$. Suppose that $M \neq I$ and $M \neq J$. By maximality of M in $\mathcal{P}$, I and J do not belong to $\mathcal{P}$, so we can find $x^i \in I \cap S$ and $x^j \in J \cap S$. Then $x^{i+j} \in IJ \subseteq M$, contradicting $M \cap S = \emptyset$. Thus $M = I$ or $M = J$ and M is prime. Since $x \notin M$, we have found a prime ideal not containing x.

We have shown that if $x \notin N$, then $x \notin M$ for some prime ideal M, and so $x \notin J$. This shows that $J \subseteq N$. Since we already showed that $N \subseteq J$, we conclude that $N = J$. □

The intersection of all of the prime ideals of a ring is also called the *prime radical*. The result we have just proved shows that for any commutative ring R, its prime radical and its nilradical are the same thing.

Example 9.8. Let $R = \mathbb{Z}/n\mathbb{Z}$ for some $n \geq 1$, and factorize n as $n = p_1^{e_1} p_2^{e_2} \dots p_m^{e_m}$, where the p_i are distinct primes and $e_i \geq 1$ for all i. We claim that the nilradical (and prime radical) of R is $r\mathbb{Z}/n\mathbb{Z}$, where $r = p_1 p_2 \dots p_m$ is the product of the primes to the first power. To demonstrate Proposition 9.7 we calculate this in two different ways.

First, if $e = \max(e_1, \dots, e_m)$ then r^e is a multiple of n, so $r^e \in n\mathbb{Z}$ and hence $(rz)^e = r^e z^e \in n\mathbb{Z}$ for any z; so $rz + n\mathbb{Z}$ is nilpotent in R for all $z \in \mathbb{Z}$. Conversely, if s is not divisible by p_i for some

i, then s^j is never divisible by p_i for all $j \geq 1$, and so $s^j \notin n\mathbb{Z}$ and hence $s + n\mathbb{Z}$ is not nilpotent in R. It follows that if $s + n\mathbb{Z}$ is nilpotent if and only if s is a multiple of r, and so $N = r\mathbb{Z}/n\mathbb{Z}$ is the nilradical as claimed.

We can also see that N is the intersection of the prime ideals of R. The prime ideals of $\mathbb{Z}$ are 0 and the ideals $p\mathbb{Z}$ for primes p. By ideal correspondence, the prime ideals of R are $p\mathbb{Z}/n\mathbb{Z}$ for all primes p such that $p\mathbb{Z}$ contains $n\mathbb{Z}$, in other words such that p divides n. Thus the prime ideals of R are exactly the $p_i\mathbb{Z}/n\mathbb{Z}$ for $1 \leq i \leq m$, and the intersection of these primes is equal to $r\mathbb{Z}/n\mathbb{Z}$ where $r = p_1 p_2 \dots p_m$, as we found before.

Since our study of groups focused heavily on finite groups, we did not ask earlier the question of whether any nontrivial group must have a maximal subgroup. One could attempt to use the same idea as in Proposition 9.6 to prove this, but it doesn't work. It is true that the union of a chain of subgroups is always a subgroup, but if all of the subgroups in the chain are proper, the union need not be. The key to the proof for ideals was that properness of an ideal is equivalent to not containing 1, and this is stable under taking unions. In fact, the corresponding result for groups is false; there do exist groups without any maximal subgroup. See Exercise 9.9.

9.1.1. *Exercises.*

Exercise 9.9. Show that $G = (\mathbb{Q}, +)$ has no maximal subgroups. (Hint: Suppose that M is a maximal proper subgroup of Q. Since Q is abelian, M is normal and we can consider Q/M. Since M is maximal, Q/M is a simple abelian group, which must be isomorphic to $\mathbb{Z}_p$ for some prime p. Thus $pQ \subseteq M$. But show that $pQ = Q$).

Exercise 9.10. Let R be a commutative ring and let $S = R[x]$. Show that $f = a_0 + a_1 x + \dots + a_m x^m$ is a unit in S if and only if a_0 is a unit in R and $a_1, \dots, a_m$ are all nilpotent in R. (Hint: If the conditions on the a_i hold, consider Exercise 8.29. Conversely, if f is a unit, then the image of f in the factor ring $R[x]/P[x] \cong R/P[x]$ is a unit for all prime ideals P of R. Use this to show that the a_i for $2 \leq i \leq m$ belong to every prime ideal of R.

Exercise 9.11. Given a poset P, one can define the *opposite poset* P^{op} whose elements are the same as in P, but where $x \leq y$ in P^{op} if and only if $y \leq x$ in P.

(a) Show that P^{op} is again a poset.

(b) A *lower bound* for a subset $X \subseteq P$ is an element $z \in P$ such that $z \leq x$ for all $x \in X$. A *minimal element* of P is $y \in P$ such that there does not exist $z \in P$ with $z < y$. Prove that if every chain in P has a lower bound, then P has a minimal element.

Exercise 9.12. A *minimal prime* in a commutative ring R is a prime ideal I of R such that there does not exist any prime ideal J with $J \subsetneq I$. In other words, I is a minimal prime if it is a minimal element of the poset of prime ideals of R under inclusion.

Prove that any commutative ring R has a minimal prime. (Hint: apply Exercise 9.11. Check the hypothesis by proving that the intersection of all of the elements in a chain of prime ideals is again a prime ideal.)

Exercise 9.13. Let R be a commutative ring, and let $I = (r_1, \ldots, r_n)$ be a nonzero finitely generated ideal of R. Prove that there is an ideal J of R which is maximal among ideals which do not contain I.

Exercise 9.14. Let R be a commutative ring. Prove that if every prime ideal of R is finitely generated, then all ideals of R are finitely generated, in the following steps:

(a). Suppose that R has an ideal which is not finitely generated. Show that there is an ideal P which is maximal under inclusion among the set of non-finitely generated ideals.

(b). Prove that P is prime: Suppose that $xy \in P$, but $x \notin P$ and $y \notin P$. Define $I = P + (x)$ and note that I is finitely generated, say $I = (p_1 + xq_1, \ldots, p_n + xq_n)$, where $p_i \in P, q_i \in R$. Let $K = (p_1, \ldots p_n)$ and let $J = \{r \in R | rx \in P\}$; note that J is also finitely generated. Show that $Jx + K = P$, and that therefore P is finitely generated, a contradiction.

9.2. **The Chinese Remainder Theorem.** The Chinese Remainder Theorem gives a way of decomposing a factor ring of a commutative ring as a direct product of simpler factor rings in some cases. It may be thought of as roughly analogous to recognizing a group as an internal direct product in group theory.

Definition 9.15. Let R be a ring. Two ideals I and J of R are said to be *comaximal* if $I + J = R$.

Note that if I and J are distinct maximal ideals of R, then $I + J$ is also an ideal which contains both I and J and thus must be R. So a pair of distinct maximal ideals are comaximal. The ideals in a comaximal pair do not have to be maximal ideals, however.

Theorem 9.16. *Let $I_1, I_2, \ldots, I_n$ be ideals of a commutative ring R and assume that the I_j are pairwise comaximal, i.e. that I_i and I_j are comaximal for every $i \neq j$. Then*

(1) $I_1 I_2 \ldots I_n = I_1 \cap I_2 \cap \cdots \cap I_n$.

(2) $R/(I_1 \cap I_2 \cap \cdots \cap I_n) \cong R/I_1 \times R/I_2 \times \cdots \times R/I_n$ *as rings.*

Proof. The statement is vacuous when $n = 1$, so assume that $n \geq 2$.

We first prove the theorem for two ideals I and J. Note that $IJ \subseteq I \cap J$ holds for any pair of ideals I and J. Now if I and J are comaximal, since $I + J = R$ we can write $1 = x + y$ for some $x \in I, y \in J$. Then if $r \in I \cap J$, $r = r1 = r(x + y) = rx + ry$. Since $r \in J$, $rx \in JI = IJ$ and since $r \in I$, $ry \in IJ$. Thus $r \in IJ$ and so $I \cap J = IJ$. Now consider the function $\phi : R \to R/I \times R/J$ defined by $\phi(r) = (r + I, r + J)$. This is easily seen to be a homomorphism of rings. The kernel of ϕ is clearly $\ker \phi = I \cap J$. Thus by the 1st isomorphism theorem, we have an isomorphism of rings $R/(I \cap J) \cong \phi(R)$. However, we can see that ϕ is surjective as follows. Given $(r + I, s + J) \in R/I \times R/J$, let $t = ry + sx$. Then $t - r = ry + sx - r = r(y - 1) + sx = -rx + sx = (s - r)x \in I$ and $t - s = ry + sx - s = ry + s(x - 1) = ry - sy = (r - s)y \in J$. It follows that $\phi(t) = (t + I, t + J) = (r + I, s + J)$ and ϕ is surjective. Thus $R/(I \cap J) \cong R/I \times R/J$ and the case of two ideals is proved.

Now consider the general case. We claim that I_1 and $I_2 I_3 \dots I_n$ are comaximal. Suppose not; then $I_1 + I_2 I_3 \dots I_n$ is a proper ideal of R, and so it must be contained in a maximal ideal M, by Proposition 9.6. Since M is maximal, it is a prime ideal. Now $I_2 I_3 \dots I_n \subseteq M$ in particular. By the characterization of prime ideals given in Lemma 8.71, this implies that $I_j \subseteq M$ for some j. But now $I_1 + I_j \subseteq M$, contradicting that I_1 and I_j are comaximal. This proves the claim.

Applying the theorem in the case of 2 ideals, we get that $I_1(I_2 I_3 \dots I_n) = I_1 \cap (I_2 I_3 \dots I_n)$. Since $I_2 I_3 \dots I_n$ is a product of a smaller number of pairwise comaximal ideals, we see that $I_1 I_2 \dots I_n = I_1 \cap (I_2 \cap \dots \cap I_n)$ by induction on the number of ideals. This proves (1) in general.

Again applying the two ideal case, we have $R/(I_1 \cap I_2 \cap \dots \cap I_n) = R/(I_1 \cap (I_2 I_3 \dots I_n)) \cong R/I_1 \times R/(I_2 \dots I_n)$. Again by induction on the number of ideals, $R/(I_2 \dots I_n) \cong R/I_2 \times \dots \times R/I_n$ and (2) is proved. □

Corollary 9.17. *Let n be a positive integer with prime factorization $n = p_1^{e_1} p_2^{e_2} \dots p_m^{e_m}$, where the p_i are distinct primes. Then*

(1) $\mathbb{Z}_n \cong \mathbb{Z}_{p_1^{e_1}} \times \dots \times \mathbb{Z}_{p_m^{e_m}}$ *as rings.*

(2) *If $\mathbb{Z}_n^\times$ is the units group of the ring $\mathbb{Z}_n$, we also get $\mathbb{Z}_n^\times \cong \mathbb{Z}_{p_1^{e_1}}^\times \times \dots \times \mathbb{Z}_{p_m^{e_m}}^\times$ as groups.*

Proof. For any nonzero integers $a, b \in \mathbb{Z}$, the reader can check that $a\mathbb{Z} + b\mathbb{Z} = \gcd(a, b)\mathbb{Z}$ and $a\mathbb{Z} \cap b\mathbb{Z} = \operatorname{lcm}(a, b)\mathbb{Z}$. Thus when $\gcd(a, b) = 1$ then $a\mathbb{Z}$ and $b\mathbb{Z}$ are comaximal. In particular, setting $I_i = \mathbb{Z}_{p_i^{e_i}}$ we see that $I_1, \dots, I_m$ are pairwise comaximal, and so (1) follows from the Chinese remainder theorem.

The units group of a direct product of rings is the direct product of the units groups of the factors. Thus part (2) follows from part (1). □

Note that the corollary proves Theorem 6.18(1), which was stated earlier without proof.

Example 9.18. Let m and n be positive integers with $\gcd(m,n) = 1$. The problem of determining a solution x to the simultaneous congruences $x \equiv a \mod m$ and $x \equiv b \mod n$ goes back at least to the writing of Chinese mathematician Sun-tzu in the 3rd Century A.D. (though not stated in the language of congruence, which is more modern). This motivating problem is what gives the Chinese remainder theorem its name.

We can solve the problem in our ring-theoretic framework as follows. Let $R = \mathbb{Z}$, let $I = m\mathbb{Z}$ and $J = n\mathbb{Z}$. Since $\gcd(m,n) = 1$, there are $s, t \in \mathbb{Z}$ such that $sm + tn = 1$, and so $I + J = R$ and I and J are comaximal. By Theorem 9.16, there is an isomorphism $\phi : R/(I \cap J) \to R/I \times R/J$. In this case $I \cap J$ consists of integers which are multiples of m and n, and hence $I \cap J = mn\mathbb{Z}$ since $\operatorname{lcm}(m,n) = mn$. We seek an element x such that $\phi(x+mn\mathbb{Z}) = (x+m\mathbb{Z}, x+n\mathbb{Z}) = (a+m\mathbb{Z}, b+n\mathbb{Z})$. This equation shows that the element x we seek is unique only up to multiples of mn.

The proof of Theorem 9.16 shows how to choose x. The key is to find s and t explicitly (which can be done by inspection for small m and n, or using the Euclidean algorithm for large ones). We then have $u + v = 1$, where $u = sm \in I$ and $v = tn \in J$. Then $x = bv + au$ is a solution.

For example, to solve the simultaneous congruences $x \equiv 4 \mod 21$ and $x \equiv 7 \mod 11$, one first notes that $(-1)(21)+(2)(11) = 1$; then $x = (22)(4)+(-21)(7) = -59$ is a solution. Of course, there is a unique positive solution for x with $1 \leq x \leq (21)(11)$, which in this case is $x = -59+231 = 172$.

A similar method can be used to solve simultaneous congruences with moduli $m_1, m_2, \ldots, m_k$ that are pairwise relatively prime.

While the original motivation behind the Chinese remainder theorem comes from its application to the integers, we will see that it has useful applications in many other rings, such as the polynomial ring $F[x]$ and other principal ideal domains (which we will define soon).

9.2.1. *Exercises.*

Exercise 9.19. Let R be a commutative ring.

(a). Show that an ideal I is equal to an intersection of finitely many maximal ideals of R if and only if R/I is isomorphic to a direct product of finitely many fields.

(b). Show that if I is an intersection of finitely many distinct maximal ideals of R, say $I = M_1 \cap \cdots \cap M_n$, then the ideals M_i are uniquely determined (up to rearrangement).

(c). Give an example showing that the same property as in (b) does not hold in groups. In other words, find a group G and a subgroup H such that H can be written as an intersection of maximal subgroups of G in multiple different ways.

Exercise 9.20. Find a solution to the system of congruences

$$x \equiv 1 \mod 7, \quad x \equiv 2 \mod 11, \quad x \equiv 3 \mod 13$$

by using the method of Example 9.18. (Hint: one way is to find x' satisfying the first two congruences, then solve the pair of congruences $x \equiv x' \mod 77$, $x \equiv 3 \mod 13$.)

9.3. **Localization.** The familiar set of rational numbers $\mathbb{Q}$ consists of fractions a/b where $a, b \in \mathbb{Z}$ and b is nonzero. Thus a rational fraction just amounts to a choice of two integers, one nonzero. However, the same fraction can be written in many different ways, so $1/2 = 50/100 = (-3)/(-6)$ for example. A careful construction of $\mathbb{Q}$ from $\mathbb{Z}$ must take this into account and check that the set of fractions is a number system with well-defined operations.

Of course $\mathbb{Q}$ has the advantage that one can divide by any nonzero element, unlike in $\mathbb{Z}$. We often face the same issue in a general ring R. There are certain elements that are not units, which it would be helpful to have inverses for, as it would give us a larger space in which to work. Localization is the formal process of adding inverses to elements in a given ring. Its name arises from the fact that for rings of functions in geometry (especially algebraic geometry), taking a localization is a way of producing a ring of functions which may be defined only locally on a neighborhood of a point rather than globally.

Let R be a commutative ring in this section. (There is a version of localization for a noncommutative ring, but it is considerably more complicated and only works in more limited circumstances.) A *multiplicative system* $X \subseteq R$ is a subset such that $1 \in X$ and if $x, y \in X$, then $xy \in X$. If one would like to add elements to a ring R so that certain elements become units, note that 1 is already a unit, and if x and y are units, then xy is also a unit. For this reason we might as well focus on adding inverses to all of the elements in a multiplicative system X.

Example 9.21. Let us first review precisely how $\mathbb{Q}$ is constructed from $\mathbb{Z}$. The goal is to embed $\mathbb{Z}$ in a field. Let $S = \{(a, b) | a, b \in \mathbb{Z}, b \neq 0\}$ be the set of ordered pairs of integers, where the second coordinate is nonzero. We write (a, b) using the suggestive notation a/b. We define an equivalence relation $\sim$ on S where $a/b \sim c/d$ means $ad = bc$. It is an easy exercise to check that $\sim$ is an equivalence relation.

Formally we let $\mathbb{Q}$ be the set of equivalence classes of S under $\sim$. Let $[a/b]$ represent the equivalence class of a/b. We define an addition and multiplication on equivalence classes by $[a/b] + [c/d] = [(ad + bc)/bd]$ and $[(a/b)(c/d)] = [ac/bd]$.

A number of things need to be checked. First, one must verify that addition and multiplication are well-defined, i.e. that the formulas do not depend on the choice of representatives for the

equivalence classes. Then one should check that $\mathbb{Q}$ satisfies the ring axioms under this $+$ and $\cdot$, where the additive identity is $[0/1]$ and the multiplicative identity is $[1/1]$. Then one shows that $\mathbb{Q}$ is a field. Finally, one notes that $\mathbb{Q}$ contains the original ring $\mathbb{Z}$ we started with as a subring, once $a \in \mathbb{Z}$ is identified with $[a/1] \in \mathbb{Q}$. All steps are straightforward.

Now we state the general problem we would like to solve. If R is any ring with a multiplicative system X, we would like to embed R in a larger ring S where the elements in X become units in S. In the example above, we accomplished this when $R = \mathbb{Z}$ and $X = \mathbb{Z} - \{0\}$. Moreover, one wants to find the most efficient choice of S. After all, one can also embed $\mathbb{Z}$ in the field $\mathbb{R}$ of real numbers, where all nonzero integers have become units, but one has added a lot of extra elements (irrational numbers) that one didn't need to make that happen. The ring $\mathbb{Q}$ is the most efficient choice in the sense that every element of $\mathbb{Q}$ is of the form ab^{-1} with $a \in R$ and $b \in X$.

It turns out to be useful to allow X to be an arbitrary multiplicative system, which creates the following problem. If $x \in X$ is a zero divisor in R, say $rx = 0$ with $r \neq 0$, $x \neq 0$, and S is a ring containing R as a subring in which x becomes a unit in S, say $xy = 1$ in S, then $0 = 0y = rxy = r1 = r$, which is a contradiction. To finesse this problem, instead of looking for a ring S containing R in which the elements of X become units, we need to settle for a ring homomorphism $\phi : R \to S$ (possibly with nonzero kernel) in which $\phi(x)$ is a unit in S for all $x \in X$.

We are now ready to state the main result, which shows that a ring of fractions with the desired properties exists and has a universal property.

Theorem 9.22. *Let R be a commutative ring with multiplicative system X. There exists a ring RX^{-1}, called the* localization of R along X*, and a ring homomorphism $\phi : R \to RX^{-1}$, such that:*

(1) *$\phi(x)$ is a unit in RX^{-1} for all $x \in X$, and every element of RX^{-1} is of the form ab^{-1} where $a \in \phi(R)$ and $b \in \phi(X)$.*

(2) *ϕ satisfies the following universal property: for every ring homomorphism $\psi : R \to D$, where D is another commutative ring and where $\psi(x)$ is a unit in D for all $x \in X$, there exists a unique ring homomorphism $\theta : RX^{-1} \to D$ such that $\theta \circ \phi = \psi$.*

(3) *$\ker \phi = \{r \in R | rx = 0$ for some $x \in X\}$.*

Proof. The proof is a straightforward generalization of the method of constructing $\mathbb{Q}$ from $\mathbb{Z}$ which was described in Example 9.21. The main difference is that the equivalence relation has to be defined in a more complicated way to account for the possibility of zerodivisors in X.

Consider all ordered pairs in the set $R \times X$, but we write the ordered pair (r, x) suggestively as r/x. We put a binary relation $\sim$ on this set, where $r_1/x_1 \sim r_2/x_2$ if there exists $s \in X$ such

that $s(r_1x_2 - x_1r_2) = 0$. This relation is trivially reflexive and symmetric. To see it is transitive, suppose also that $r_2/x_2 \sim r_3/x_3$, so $t(r_2x_3 - x_2r_3) = 0$ with $t \in X$. Then

$$stx_2x_3r_1 = tx_3(sr_1x_2) = tx_3(sx_1r_2) = sx_1(tr_2x_3) = sx_1(tx_2r_3) = stx_2r_3x_1,$$

and so $stx_2(r_3x_1 - x_3r_1) = 0$, where $stx_2 \in X$ since X is multiplicatively closed. We conclude that $\sim$ is an equivalence relation. Let $[r/x]$ indicate the equivalence class of the element r/x, and let RX^{-1} be defined as the set of all equivalence classes of elements of $R \times X$ under this relation.

We claim that the operations $[r_1/x_1] + [r_2/x_2] = [(r_1x_2 + r_2x_1)/(x_1x_2)]$ and $[r_1/x_1] \cdot [r_2/x_2] = [(r_1r_2)/(x_1x_2)]$ make RX^{-1} into a ring. First, one must show that these are well defined operations on equivalence classes. If $[r_1/x_1] = [p_1/y_1]$ and $[r_2/x_2] = [p_2/y_2]$, then $s(r_1y_1 - x_1p_1) = 0$ and $t(r_2y_2 - x_2p_2) = 0$ for some $s, t \in X$. Thus

$$(r_1x_2+r_2x_1)(sty_1y_2) = str_1x_2y_1y_2+str_2x_1y_1y_2 = stx_1p_1x_2y_2+stx_1y_1x_2p_2 = (p_1y_2+y_1p_2)(stx_1x_2).$$

Then $st\big((r_1x_2+r_2x_1)(y_1y_2)-(p_1y_2+y_1p_2)(x_1x_2)\big) = 0$, in other words we have $[(r_1x_2+r_2x_1)/(x_1x_2)] = [(t_1y_2 + y_1t_2)/(y_1y_2)]$ and addition is well-defined. Showing that multiplication is well-defined is similar and left to the reader. Now that we have well-defined operations, checking the ring axioms for RX^{-1} is routine, where the identity for addition is $[0/1]$ and the identity for multiplication is $[1/1]$. It is a good exercise for the reader to check the details.

(1) We define the map $\phi : R \to RX^{-1}$ by $\phi(r) = [r/1]$. It is clear that ϕ is a ring homomorphism. If $x \in X$, then $\phi(x) = [x/1]$, and this is a unit in RX^{-1}, since $[x/1][1/x] = [x/x] = [1/1]$, so $[x/1]^{-1} = [1/x]$. We also have for a general element $[r/x]$ of RX^{-1} that $[r/x] = [r/1][1/x] = \phi(r)\phi(x)^{-1}$.

(2) Suppose that $\psi : R \to D$ is another ring homomorphism such that $\psi(x)$ is a unit in D for all $x \in X$. Define $\theta : RX^{-1} \to D$ by $\theta([r/x]) = \psi(r)\psi(x)^{-1}$. The element $\psi(x)^{-1}$ makes sense because $\psi(x)$ is a unit in D. This function is well-defined, since if $[r_1/x_1] = [r_2/x_2]$, this implies $s(r_1x_2 - x_1r_2) = 0$ for some $x \in S$, so $\psi(s)\psi(r_1)\psi(x_2) = \psi(s)\psi(x_1)\psi(r_2)$, and hence $\psi(r_1)\psi(x_1)^{-1} = \psi(r_2)\psi(x_2)^{-1}$ because $\psi(s), \psi(x_1)$, and $\psi(x_2)$ are units.

It is easy to check that θ is a ring homomorphism. Obviously $\theta\phi(r) = \theta([r/1]) = \psi(r)\psi(1)^{-1} = \psi(r)$ and so $\theta\phi = \psi$. Finally, θ is unique: If θ' is any homomorphism with $\theta'\phi = \psi$, since any ring homomorphism preserves multiplicative inverses, we have $\theta'([r/x]) = \theta'([r/1])\theta'([x/1])^{-1} = \theta'\phi(r)(\theta'\phi(x))^{-1} = \psi(r)\psi(x)^{-1}$ and hence $\theta' = \theta$.

(3) We have $\phi(r) = [r/1] = [0/1]$ in RX^{-1} if and only if $0 = x(r(1) - (1)(0)) = xr$ for some $x \in X$, by the definition of the equivalence relation. $\square$

The ring RX^{-1} is called the *localization of R along X*. When the localization RX^{-1} is used in practice, one tends to write its elements as fractions r/x or $\frac{r}{x}$ without the equivalence class formalism. One simply remembers that a particular fraction can be written in many different ways (other elements of the equivalence class), as we do with the rational numbers.

Remark 9.23. In many common situations X is a set of nonzerodivisors in R. When this is the case, $r_1/x_1 = r_2/x_2$, which means by definition $s(r_1x_2 - x_1r_2) = 0$ for some $s \in X$, is equivalent to $r_1x_2 - x_1r_2 = 0$. Thus when X is a set of nonzerodivisors, one can define the localization using the simpler and more natural equivalence relation we used in Example 9.21. Also, in this case by part (3) of the theorem the kernel of $\phi : R \to RX^{-1}$ is 0, so one can think of R as a subring of its localization RX^{-1} via the injective homomorphism ϕ.

Example 9.24. Let R be any integral domain. Then $X = R \setminus \{0\}$ is a multiplicative system. In this case RX^{-1} is called the *field of fractions* of R. It comes along with the canonical injective ring homomorphism $\phi : R \to RX^{-1}$, and usually one identifies R with its image and thinks of R as a subring of RX^{-1}. In this way we can just write r for the fraction $r/1 = \phi(r)$. It is easy to see that RX^{-1} is a field, since if $r/x \neq 0$, we must have $r \neq 0$. Then $r \in X$, so x/r is an element of RX^{-1} and clearly $x/r = (r/x)^{-1}$. So every nonzero element is a unit.

We see from this that every integral domain can be embedded in a field. When $R = \mathbb{Z}$ we recover $\mathbb{Q}$ as its field of fractions. When F is a field and we take $R = F[x]$ to be the polynomial ring, then its field of fractions is written as $F(x)$ and called the *field of rational functions in one variable over F*. The elements of $F(x)$ are formal ratios of polynomials $f(x)/g(x)$ where $g(x)$ is not 0.

Example 9.25. Since we allowed X to be any multiplicative system in R, at the opposite extreme from the case where X consists of zerodivisors is the case where $0 \in X$. Then $0(r1 - 0x) = 0$ and so $r/x = 0/1$ in RX^{-1} for all $r/x \in RX^{-1}$. Thus RX^{-1} collapses to the zero ring. This makes sense since the zero ring is the only ring in which 0 can be a unit.

9.3.1. *Exercises.*

Exercise 9.26. Prove that any field of characteristic 0 contains a unique subring isomorphic to $\mathbb{Q}$.

Exercise 9.27. Consider the ring $\mathbb{Z}_n$ for some $n \geq 2$. Let $\overline{a} \in \mathbb{Z}_n$ and let $X = \{\overline{1}, \overline{a}, \overline{a}^2, \dots\}$ be the set of powers of $\overline{a}$. Then X is a multiplicative system in $\mathbb{Z}_n$. Show that $\mathbb{Z}_nX^{-1}$ is isomorphic to $\mathbb{Z}_d$ for some divisor d of n and explain how to determine d.

Exercise 9.28. Let R be a commutative ring. The ring of formal Laurent series over R is the ring $R((x))$ given by

$$R((x)) = \{\sum_{n \geq N}^{\infty} a_n x^n | a_n \in R, N \in \mathbb{Z}\}.$$

Note that this is similar to the power series ring $R[[x]]$, except that Laurent series are allowed to include finitely many negative powers of x. The product and sum in this ring are defined similarly as for power series.

(a) Prove that if F is a field, then $F((x))$ is a field.

(b) Prove that if F is a field, then $F((x))$ is isomorphic to the field of fractions of $F[[x]]$. (Hint: use the universal property of the localization to show there is a map from the field of fractions to $F((x))$, then show it is surjective).

(c) Show that $\mathbb{Q}((x))$ is *not* the field of fractions of its subring $\mathbb{Z}[[x]]$. (Hint: consider the power series representation of e^x.)

Exercise 9.29. Recall that a commutative ring R is *local* if it has a unique maximal ideal M.

(a) Let P be a prime ideal of R. Let $X = R - P$ be the set of elements in R which are not in P. Consider the localization RX^{-1}. Show that RX^{-1} is a local ring, with unique maximal ideal $PX^{-1} = \{r/x \,|\, r \in P, x \in X\}$.

(b) Note that R/P is a domain, since P is prime. Show that RX^{-1}/PX^{-1} is isomorphic to the field of fractions of R/P.

Exercise 9.30. Let R be an integral domain with multiplicative system X not containing 0.

(a) For any ideal I of R, define $IX^{-1} = \{r/x \in RX^{-1} | r \in I\}$. Show that I is an ideal of RX^{-1}.

(b) Show that every ideal of RX^{-1} has the form IX^{-1} for some ideal I of R.

(c) Show that if P is a prime ideal of RX^{-1}, then $P = IX^{-1}$ for some prime ideal I of R with $I \cap X = \emptyset$.

10. Factorization theory in commutative domains

10.1. **Euclidean Domains.** The integers $\mathbb{Z}$ satisfy a number of important results that are keys to understanding their structure. First, there is division with remainder: for any integers a, b with $b \neq 0$, there is a quotient q and remainder r in $\mathbb{Z}$, with $0 \leq r < |b|$, such that $a = qb + r$. Second, any two integers a, b, not both zero, have a greatest common divisor $\gcd(a, b)$ which is an integral linear combination of a and b. The GCD can be calculated using the Euclidean algorithm, which is based simply on repeated applications of division with remainder. We have also seen above that the ideals of $\mathbb{Z}$ have a very simple structure—they are precisely the *principal* ideals $m\mathbb{Z}$ for $m \geq 0$.

This is another consequence of division with remainder. A third important idea is that any positive integer can be written uniquely as a product of primes. This can also be used to show that any two integers have a greatest common divisor.

The next goal is to show that all of the results above can be generalized and shown to hold for certain classes of integral domains. The existence of something like division with remainder is the most special condition, and will hold for a class of rings called *Euclidean Domains*. Integral domains such that every ideal is generated by one element are called *principal ideal domains* or PIDs, and every Euclidean domain is a PID. Finally, rings which have an analog of unique factorization into primes are called unique factorization domains or UFDs. Every PID is UFD, but it turns out that UFDs are a much more general class of rings, as PIDs are "small" in a certain sense.

The main thing we have to be more careful about when defining and studying these concepts for more general rings is the possible existence of a lot more units in the ring. The units group of $\mathbb{Z}$ is just $\{1, -1\}$, so multiplication by a unit either does nothing or negates an element, and this can be easily controlled. In more general rings, we will have to explicitly allow for unknown unit multiples in the definitions.

In the next sections we will consider these concepts in the order discussed above, from most special to the most general.

Definition 10.1. Let R be an integral domain. We say that R is a *Euclidean domain* if there is a function $d : R \to \mathbb{N} = \{0, 1, 2 \dots\}$, such that for any $a, b \in R$ with $b \neq 0$, there exist q, r such that $a = qb + r$ with either $r = 0$ or $d(r) < d(b)$.

The function d is called the *norm function* for the Euclidean domain. Because the two possible conclusions are $r = 0$ or $d(r) < d(b)$, the value of $d(0)$ is actually irrelevant. Some authors decline to define d at 0, or specify that $d(0) = 0$, but it doesn't make any difference.

Example 10.2. Let $R = \mathbb{Z}$ and define $d : R \to \mathbb{N}$ to be the absolute value function $d(a) = |a|$. Then R is a Euclidean domain. For by the usual division with remainder, if $a, b \in \mathbb{Z}$ with $b \neq 0$, we have $a = qb + r$ for unique q and r with $0 \leq r < |b|$, so $r = 0$ or $r < |b|$.

Note that in the example above the elements q and r are uniquely determined, but there is no requirement that this be the case for a Euclidean domain in general. Also, for the case of $\mathbb{Z}$, the required norm function can be taken to be something canonical and familiar—the absolute value—but other less natural norm functions would work, such as $d(a) = 2|a|$.

After the integers, the simplest example of a Euclidean domain is the ring of polynomials over a field.

Example 10.3. Let F be a field and let $R = F[x]$. For $0 \neq f \in F[x]$ define $d(f) = \deg(f)$, and let $d(0) = 0$. Then R is a Euclidean domain with respect to this norm function. This follows from polynomial long division: Given $f, g \in F[x]$ with $g \neq 0$, there are unique $q, r \in F[x]$ such that $f = qg + r$, with $r = 0$ or $\deg(r) < \deg(g)$.

The reader may have learned how to divide one polynomial by another but not have seen a proof that this always works, so we give a proof here.

Lemma 10.4. *Consider the setup in Example 10.3. Then a unique q and r with the claimed properties exist.*

Proof. Let $S = \{f - tg | t \in F[x]\}$. If $0 \in S$, take $r = 0$. Otherwise, let r be an element of S with minimal value of $d(r) = \deg(r)$ among elements of S. Write $r = a_0 + a_1 x + \cdots + a_m x^m$ and $g = b_0 + b_1 x + \cdots + b_n x^n$, where $a_m \neq 0$ and $b_n \neq 0$, so that $m = d(r)$ and $n = d(g)$. Now if $m \geq n$, the leading terms in the difference $h = r - (a_m b_n^{-1}) x^{m-n} g$ cancel, so that $d(h) < d(r) = m$. Since $h \in S$, this contradicts the choice of r. Thus $d(r) < d(g)$. Since $r = f - qg$ for some $q \in F[x]$, we now have $f = qg + r$ with either $r = 0$ or $d(r) < d(g)$, as required.

For uniqueness, suppose that $f = q'g + r'$ with $d(r') < d(g)$ or $r' = 0$. Then $(q - q')g = r' - r$. Suppose that $r' - r \neq 0$. Then $q - q' \neq 0$ as well and we get $d(q - q') + d(g) = d(r' - r)$, by Lemma 8.25. Since either r or r' is nonzero, in any case we have $d(r' - r) \leq \max(d(r'), d(r)) < d(g)$. This forces $d(q - q') < 0$ which is a contradiction. Hence $r' - r = 0$, which implies that $q - q' = 0$ as well. $\square$

More interesting examples of Euclidean domains are provided by certain *quadratic integer rings* which are important in number theory. Let D be a squarefree integer. For our purposes, it is convenient to take this to mean either $D = \pm p_1 p_2 \dots p_m$ for some nonempty set of distinct primes $p_1, \dots, p_m$, or else $D = -1$. Let $\sqrt{D}$ be a square root of D in $\mathbb{C}$ (choose either square root). We define $\mathbb{Q}(\sqrt{D}) = \{a + b\sqrt{D} | a, b \in \mathbb{Q}\}$, as a subset of $\mathbb{C}$. Note that $(a + b\sqrt{D})(c + d\sqrt{D}) = (ac + dbD + (ad + bc)\sqrt{D})$, and clearly $\mathbb{Q}(\sqrt{(D)}$ is closed under subtraction, so $\mathbb{Q}(\sqrt{D})$ is a subring of $\mathbb{C}$. In fact, $\mathbb{Q}(\sqrt{D})$ is a field, as follows. We define the norm of an element $a + b\sqrt{D} \in \mathbb{Q}(\sqrt{D})$ as $N(a + b\sqrt{D}) = (a + b\sqrt{D})(a - b\sqrt{D}) = (a^2 - b^2 D) \in \mathbb{Z}$. If $N(a + b\sqrt{D}) = 0$, then $a^2 = b^2 D$ in $\mathbb{Z}$; if both sides are nonzero, after clearing denominators, unique factorization in $\mathbb{Z}$ implies that D is a square, contradicting the choice of D. Thus $a = b = 0$ and $a + b\sqrt{D} = 0$. So $N(x) = 0$ implies $x = 0$, as we expect of something called a norm. In particular, if $0 \neq x = a + b\sqrt{D}$, then $N = N(x) = a^2 - b^2 D \neq 0$, so that $((a/N) - (b/N)\sqrt{D}) = x^{-1}$ in $\mathbb{Q}(\sqrt{D})$.

The norm is also multiplicative:

$$N((a+b\sqrt{D})(c+d\sqrt{D})) = N((ad+bcD)+(bc+ad)\sqrt{D})$$

$$= (ac+bdD)^2-(bc+ad)^2D = (a^2-b^2D)(c^2-d^2D) = N(a+b\sqrt{D})N(c+d\sqrt{D}).$$

In fact, when $D<0$ so that $\sqrt{D}$ is imaginary, then $a-b\sqrt{D} = \overline{a+b\sqrt{D}}$ and $N(x) = x\overline{x} = ||x||^2$ where $||\ ||$ is the complex norm, so multiplicativity is a consequence of the multiplicativity of the complex norm in that case.

Definition 10.5. Let D be a squarefree integer. We define the *quadratic integer ring* $\mathcal{O}_{\mathbb{Q}(\sqrt{D})} = \{a+b\omega | a,b\in\mathbb{Z}\}$, where $\omega = \sqrt{D}$ if $D \not\equiv 1 \mod 4$, while $\omega = (1+\sqrt{D})/2$ if $D \equiv 1 \mod 4$.

We also define $\mathbb{Z}[\sqrt{D}] = \{a+b\sqrt{D} | a,b\in\mathbb{Z}\}$ for any such D, so $\mathbb{Z}[\sqrt{D}] \subseteq \mathcal{O}_{\mathbb{Q}(\sqrt{D})}$, with equality unless $D \equiv 1 \mod 4$. All of the rings in question are subrings of $\mathbb{Q}(\sqrt{D})$. The motivation for the definition of $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ comes from number theory. The ring $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is the *integral closure* of $\mathbb{Z}$ inside $\mathbb{Q}(\sqrt{D})$. Explicitly, this means that $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is the set of all $\alpha \in \mathbb{Q}(\sqrt{D})$ such that α is a root of a *monic* polynomial $f = x^m + a_{m-1}x^{m-1} + \cdots + a_0 \in \mathbb{Z}[x]$, that is, a polynomial whose leading coefficient is 1. Such rings and their factorization theory are relevant to the study of certain diophantine equations. Integral closures are important in commutative algebra more generally.

We claim that if $x \in \mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ then $N(x) \in \mathbb{Z}$. This is obvious if $D \not\equiv 1 \mod 4$. If $D \equiv 1 \mod 4$, then $x = a+b\omega = (a+b/2)+(b/2)\sqrt{D}$ so

$$N(x) = (a+b/2)^2 - (b/2)^2D = a^2+ab+b^2/4-Db^2/4 = a^2+ab+b^2(1-D)/4 \in \mathbb{Z}$$

since $D-1$ is a multiple of 4, proving the claim. Now suppose that x is a unit in $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$. Then $1 = N(1) = N(x)N(x^{-1})$. Since $N(x)$ and $N(x^{-1})$ are integers, $N(x) = \pm 1$. Conversely, if $N(x) = \pm 1$ then $x^{-1} = N(x)[(a+b/2)-b/2\sqrt{D}] = N(x)[(a+b)-b\omega] \in \mathcal{O}_{\mathcal{Q}(\sqrt{D})}$, so x is a unit. We conclude that the units group of $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is $\{x \in \mathcal{O}_{\mathcal{Q}(\sqrt{D})} | N(x) = 1\}$.

The special case where $D=-1$ is called the *Gaussian integers*. In this case $\mathcal{O}_{\mathbb{Q}(\sqrt{-1})} = \mathbb{Z}[i] = \{a+bi | a,b\in\mathbb{Z}\}$. By the remarks above, this ring has units group $U(\mathbb{Z}[i]) = \{\pm 1, \pm i\}$.

Example 10.6. The ring of Gaussian integers $\mathbb{Z}[i]$ is a Euclidean domain.

Proof. We define $d(a+bi) = N(a+bi) = a^2+b^2 = ||a+bi||^2$, where $||\ ||$ is the complex norm. Let $x = a+bi$ and $y = c+di$ with $y \neq 0$. We seek $q, r \in \mathbb{Z}[i]$ such that $x = qy+r$, with $r=0$ or $N(r) < N(y)$. We know that $\mathbb{Q}[i]$ is a field, so in this ring xy^{-1} makes sense; write $z = xy^{-1} = s+ti$ where $s,t\in\mathbb{Q}$. The idea is to take q to be an element of $\mathbb{Z}[i]$ which approximates $z \in \mathbb{Q}[i]$ as closely as possible. Since $x-zy = 0$, the "error term" $r = x-qy$ should then be small.

Every rational number lies at a distance of no more than $1/2$ from some integer. Choose $q = e + fi \in \mathbb{Z}[i]$ such that $|e - s| \leq 1/2$ and $|f - t| \leq 1/2$. Then

$$||(z-q)||^2 = ||(e+fi)-(s+ti)||^2 = ||(e-s)+(f-t)i||^2 = (e-s)^2+(f-t)^2 \leq 1/4+1/4 = 1/2.$$

Now $x = zy$ and so $r = x - qy = zy - qy = (z-q)y$. Then $||r||^2 = ||(z-q)||^2||y||^2 \leq ||y||^2/2 < ||y||^2$. Thus $x = qy + r$ with $r = 0$ or $N(r) < N(y)$, as required. □

Note that in this case the choice of q and r are not necessarily unique, because there is some freedom in the choice of e and f in the proof when s or t is halfway betweeen two integers. For example, if $x = 1$ and $y = (1 + i)$, then $1 = (1 - i)(1 + i) - 1$ and $1 = (-i)(1 + i) + i$, where $N(-1) = N(i) = 1 < N(y) = 2$.

One may show in a similar way that the rings $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ are Euclidean domains for a finite number of small values of D (see Exercise 10.8), but for most D these rings are not Euclidean domains (or even unique factorization domains in the sense we will study shortly). They are all *Dedekind Domains*, rings which satisfy a looser kind of unique factorization property.

10.1.1. *Exercises.*

Exercise 10.7. Consider the ring $\mathcal{O}_{\mathbb{Q}(\sqrt{2})} = \mathbb{Z}[\sqrt{2}]$. If $u = 3+2i$ then clearly $N(u) = (3^2)-2(2^2) = 1$, so u is a unit. Show that u has infinite order in the units group and hence the units group is infinite. (It is a fact that the units group of $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is always infinite when $D > 0$.)

Exercise 10.8. Recall that when D is a squarefree integer, then the *ring of integers* in the field $\mathbb{Q}(\sqrt{D}) = \{x + y\sqrt{D} | x, y \in \mathbb{Q}\}$ is the subring $\mathcal{O} = \{a + b\omega | a, b \in \mathbb{Z}\}$ of $\mathbb{Q}(\sqrt{D})$, where $\omega = \sqrt{D}$ if D is congruent to 2 or 3 modulo 4, while $\omega = (1 + \sqrt{D})/2$ if D is congurent to 1 modulo 4. The field $\mathbb{Q}(\sqrt{D})$ has the norm $N(a + b\sqrt{D}) = a^2 - Db^2$, which is multiplicative, i.e. $N(z_1 z_2) = N(z_1)N(z_2)$ for $z_1, z_2 \in \mathbb{Q}(\sqrt{D})$.

(a) Consider the ring of integers $\mathcal{O}$ in $\mathbb{Q}(\sqrt{D})$. Suppose that for every $z \in \mathbb{Q}(\sqrt{D})$, there exists an element $y \in \mathcal{O}$ such that $|N(z - y)| < 1$. Prove that $\mathcal{O}$ is a Euclidean domain with respect to the function $d : \mathcal{O} \to \mathbb{N}$ given by $d(x) = |N(x)|$. (Hint: follow the method of proof we used to show that $\mathbb{Z}[i]$ is a Euclidean domain).

(b) Show that the ring of integers $\mathcal{O}$ is a Euclidean domain when $D = -2, 2, -3, -7$, or -11. (In each case show that part (a) applies).

10.2. **Principal Ideal Domains (PIDs).** After fields, which have no nontrivial proper ideals at all, the commutative domains with the simplest ring theory are the principal ideal domains, which

every ideal is generated by one element. We will see that such rings have a number of very nice properties which are similar to the ring $\mathbb{Z}$ of integers.

Definition 10.9. Let R be an integral domain. The ring R is a *principal ideal domain* or *PID* if every ideal I of R has the form $(a) = aR$ for some $a \in R$.

We noted that $\mathbb{Z}$ is a PID in Example 8.68. More generally, we have the following result.

Proposition 10.10. *Let R be a Euclidean domain with respect to the function $d : R \to \mathbb{N}$.*

(1) *R is a PID.*

(2) *If I is a nonzero ideal of R, then $I = (b)$ where b is any nonzero element with $d(b)$ minimal among nonzero elements of I.*

Proof. (1) If $I = 0$, then $I = (0)$ is certainly principal. Assume now that I is nonzero. Let $m = \min(d(a)|0 \neq a \in I)$ and pick any $b \in I$ with $d(b) = m$. We claim that $I = bR$. Certainly $bR \subseteq I$, since $b \in I$. If $a \in I$, we can find $q, r \in R$ such that $a = bq + r$, where $r = 0$ or $d(r) < d(b)$. Note that $r = a - bq \in I$, since $a, b \in I$. If $d(r) < d(b)$ we contradict the choice of b, which forces $r = 0$. But now $a = bq \in bR$, so $I \subseteq bR$. We have $I = bR$, as claimed, and so R is a PID.

(2) This was shown in the course of the proof of (1). □

Example 10.11. Let $\phi : \mathbb{R}[x] \to \mathbb{C}$ be the evaluation map $\phi(f(x)) = f(i)$, where $i = \sqrt{-1} \in \mathbb{C}$. (Recall from Example 8.41 that we can define an evaluation homomorphism which evaluates at an element in a commutative ring containing the coefficient field as a subring.)

Since ϕ is a homomorphism, $I = \ker \phi$ is an ideal of the Euclidean domain $\mathbb{R}[x]$. If $f = a + bx$ for $a, b \in \mathbb{R}$, then $\phi(f) = a + bi$, which is not 0 in $\mathbb{C}$ unless $a = b = 0$ and so $f = 0$. On the other hand $\phi(x^2 + 1) = 0$ and so $x^2 + 1 \in I$. By Proposition 10.10(2), since $x^2 + 1$ is an element of minimal degree among nonzero elements of I, we must have $I = (x^2 + 1)$.

Moreover, ϕ is clearly surjective, since $a + bi = \phi(a + bx)$. Thus from the first isomorphism theorem we conclude that $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$. This shows how to "construct" $\mathbb{C}$ from $\mathbb{R}$ in some sense. Also, we see that $(x^2 + 1)$ must be a maximal ideal of $\mathbb{R}[x]$.

Example 10.12. Consider the map $\phi : \mathbb{Z}[i] \to \mathbb{Z}_5$ given by $\phi(a + bi) = \overline{a + 2b}$. An easy calculation shows that ϕ is a homomorphism of rings. It is clear that ϕ is surjective. Let $I = \ker \phi$. By the first isomorphism theorem, $\mathbb{Z}[i]/I \cong \mathbb{Z}_5$. So I is a maximal ideal because $\mathbb{Z}_5$ is a field.

We know that $I = (x)$ is prinicpal, generated by $x = a + bi$ with minimal value of $N(x) = a^2 + b^2$ among nonzero elements of I. We see that $\phi(2 - i) = 0$ and so $2 - i \in I$, with $N(2 - i) = 5$. The

only nonzero elements with a smaller norm are $(\pm 1 \pm i)$, ± 1, and $\pm i$, none of which is in I. Thus $I = (2 - i)$ and we conclude that $\mathbb{Z}[i]/(2 - i) \cong \mathbb{Z}_5$.

Euclidean domains are our only examples of PIDs so far, so one may well wonder whether every PID must be a Euclidean domain. The answer is no: the quadratic integer ring $\mathcal{O}_{\mathbb{Q}(\sqrt{-19})} = \mathbb{Z} + \mathbb{Z}((1 + \sqrt{-19})/2)$ is a PID which is not Euclidean; see Dummit and Foote, sections 8.1, 8.2. We view this as mostly a curiosity, as many quadratic integer rings are not PIDs at all, and so the more advanced techniques of Dedekind domains must be used to study them anyway. And the simple examples of PIDs of greatest importance in this first course—in particular the polynomial ring $F[x]$ where F is a field—are Euclidean.

We show now that in an arbitrary PID we have a theory of divisors, gcds, and lcms which behaves very analogously to the familiar special case of $\mathbb{Z}$.

Definition 10.13. Let R be an integral domain. We write $d|b$ for $d, b \in R$ and say d *divides* b if $b = cd$ for some $c \in R$. Given $a, b \in R$, we say that $d \in R$ is a *greatest common divisor* or *gcd* of a and b if (i) $d|a$ and $d|b$; and (ii) for any $c \in R$ such that $c|a$ and $c|b$, then $c|d$. If d is a gcd of a and b then we write $d = \gcd(a, b)$.

Traditionally when working in the ring of integers $\mathbb{Z}$, one insists that gcds should be positive; with this convention there is a unique gcd of two integers a and b (not both 0), and this gcd is literally the greatest (i.e. largest) common divisor of a and b. In a general PID, the term "greatest" is maintained, but it has no literal meaning; note that the definition of gcd is made purely in terms of divisibility with no reference to any ordering of the elements. We no longer insist on a unique gcd but just refer to "a" gcd. Even in $\mathbb{Z}$, with our definition above, either 6 or -6 is a gcd of 12 and 18, for example. Note that we also allow $a = b = 0$ in the definition—this is often avoided in $\mathbb{Z}$ because every number is a common divisor of both 0 and 0, so there is no "greatest"; however, $\gcd(0, 0)$ makes sense according to our definition and is equal to 0.

It is useful to recast divisibility in terms of ideals. Note that $d|b$ means $b = cd$ for some $c \in R$, so that $b \in (d)$. Then $(b) \subseteq (d)$ since (b) is the unique smallest ideal containing b. Conversely, if $(b) \subseteq (d)$ then $b \in (b) \subseteq (d)$ and so $b = cd$ for some c. We conclude that $d|b$ if and only if $b \in (d)$ if and only if $(b) \subseteq (d)$. This means that d is a common divisor of a and b if and only if $(b) \subseteq (d)$ and $(a) \subseteq (d)$, or equivalently $(a) + (b) = (a, b) \subseteq (d)$. So d is a greatest common divisor of a and b if for all principal ideals (c) with $(a, b) \subseteq (c)$, we have $(d) \subseteq (c)$. In other words, $d = \gcd(a, b)$ is equivalent to the statement that the ideal (d) is uniquely minimal among principal ideals that contain (a, b).

As mentioned above, $d = \gcd(a, b)$ (when it exists) is not uniquely determined, However, as the discussion in the previous paragraph makes clear, the ideal (d) generated by the gcd is uniquely determined by a and b, as it is the uniquely minimal principal ideal containing (a, b). Thus the other possible choices of $\gcd(a, b)$ are exactly the other elements d' such that $(d') = (d)$. Let us tease out further exactly how this can happen.

Definition 10.14. Let R be an integral domain. We say that a is an *associate* of b if $a = ub$ for some unit $u \in R$.

A quick argument shows that the relation "a is an associate of b" is an equivalence relation. We often say that "a and b are associates" without preferencing one over the other.

Lemma 10.15. *Let R be any integral domain. Then $(a) = (b)$ if and only a and b are associates.*

Proof. Suppose that $(a) = (b)$. If $a = 0$ then (a) is the zero ideal and so $b = 0$, and vice versa. Obviously a and b are associates in this case.

Now assume that a and b are nonzero. Since $a \in (a) = (b)$, we have $a = bx$ for some $x \in R$. Similarly, since $b \in (b) = (a)$ we have $b = ay$ for $y \in R$. Hence $a = bx = ayx$ and so $a(yx - 1) = 0$. Since R is a domain and $a \neq 0$, we get $yx = 1$ and thus x is a unit. Thus a and b are associates.

Conversely, if $a = ub$ for some unit u, then for any $r \in R$ we have $ar = b(ur) \in (b)$, so $(a) \subseteq (b)$. But $b = u^{-1}a$ and thus $(b) \subseteq (a)$ by the same argument. We conclude that $(a) = (b)$. □

In particular, we see that the set of possible gcd's of a pair of elements a, b is an equivalence class of associates. For example, $\mathbb{Z}^\times = \{-1, 1\}$, so in the integers the only freedom is the sign of the gcd. In the Gaussian integers $\mathbb{Z}[i]$ the units are $\{\pm 1, \pm i\}$ and so the set of associates of an element $a + bi$ is $\{\pm a \pm bi\}$.

Let us return to PIDs now.

Proposition 10.16. *Let R be PID. Given elements $a, b \in R$, then $d = \gcd(a, b)$ exists, and moreover $(d) = (a, b) = (a) + (b)$. Thus $d = ax + by$ for some $x, y \in R$.*

Proof. Since R is a PID, $(a, b) = (d)$ for some d. Thus since $(a, b) = (d)$ is already principal, clearly (d) is uniquely minimal among principal ideals containing (a, b). That $d = ax + by$ for some $x, y \in R$ is just a restatement of $d \in (a, b)$. □

We note that in an integral domain R which is not a PID, it is possible that a pair of elements a, b has a gcd d, but that $(a, b) \subsetneq (d)$. It is also possible that no gcd of those elements exist, as we will see in Example 10.35.

It is also easy to develop of theory of least common multiple (lcm) in an integral domain. In any PID R, the lcm of any 2 elements a, b exists, and if $m = \operatorname{lcm}(a, b)$ then $(m) = (a) \cap (b)$. Moreover, one has the nice formula $(ab) = (\gcd(a, b) \operatorname{lcm}(a, b))$ as one gets in the integers, or in terms of elements, ab and $\gcd(a, b) \operatorname{lcm}(a, b)$ are associates. We leave this to the exercises.

10.2.1. *Calculating the GCD.* In this optional section we describe how one might calculate GCDs in practice.

Since a Euclidean domain is a PID, gcd's always exist in a Euclidean domain. Assuming that there is an algorithm for computing q and r such that $a = qb + r$ with $r = 0$ or $d(r) < d(b)$, then there is an algorithm for calculating the gcd, modelled on the Euclidean algorithm for finding the gcd of two integers. Suppose that R is Euclidean with respect to the norm function $d : R \to \mathbb{N}$. Given $a, b \in R$ with $b \neq 0$, we can find q, r such that $a = qb + r$, where $d(r) < d(b)$ or $r = 0$. Note that $r = a - qb \in (a, b)$, so $(r, b) \subseteq (a, b)$. Conversely, $a = qb + r \in (b, r)$, so $(a, b) \subseteq (b, r)$. We see that $(a, b) = (b, r)$ and thus $\gcd(a, b) = \gcd(b, r)$.

Now in general, given a, b for which we want to find a gcd, assume both are nonzero, since $\gcd(0, b) = b$ is trivial to calculate. Let $0 \neq a_1 = a, 0 \neq a_2 = b$, and calculate $a_1 = q_1 a_2 + a_3$ as above, with $d(a_3) < d(a_2)$ or $a_3 = 0$. Then $\gcd(a_1, a_2) = \gcd(a_2, a_3)$. If $a_3 \neq 0$, continue in this way, writing $a_2 = q_2 a_3 + a_4$, with $d(a_4) < d(a_3)$ or $a_4 = 0$. We create a sequence $a_1, a_2, a_3, \ldots, a_n$ for which $d(a_{i+1}) < d(a_i)$ for all $i \geq 2$. Necessarily there is n such that $a_n = 0$ but $a_i \neq 0$ for $i < n$. Then $\gcd(a, b) = \gcd(a_1, a_2) = \gcd(a_2, a_3) = \cdots = \gcd(a_{n-1}, a_n) = \gcd(a_{n-1}, 0) = a_{n-1}$. So the last nonzero term of the sequence is a gcd of a and b. It is also possible to use the results of this calculation to find explicit $x, y \in R$ such that $ax + by = \gcd(a, b)$. For the last two nontrivial steps gave $a_{n-3} - q_{n-3} a_{n-2} = a_{n-1}$ and $a_{n-4} - q_{n-4} a_{n-3} = a_{n-2}$. Substituting the second in the first we obtain

$$a_{n-1} = a_{n-3} - q_{n-3}(a_{n-4} - q_{n-4} a_{n-3}) = (1 + q_{n-3} q_{n-4}) a_{n-3} + (-q_{n-3}) a_{n-4}.$$

Continuing inductively in this way we obtain an explicit expression for a_{n-1} as an R-linear combination of a_{n-i} and a_{n-i+1} for all $i \leq n - 1$; when $i = n - 1$ we get a_{n-1} as an R-linear combination of a and b.

Example 10.17. Let $R = \mathbb{Q}[x]$. Let us calculate $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1)$. Each step of the Euclidean algorithm can be performed by polynomial long division with remainder (we leave the details of these calculations to the reader). Let $a_1 = x^5 - x^2 + 5x - 5$ and $a_2 = x^4 - 1$. Then $x^5 - x^2 + 5x - 5 = x(x^4 - 1) + (-x^2 + 6x - 5)$, so set $a_3 = -x^2 + 6x - 5$. Now $x^4 - 1 = (-x^2 - 6x - 31)(-x^2 + 6x - 5) + (156x - 156)$, so set $a_4 = 156x - 156$. Next, $-x^2 + 6x - 5 =$

$(-(1/156)x + 5/156)(156x - 156) + 0$. So $a_5 = 0$ and $a_4 = 156x - 156$ is the gcd. Since nonzero scalars are units in $\mathbb{Q}$, $x - 1$ is also a gcd. So $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1) = x - 1$.

10.2.2. *Exercises.*

Exercise 10.18. Let R be an integral domain. We take m *is a multiple of* a to mean the same thing as a divides m, i.e. $a|m$. The element m is a *least common multiple* of a and b if (i) $a|m$ and $b|m$; and (ii) for all $x \in R$ such that $a|x$ and $b|x$, we have $m|x$. We write $m = \operatorname{lcm}(a, b)$ in this case.

(a). Show that m is a least common multiple of a and b if and only if (m) is uniquely maximal among principal ideals contained in $(a) \cap (b)$.

(b). Prove that a and b have a least common multiple if and only if a and b have a greatest common divisor, and that in this case $(ab) = (\gcd(a, b) \operatorname{lcm}(a, b))$.

(c). Show that in a PID, $m = \operatorname{lcm}(a, b)$ exists for any elements a, b, and $(m) = (a) \cap (b)$.

Exercise 10.19. A *Bezout domain* is an integral domain R in which every ideal generated by 2 elements is principal; that is, given $a, b \in R$ we have $(a, b) = (d)$ for some d.

(a). Prove that an integral domain R is a Bezout domain if and only if every pair of elements a, b has a GCD $d \in R$ such that $d = ax + by$ for some $x, y \in R$.

(b). Prove that every finitely generated ideal of a Bezout domain is principal.

Exercise 10.20. Use the calculation in Example 10.17 to write find $u(x), v(x) \in \mathbb{Q}[x]$ such that $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1) = u(x)(x^5 - x^2 + 5x - 5) + v(x)(x^4 - 1)$.

10.3. **Unique Factorization Domains (UFD's).** We now study factorization of elements in an integral domain as products of simpler elements. We will see that there is a large class of rings for which factorization behaves in a similar way as the factorization of integers as products of primes in $\mathbb{Z}$.

Definition 10.21. Let R be an integral domain. Let a be element of R with $a \neq 0$ and a not a unit. We say that a is *irreducible* if whenever $a = bc$ in R, then either b or c is a unit in R. We say that a is *prime* if whenever $a|(bc)$ then $a|b$ or $a|c$.

Example 10.22. Let $R = \mathbb{Z}$. Since the units in $\mathbb{Z}$ are just ± 1, a is irreducible in $\mathbb{Z}$ if the only ways to write a in $\mathbb{Z}$ as a product of other elements are $a = (1)(a)$ or $a = (-1)(-a)$. Clearly this holds if and only if $a = \pm p$ for a prime number p.

If $a = \pm p$ for a prime number p, then Euclid's lemma states that if $a|bc$ then $a|b$ or $a|c$, so a is a prime element in $\mathbb{Z}$. Conversely if a is a composite number, then $a = bc$ where $|b| < |a|$ and $|c| < |a|$, and so $a|(bc)$ but clearly $a \not| b$ and $a \not| c$, so a is not a prime element.

We conclude that the irreducible and prime elements in $\mathbb{Z}$ are the same, both consisting of the numbers $\pm p$ for prime numbers p.

We see that both prime and irreducible elements are reasonable ways to try to generalize the idea of a prime number in the integers. It turns out that they give distinct concepts in arbitrary integral domains, which is why it is useful to study both of them. This is actually a common situation in algebra: when trying to generalize a concept, there may be several different but equivalent ways to formulate the original idea, where the natural generalizations of these different ways lead to distinct notions in the more general setting. Sometimes one of the generalizations is clearly the most useful one to consider; other times they all give potentially interesting concepts worth investigating. In the case at hand, we will see that in rings where factorization behaves best (unique factorization domains), prime and irreducible will turn out to be equivalent concepts.

Example 10.23. Let F be a field and let $R = F[x]$. An irreducible element of R is called an *irreducible polynomial*. Note that if $\deg f = 1$ then f is irreducible; for if we write $f = gh$, then $\deg f = \deg g + \deg h$, and there is no choice but to have $\deg g = 1$ and $\deg h = 0$ or $\deg g = 0$ and $\deg h = 1$. Since the polynomials of degree 0 are the nonzero constants, which are units in R, either g or h is a unit.

The polynomial $x^2 + 1$ is not irreducible in $\mathbb{C}[x]$, since $x^2 + 1 = (x - i)(x + i)$ in this ring, and neither $x - i$ or $x + i$ is a unit since only the nonzero constant polynomials are units. On the other hand, $x^2 + 1$ is irreducible in $\mathbb{R}[x]$, which we can see as follows. if not, it clearly would be a product of two degree 1 polynomials in $\mathbb{R}[x]$, say $x^2 + 1 = (ax + b)(cx + d)$. Since $bd = 1$, b and d are nonzero, so $x^2 + 1 = ac(x + b/a)(x + d/c)$, but $ac = 1$, so $x^2 + 1 = (x + r)(x + s)$ for $r, s \in \mathbb{R}$. Now we must have $r + s = 0$ and $rs = 1$, leading to $r(-r) = 1$ or $r^2 = -1$, which has no solution with $r \in \mathbb{R}$.

Example 10.24. Let $R = \mathbb{Z}[i]$. We claim that $3 \in \mathbb{Z}[i]$ is irreducible. If we write $3 = xy$, then $N(3) = N(x)N(y)$ as the norm $N(a + bi) = a^2 + b^2$ is multiplicative. Thus $9 = N(x)N(y)$. No element in R has norm 3, since $a^2 + b^2 = 3$ clearly has no solutions in integers. Thus either $N(x) = 1$ or $N(y) = 1$. However, an element of norm 1 in R is a unit.

We are now ready to define the rings with well-behaved factorization.

Definition 10.25. Let R be an integral domain. Then R is a *unique factorization domain* or *UFD* if

(1) Every element $a \in R$ which is nonzero and not a unit has an expression $a = p_1 p_2 \dots p_n$ for some $n \geq 1$ where each p_i is irreducible in R.

(2) If $p_1 p_2 \dots p_n = q_1 q_2 \dots q_m$ where each p_i and q_j is irreducible, then $n = m$ and possibly after rearranging the q_i, p_i is an associate of q_i for all i.

Example 10.26. $\mathbb{Z}$ is a UFD. The irreducibles in $\mathbb{Z}$ are the primes and their negatives. It is a familiar theorem that any positive number greater than 1 has a unique expression as a product of positive primes; this extends in an obvious way to all nonzero, nonunit integers if we allow all prime elements and only require uniqueness up to associates. For example, $10 = (2)(5) = (-5)(-2)$ are two factorizations of 10 as products of irreducibles, but after rearrangement the two factorizations are the same up to associates.

In a general integral domain, asking for any two factorizations to be the same "up to associates" is the best we can hope for. For, note that if p is an irreducible and u is a unit, then pu is again an irreducible which is an associate of p. Thus, for example, any product of two irreducibles $p_1 p_2$ is also the product of irreducibles p_1', p_2' where $p_1' = up_1$, $p_2' = u^{-1}p_2$ for any unit u, so this kind of ambiguity cannot be avoided. So the definition of UFD captures those domains in which every nonzero, nonunit element can be written as a product of irreducibles in a way that is as unique as we can reasonably ask for.

Our next main goal is prove that any PID is also a UFD. We will see later that the class of UFD's is considerably more general than the class of PIDs. We first need some preliminary results. Here are some basic properties of prime and irreducible elements.

Lemma 10.27. *Let R be an integral domain.*

(1) *$a \in R$ is a prime element if and only if (a) is a nonzero prime ideal of R.*

(2) *If a is prime, then a is irreducible.*

(3) *If R is a PID, then a is prime if and only if a is irreducible, if and only if (a) is maximal and not zero. Thus all nonzero prime ideals are maximal.*

Proof. (1) This follows more or less from the definitions. If (a) is a nonzero prime ideal, then by definition (a) is proper so a is not a unit. If $a = bc$ then $bc \in (a)$, so either $b \in (a)$ or $c \in (a)$ and thus $a|b$ or $a|c$. Thus a is a prime element. The converse is similar.

(2) Suppose that a is prime, so $a \neq 0$ and a is not a unit. If $a = bc$ then $a|(bc)$ so either $a|b$ or $a|c$. If $a|b$, then $b = ad$, say, so $a = adc$ and $a(1 - dc) = 0$. Since we are in a domain, $cd = 1$ and thus c is a unit. By symmetry, if $a|c$ we conclude that b is a unit.

(3) Now let R be a PID. If a is an irreducible element, consider (a). Since by definition a is not a unit, (a) is a proper ideal. If $(a) \subseteq I \subseteq R$ for some ideal I, we can write $I = (b)$ for some b. Then $b|a$, so $a = bc$. Since a is irreducible, either b or c is a unit. If b is unit, then $(b) = R$. If c is a unit, then a and b are associates and $(a) = (b)$. We see that either $I = (a)$ or $I = R$ and hence (a) is maximal ideal, which is nonzero since $a \neq 0$. Now any nonzero maximal ideal (a) is a nonzero prime ideal, and hence a is a prime element by (1). Finally a prime element is irreducible by (2). □

We see from the result above that the picture of the prime ideals in a PID is quite simple. Note that a field F is trivially a PID, and in this case (0) is maximal and the only prime ideal of F; F has no prime or irreducible elements and the previous result is vacuous. If R is a PID which is not a field, then it has some nonzero proper ideal and hence at least one nonzero maximal ideal. Then (0) is the only prime of R which is not maximal, and all of the other primes are maximal ideals (a) generated by irreducible elements a. There is one maximal ideal for each associate equivalence class of irreducible elements. In general the set of prime ideals of a commutative ring, considered as a poset under inclusion, is called its *prime spectrum*.

10.3.1. *the noetherian property.* The final element we need for the proof that PIDs are UFDs is the following notion which is very important in the theory of rings and modules in general. We take a small detour to explore this concept a bit beyond what we technically need at this point.

Definition 10.28. Let R be a commutative ring. Then R is called *noetherian* if given a chain of ideals I_i of R for all $i \geq 1$ with $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots \subseteq I_n \subseteq \ldots$, then there exists n such that $I_m = I_n$ for all $m \geq n$ (we say the chain *stabilizes*). This condition is also known as the *ascending chain condition* or *ACC* as well as the noetherian property.

Note that only chains indexed by the natural numbers are needed here; these are not the general chains (totally ordered sets) considered in Zorn's Lemma. It is important to remember that it does not suffice to consider chains of this special sort when verifying the hypothesis of Zorn's lemma.

The term noetherian honors Emmy Noether, a German mathematician who in her last years moved to America and taught at Bryn Mawr college. She was one of the most important figures in the development of commutative ring theory in the early twentieth century. As it turns out many of the rings one naturally tends to encounter in practice are noetherian; the fact that the condition is so common is one of the things that makes it the most useful. It is easy to prove this for PIDs.

Lemma 10.29. *A PID is a noetherian ring.*

Proof. Let $I_1 \subseteq I_2 \subseteq \dots$ be a chain of ideals in the PID R. Then $I = \bigcup_{i \geq 1} I_i$ is again an ideal of R. Since R is a PID, $I = (a)$ for some a. Now $a \in I_n$ for some n. Then for $m \geq n$, we have $(a) \subseteq I_n \subseteq I_m \subseteq I = (a)$ and so $I_n = I_m$ for all $m \geq n$. Thus the chain stabilizes and R is noetherian. □

Let us prove several different characterizations of the noetherian property, all of which are useful and interesting.

Proposition 10.30. *Let R be a commutative ring. The following are equivalent:*

(1) *R is noetherian; i.e. R has the ascending chain condition on ideals.*

(2) *Every nonempty collection of ideals of R has a maximal element (under inclusion).*

(3) *Every ideal I of R is finitely generated, i.e. $I = (a_1, \dots, a_k)$ for some $a_i \in R$.*

Proof. (1) $\implies$ (2). Let S be some nonempty collection of ideals of R. Suppose that S has no maximal element. Pick any $I_1 \in S$. Since I_1 is not a maximal element of S under inclusion, there must be $I_2 \in S$ with $I_1 \subsetneq I_2$. Now I_2 is also not maximal in S, so there is $I_3 \in S$ with $I_2 \subsetneq I_3$. Continuing inductively, we have an ascending chain $I_1 \subsetneq I_2 \subsetneq I_3 \subsetneq \cdots \subsetneq I_n \subsetneq \dots$, which shows that the ascending chain condition fails.

(2) $\implies$ (3). Let I be an ideal of R. Consider the collection S of all finitely generated ideals of R which are contained in I. Note that this is a nonempty collection since $(0) \subseteq I$. Now by hypothesis S has a maximal element $J \subseteq I$, say with $J = (a_1, \dots, a_k)$. Suppose that $J \subsetneq I$. Pick any $a_{k+1} \in I \setminus J$. Then $J \subsetneq (a_1, \dots, a_k, a_{k+1}) \subseteq I$, which shows that J was not maximal after all. This contradiction implies that $J = I$ and so I is finitely generated.

(3) $\implies$ (1). This is similar to the proof of Lemma 10.29; indeed, that proof could have been skipped as this result is more general. If $I_1 \subseteq I_2 \subseteq \dots$ is a chain of ideals, then $I = \bigcup_{i \geq 1} I_i$ is an ideal of R, so $I = (a_1, \dots a_k)$ for some $a_i \in R$, by condition (3). Now each a_i is contained in some I_j; since the ideals form a chain, there is n such that $a_i \in I_n$ for all i. Then for $m \geq n$ we have $(a_1, \dots, a_k) \subseteq I_n \subseteq I_m \subseteq I = (a_1, \dots, a_k)$ and so $I_n = I_m$ for all $m \geq n$. □

Condition (2) in the previous result is called the *maximal condition*. It is useful to compare it with Zorn's Lemma. Our study of applications of Zorn's Lemma showed why it is useful to be able to choose maximal elements of posets. Zorn's Lemma potentially applies to posets of ideals in arbitrary commutative rings, but in order to apply it one needs that poset to satisfy the condition that chains have upper bounds. Some posets of ideals of interest do not satisfy this condition, and so Zorn's Lemma cannot be used. In a noetherian ring, any poset of ideals has a maximal element

and so we never need to use Zorn's Lemma, but instead we have restricted the kind of ring that our results apply to.

Condition (3) shows that in some sense noetherian rings generalize PIDs. The definition of a PID, where every ideal must be generated by one element, is generalized to the weaker condition that every ideal must be generated by some finite set of elements.

10.3.2. *PIDs are UFDs.* We are now ready to prove the main goal of this section, that PIDs have the unique factorization property. In fact, we are able to prove a somewhat more general statement.

Theorem 10.31. *Let R be an integral domain.*

(1) *Suppose that R is noetherian, and that all irreducibles in R are prime. Then R is a UFD.*
(2) *If R is a PID, then R is a UFD.*

Proof. (1) We first have to show that if a is a nonzero, nonunit element of R, then a can be written as a finite product of irreducibles. Consider the set of ideals

$$S = \{(a) | a \text{ is nonzero, nonunit, and not a finite product of irreducibles}\}.$$

Suppose that the collection S is nonempty. Since R is noetherian, it satisfies the maximal condition (condition (2) in Proposition 10.30) and so S has a maximal element, say (a). Now a is not itself irreducible (note that we consider a single irreducible to be a "product" of 1 irreducible) and so we can write $a = bc$ where b and c are both not units. Then $(a) \subsetneq (b)$, for if $(a) = (b)$, then c would be forced to be a unit. Similarly, $(a) \subsetneq (c)$. Since (a) is a maximal element of S, neither (b) nor (c) belongs to S, and neither b nor c is zero or a unit. Thus b and c are both finite products of irreducibles. But then $a = bc$ is a finite product of irreducibles as well, a contradiction. It follows that $S = \emptyset$ and so every nonzero nonunit element of R is a finite product of irreducibles.

Now suppose that $p_1 p_2 \dots p_m = q_1 q_2 \dots q_n$, where each p_i and q_j is irreducible, and hence also prime by hypothesis. Note that we allow the case that $m = 0$ or $n = 0$, so that one or the other product is empty and by convention equal to 1. We prove by induction on m that $m = n$, and after relabeling the q_j we have p_i is an associate of q_i for all i. If $m = 0$ then we have $1 = q_1 q_2 \dots q_n$; if $n \neq 0$, then each q_i is irreducible and a unit, a contradiction. So $n = 0$ and there is nothing further to show. Now we assume $m \geq 1$; similarly, this forces $n \geq 1$. Since p_1 is prime, the definition of prime extends by induction to prove that since $p_1 | q_1 q_2 \dots q_n$, we have $p_1 | q_i$ for some i. Relabel the q's so that q_i becomes q_1. Now $p_1 | q_1$ means $q_1 = p_1 x$, but since q_1 is irreducible, either p_1 or x is a unit. The element p_1 is irreducible and hence not a unit, so x is a unit and p_1, q_1 are associates.

Since we are in a domain, We may now cancel p_1 from both sides to get $p_2p_3 \dots p_m = (xq_2)q_3 \dots q_n$ (some product could be empty). Since x is a unit and q_2 is irreducible, xq_2 is irreducible. By induction we obtain that $m-1 = n-1$ and possibly after relabeling, p_i is an associate of q_i for all i (note that an associate of xq_2 is also an associate of q_2). Since we already showed that p_1 is an associate of q_1, we are done.

(2) We proved that PID's are noetherian in Lemma 10.29, and that irreducible elements are prime in a PID in Lemma 10.27. Thus (1) applies and shows that a PID is a UFD. □

10.3.3. *Properties of UFDs.* Some of the nice properties we proved for PIDs in the preceding section hold for general UFD's. First, we have that there is no distinction between irreducible and prime elements.

Lemma 10.32. *Let R be a UFD. Then $a \in R$ is prime if and only if it is irreducible.*

Proof. We already saw that a prime element in an integral domain is irreducible in Lemma 10.27.

Now let a be irreducible. Suppose that $a|(bc)$. Write $bc = ad$ for some $d \in R$. Write $b = p_1p_2 \dots p_m$, $c = q_1q_2 \dots q_n$, and $d = r_1r_2 \dots r_t$, for some irreducibles p_i, q_i, and r_i. Now we have $ar_1r_2 \dots r_t = p_1p_2 \dots p_mq_1q_2 \dots q_n$. By the uniqueness condition in the definition of UFD, we must have that a is an associate of some p_i or some q_i. Then $a|b$ or $a|c$, and so a is a prime element. □

For the next result and other applications it is useful to make the following observation. Suppose that $a = p_1p_2 \dots, p_k$ is a product of irreducible elements p_i. Some of the p_i may be associates of each other; if we multiply these together we will get a unit multiple of a power of a single p_i. Doing this for each class of associates and renaming the irreducibles, we get $a = uq_1^{e_1}q_2^{e_2} \dots q_m^{e_m}$ for some $e_i \geq 1$, where q_i and q_j are not associates for $i \neq j$, and for some unit u. By the uniqueness property of the UFD, we get that this expression for a is unique up to replacing the q_i with associates and changing the unit u. Note that the unit u cannot be removed in general as it cannot necessarily be "absorbed" into a prime power. For example, in $\mathbb{Z}$ we have $-36 = (-1)(2^2)(3^2)$, and replacing 2 by -2 or 3 by -3, the only possible associates, does not remove the unit in front.

Now we can also easily get that gcd's exist in a UFD.

Lemma 10.33. *Let R be a UFD. Then for every pair of elements $a, b \in R$, $\gcd(a, b)$ exists.*

Proof. If $a = 0$ then $\gcd(0, b) = b$. If a or b is a unit then $(a, b) = R$ and so $1 = \gcd(a, b)$. So we can assume that a and b are nonzero, nonunits, and thus we can express each as a unit times a product of powers of pairwise non-associate irreducibles. In fact, if we make the convention that $p^0 = 1$ for any irreducible p, then we can write each of a and b using the same overall set of irreducibles by

taking the union of all associate classes of irreducibles that appear in either a or b. In this way we can write $a = up_1^{e_1}p_2^{e_2}\dots p_m^{e_m}$ and $b = vp_1^{f_1}p_2^{f_2}\dots p_m^{f_m}$ where the p_i are pairwise non-associate irreducibles; $e_i \geq 0$ and $f_i \geq 0$, and u, v are units in R. Note that the exponents e_i and f_i are uniquely determined by a and b.

Now define $g_i = \min(e_i, f_i)$ for all i. Then $d = p_1^{g_1}p_2^{g_2}\dots p_m^{g_m}$ is a gcd of a and b. We leave it to the reader to check the details. □

10.3.4. *Examples.* There are many examples of integral domains which are not UFDs. We think the following example is one of the simplest.

Example 10.34. Let F be a field. Let

$$R = \{f \in F[x] | f(x) = a_0 + a_1x + a_2x^2 + \dots + a_mx^m \text{ with } a_1 = 0\}.$$

It is easy to check that R is a subring of $F[x]$, as we never create a nonzero x-term by multiplying or adding polynomials without an x-term. R is a domain since it is a subring of a domain.

Now R contains no polynomials of degree 1. Hence if $f \in R$ has degree 2 or 3, if we write $f = gh$ for $g, h \in R[x]$, then $\deg f = \deg g + \deg h$ forces either $\deg g = 0$ or $\deg h = 0$. But R contains all of the scalars in $F[x]$ and so every nonzero element in R with degree 0 is a unit. It follows that all elements in R with degree 2 or degree 3 are irreducible in R.

Now $x^6 = (x^2)(x^2)(x^2) = (x^3)(x^3)$ gives two factorizations of $x^6 \in R$ as a product of irreducibles, where the number of irreducibles is not even the same in the two expressions. Thus R is not a UFD.

Most quadratic integer rings are not UFDs, so these are also an easy source of examples of non-UFDs. The following is one example, but there are lots of similar ones.

Example 10.35. Let $R = \mathcal{O}_{\mathbb{Q}(\sqrt{-10})}$. Thus $R = \mathbb{Z}[\sqrt{-10}] = \{a + b\sqrt{-10} | a, b \in \mathbb{Z}\}$ since -10 is not congruent to 1 modulo 4. In this ring we have the norm $N(a + b\sqrt{-10}) = a^2 + 10b^2$. Since an element is a unit if and only if it has norm 1, it is clear that R has group of units $R^\times = \{\pm 1\}$.

Note that $-10 = (-2)(5) = (\sqrt{-10})(\sqrt{-10})$ in R. We claim that -2, 5, and $\sqrt{-10}$ are all irreducibles in R. Because we know the units in R it is clear that none of these are associates of each other, so this will then imply that factorization in R is not unique.

Since $N(-2) = 4$, if $-2 = xy$ with $x, y \in R$ both nonunits, since $N(-2) = N(x)N(y)$ we must have $N(x) = N(y) = 2$. But $a^2 + 10b^2 = 2$ has no solutions. So -2 is irreducible in R. Similarly, there are no elements of norm 5 and so 5 is irreducible in R. If $\sqrt{-10} = xy$ with x and y nonunits, then $N(x)N(y) = 10$ and again if x and y are to be nonunits then $N(x) = 2$ and $N(y) = 5$ or vice

versa; but we know there are no elements of such norms. Thus -2, 5, and $\sqrt{-10}$ are all irreducible as claimed. We conclude that R is not a UFD.

We can also see that R has irreducible elements which are not prime (which gives an additional proof that R is not a UFD, by Lemma 10.32). We already saw that 5 is irreducible and that $5|(\sqrt{-10})(\sqrt{-10})$. Suppose that 5 is prime. Then $5|\sqrt{-10}$. But if $\sqrt{-10} = 5x$ for $x \in R$ then taking norms we get $10 = 25N(x)$ which is clearly impossible. So 5 is an irreducible element which is not prime. Similar arguments show that 2 and $\sqrt{-10}$ also have this property.

Using the same idea we can also give an example of a pair of elements in an integral domain which have no greatest common divisor. Let $a = 10$ and $b = 2\sqrt{-10}$. One may check that both principal ideals (2) and $(\sqrt{-10})$ contain (a, b) and are minimal among principal ideals containing it. Thus there is no uniquely minimal principal ideal containing (a, b).

10.3.5. *Exercises.*

Exercise 10.36. Finish the proof of Lemma 10.33.

Exercise 10.37. Let R be an integral domain. Let X be a multiplicative system in R not containing 0, and let $D = RX^{-1}$. Show that if R is a Euclidean domain, so is D. (Hint: use factorization into irreducibles to define the norm function.)

Exercise 10.38. Let $G = (\mathbb{R}_{>0}, \cdot)$ be the group of positive real numbers under multiplication. Then G is an *ordered group*: it is a totally ordered set such that if $\alpha < \beta$ and $\gamma \in G$ then $\alpha\gamma < \beta\gamma$. Let F be any field and let FG be the group ring. Let R be the subset of FG consisting of the F-span of $\mathbb{R}_{\geq 1}$. It is easy to see that R is a subring of FG.

(a). Prove that R is an integral domain, and the only units in the ring R are those of the form $\lambda 1_{\mathbb{R}}$, where $0 \neq \lambda \in F$.

(b). Show that any element x in the F-span of $\mathbb{R}_{>1}$ is a product of two elements in $\mathbb{R}_{>1}$. Conclude that no such element can be written as a finite product of irreducibles. Thus R is not a UFD.

(c). Show that R is not noetherian, and find an explicit properly ascending chain of ideals in R.

Exercise 10.39. Let n be a squarefree integer with $n > 3$ and let $R = \mathbb{Z}[\sqrt{-n}] = \{a + b\sqrt{-n} | a, b \in \mathbb{Z}\}$. (Note this is different from the ring of integers $\mathcal{O}_{\mathbb{Q}(\sqrt{-n})}$ when $n \equiv 1 \mod 4$).

(a). Prove that 2, $\sqrt{-n}$, $1 + \sqrt{-n}$, and $1 - \sqrt{-n}$ are all irreducible in R.

(b). Show that R is not a UFD.

(c). Find an element in R which is irreducible and not prime.

Exercise 10.40. Consider the ring $R = \mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} | a, b \in \mathbb{Z}\}$, in other words the ring of integers $\mathcal{O}_{\mathbb{Q}(\sqrt{-5})}$.

(a). Consider the ideals $I_2 = (2, 1 + \sqrt{-5})$, $I_3 = (3, 2 + \sqrt{-5})$, $I_3' = (3, 2 - \sqrt{-5})$. Show that $R/I_2 \cong \mathbb{Z}_2$, and $R/I_3 \cong R/I_3' \cong \mathbb{Z}_3$. Conclude that all three ideals are maximal ideals.

(b). Show that $R/(3) \cong \mathbb{Z}_3 \times \mathbb{Z}_3$ as rings. (Hint: Chinese Remainder theorem).

(c). Is $R/(2) \cong \mathbb{Z}_2 \times \mathbb{Z}_2$?

Exercise 10.41. This problem continues the investigations of the ring R in the previous problem.

(a). Prove that I_2, I_3, I_3' are all not principal ideals of R.

(b). Prove that $I_2^2 = (2)$, $I_2 I_3 = (1 - \sqrt{-5})$, $I_2 I_3' = (1 + \sqrt{-5})$, and $I_3 I_3' = (3)$. In particular, this gives multiple examples showing that a product of nonprincipal ideals can be principal.

(c). Consider the equality of products of principal ideals $(2)(3) = (1 + \sqrt{-5})(1 - \sqrt{-5})$. Show that expressing each of the ideals in this equation as a product of maximal ideals, one gets the same result on both sides of the equation up to rearrangement of the ideals.

Remark. The ring R is an example of a Dedekind domain. *Although unique factorization fails in the sense that R is not a UFD, there is a different kind of unique factorization: every nonzero ideal is a product of maximal ideals in a unique way up to the order of the factors. This is demonstrated by part (c): even though the element* 6 *factors in two essentially different ways (hence R is not a UFD), the equality of products of principal ideals* $(2)(3) = (1 + \sqrt{-5})(1 - \sqrt{-5})$ *leads to the same answer once everything is expressed in terms of products of maximal ideals. Dedekind domains are important in algebraic geometry and number theory and we will study them in more detail in Math 200c.*

Exercise 10.42. Suppose that R is a UFD with field of fractions F. A polynomial f is *monic* if it has leading coefficient 1; in other words $f(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + x^n$.

(a). Suppose that $f \in R[x]$ factors as $f = gh$ with $g, h \in F[x]$. Show that the product of any coefficient of g with any coefficient of h is in R.

(b). Suppose that f, g, and h are as in part (a) and that moreover g and h are monic. Show that $g \in R[x]$ and $h \in R[x]$.

(c). Show that the ring $S = \mathbb{Z}[2\sqrt{2}] = \{a + b2\sqrt{2} | a, b \in \mathbb{Z}\}$ is not a UFD by finding $f \in S[x], g, h \in F[x]$, where F is the field of fractions of S, which violate the results above.

10.4. **Applications to number theory.** In this optional section we show how factorization in the Gaussian integers $\mathbb{Z}[i]$ can be used to solve some classical problems in number theory. Factorization in other quadratic integer rings has similar applications.

Let $R = \mathbb{Z}[i]$. The norm function is $N(a+bi) = a^2+b^2 = ||a+bi||^2$. We have seen that an element is a unit in R if and only if it has norm 1, so the units are $\{\pm 1, \pm i\}$. The norm is multiplicative, so if $N(x) = p$ is a prime number, then x must be irreducible: if $x = yz$ then $N(x) = N(y)N(z)$ and so either $N(y)$ or $N(z)$ equals one and hence y or z is a unit.

Consider those integers n for which $N(x) = n$ for some $x \in R$. Then if $x = a + bi$ we have $N(x) = a^2 + b^2 = n$. It follows that if we understand which positive integers are norms, we will have solved the classical problem of which integers can be represented as sums of two squares. The answer was originally given by Fermat (without proof) and Euler first proved it was correct.

The main step is to understand which prime numbers in $\mathbb{Z}$ are irreducible elements of $R = \mathbb{Z}[i]$.

Lemma 10.43. *Let p be a prime integer. Then the following are equivalent:*

(1) *p is reducible in $\mathbb{Z}[i]$.*

(2) *$p = a^2 + b^2$ for $a, b \in \mathbb{Z}$.*

(3) *$p = 2$ or $p \equiv 1 \mod 4$.*

Proof. (1) $\implies$ (2). Suppose that p is reducible, so $p = xy$ where $x, y \in R = \mathbb{Z}[i]$ and neither x nor y is a unit. Now $N(p) = p^2 = N(x)N(y)$ and neither x nor y can have norm 1. Thus $N(x) = N(y) = p$. Writing $x = a + bi$, we have $p = N(x) = a^2 + b^2$ for $a, b \in \mathbb{Z}$.

(2) $\implies$ (3). If $a \in \mathbb{Z}$, then either a is even, so $a^2 \equiv 0 \mod 4$, or a is odd, so $a^2 \equiv 1 \mod 4$. It follows that for any $a, b \in \mathbb{Z}$, $a^2 + b^2$ is congruent to 0, 1, or 2 modulo 4. Thus p is either 2 or an odd prime with $p \equiv 1 \mod 4$.

(3) $\implies$ (1). If $p = 2$ then $p = (1 + i)(1 - i)$ is reducible. Now assume that $p \equiv 1 \mod 4$. We will use in this proof that the units group $\mathbb{Z}_p^\times$ is cyclic of order $p - 1$. It is a consequence of the fact that any finite field has cyclic multiplicative group, which will be proved later. We have that $|\mathbb{Z}_p^\times|$ is a multiple of 4. It therefore has an element $\overline{c}$ which has order exactly 4 in this group. Since $\overline{-1}$ is the unique element with order exactly 2, we have $\overline{c}^2 = \overline{-1}$. Thus $p|(c^2 + 1)$. Now in $\mathbb{Z}[i]$ we have $c^2 + 1 = (c + i)(c - i)$. Suppose that p is irreducible in $\mathbb{Z}[i]$. Since $\mathbb{Z}[i]$ is a PID, then p is a prime element of $\mathbb{Z}[i]$. From $p|(c + i)(c - i)$ we deduce that $p|(c + i)$ or $p|(c - i)$ in $\mathbb{Z}[i]$. But this is clearly false as a multiple of p has the form $p(a + bi) = pa + pbi$ with both real and imaginary part multiples of p. So p is reducible in $\mathbb{Z}[i]$. □

Having proved which prime numbers are sums of 2 squares, it is not hard to characterize which integers are sums of 2 squares in general.

Theorem 10.44. *Let $n > 1$. Then n is a sum of 2 squares in $\mathbb{Z}$ if and only if for every prime p dividing n such that $p \equiv 3 \mod 4$, the maximal power p^e dividing n has even exponent e.*

Proof. Suppose first that $n = a^2 + b^2$ is a sum of 2 squares in $\mathbb{Z}$. Then $n = (a + bi)(a - bi)$ in $R = \mathbb{Z}[i]$. Let p be a prime dividing n with $p \equiv 3 \mod 4$. Let p^e be the maximal power of p dividing n in $\mathbb{Z}$. By Lemma 10.43, p is also an irreducible and hence prime element of R. Let $(a+bi) = p_1 p_2 \dots p_m$ where each p_i is irreducible in R. Now the complex conjugation map $R \to R$ given by $(a+bi) \mapsto \overline{a+bi} = (a-bi)$ is a ring isomorphism, so $(a-bi) = \overline{p_1} \dots \overline{p_m}$ where each $\overline{p_i}$ is irreducible as well. The number n has an irreducible factorization in R of the form $n = p^e q_1 \dots q_s$ where the q_i are non-associates of p. We also have $n = p_1 p_2 \dots p_m \overline{p_1} \dots \overline{p_m}$. If k of the p_i are associates of p, then k of the $\overline{p_i}$ are associates of p, since $p = \overline{p}$. It follows from unique factorization that $e = 2k$.

Conversely, suppose the given condition holds. Factor n in $\mathbb{Z}$ as $n = m p_1^{e_1} p_2^{e_2} \dots p_s^{e_s}$ where the p_i are distinct primes with $p_i \equiv 3 \mod 4$, each e_i is even, and m is not divisible by any prime q with $q \equiv 3 \mod 4$. If we find a and b such that $a^2 + b^2 = m$, then $(ac)^2 + (bc)^2 = mc^2 = n$ where $c = p_1 p_2 \dots p_s$. Moreover, note that if $m_1 = a^2 + b^2$ and $m_2 = c^2 + d^2$ are both sums of squares, then $m_1 m_2 = (a+bi)(a-bi)(c+di)(c-di) = (a+bi)(c+di)(a-bi)(c-di) = (x+yi)(x-yi) = x^2 + y^2$ where $(a+bi)(c+di) = (x+yi)$. It follows that if we write m as a product of primes, we just need to express each prime factor of m as a sum of 2 squares. But this is possible by Lemma 10.43. □

10.4.1. *Exercises.*

Exercise 10.45. Let R be the ring $\mathbb{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} | a, b \in \mathbb{Z}\}$. By using similar arguments as we used to study the Gaussian integers $\mathbb{Z}[i]$, show that the following are equivalent for an odd prime number $p \in \mathbb{Z}$:

(i) p is not irreducible in R.

(ii) $p = a^2 + 2b^2$ for some $a, b \in \mathbb{Z}$.

(iii) $\overline{-2}$ is a square in $\mathbb{Z}_p$.

(By the way, it is also known that -2 is a square mod p as in condition (iii) if and only if p is congruent to either 1 or 3 modulo 8.)

Exercise 10.46. Recall that the *characteristic* of a ring R is the order of the element 1 in the additive group of R, when this is a finite number; otherwise we say that R has characteristic 0. Using the Eisenstein criterion, prove that the following elements are irreducible in the indicated ring.

(a). The element $x^n - p \in (\mathbb{Z}[i])[x]$, where p is an odd prime in $\mathbb{Z}$ and $n \geq 1$.

(b). The element $x^2 + y^2 - 1 \in F[x, y]$, where F is any field of characteristic not 2.

11. Polynomial extensions

11.1. **Gauss's Lemma.** In this section we will prove that if R is a UFD, then so is the polynomial ring $R[x]$. Since this process can be iterated, this produces a large collection of examples of UFDs. On the other hand, we will see that $R[x]$ is not a PID unless R is a field.

The main technical element needed for the proof is a Lemma of Gauss which is interesting in its own right. We begin now with some preliminary results directed towards that result.

Throughout this section we assume that R is a UFD. We would like to understand factorization in $R[x]$ and how it relates to factorization in R. It will turn out to be very useful to let F be the field of fractions of R (which exists since R is a domain), and think of R as a subring of F. Then $R[x]$ is naturally a subring of $F[x]$, and the ring $F[x]$ is a PID as we have seen, and so has a relatively simple factorization theory. We will be able to use factorization in $F[x]$ to help us understand factorization in $R[x]$.

Example 11.1. Let $R = \mathbb{Z}$, so $F = \mathbb{Q}$. Consider $f(x) = 5x-10 \in \mathbb{Z}[x]$. Then $f(x)$ is not irreducible in $\mathbb{Z}[x]$, for this ring has only ± 1 as units, while $f = 5(x-2)$ is a product of 2 irreducible elements in $\mathbb{Z}[x]$. On the other hand, if we consider f as an element of $\mathbb{Q}[x]$, then in this ring 5 is a unit and so is ignored when considering factorization. Then the element $5x - 10$ is already itself irreducible, as is true for any degree 1 polynomial in a polynomial ring over a field.

We see from the preceding example that one of the main differences between factorization in $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ is that there are constant polynomials in $\mathbb{Z}[x]$ that are themselves irreducibles.

Example 11.2. Let $f(x) = x^2 - 5x + 6 \in \mathbb{Z}[x]$. Although this polynomial has integer coefficients, we can consider it as an element of $\mathbb{Q}[x]$. As such, there are many factorizations of it as a product of two linear terms, for example $f(x) = ((2/3)x - (4/3))((3/2)x - (9/2))$. Since any linear polynomial is irreducible in $\mathbb{Q}[x]$, this is a factorization of f as a product of irreducibles in $\mathbb{Q}[x]$. But it doesn't tell us about factorization in $\mathbb{Z}[x]$ because the polynomials have coefficients that are not in $\mathbb{Z}$. On the other hand, we can multiply the first factor by $3/2$ and the second by $2/3$ to obtain $f(x) = (x-2)(x-3)$, which is a factorization in $\mathbb{Z}[x]$. Because no constants in $\mathbb{Z}$ factor out of $x-2$ or $x-3$, it is easy to see that these polynomials are irreducible in $\mathbb{Z}[x]$, so we have found a factorization into irreducibles in $\mathbb{Z}[x]$.

The example above already shows the main idea of Gauss's lemma. If we factor a polynomial in $R[x]$ over $F[x]$, we will see that we will be able to adjust the terms by scalars to get a factorization in $R[x]$.

In the previous section we saw that in a UFD R, $\gcd(a,b)$ is defined (up to associates as always) for any $a, b \in R$. It is easy to extend this definition to define $d = \gcd(a_1, \dots a_n)$ for any elements $a_i \in R$. This is an element such that $d|a_i$ for all i, and if $c|a_i$ for all i, then $c|d$. To show that it exists, one may define it as $\gcd(a_1, a_2, \dots, a_n) = \gcd(\gcd(a_1, \dots, a_{n-1}), a_n)$ by induction and then show it has the required properties. Alternatively, one can generalize Lemma 10.33 directly to the case of finitely many elements.

Definition 11.3. Let $f \in R[x]$ for a UFD R. Write $f = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m$ with $a_m \neq 0$. The *content* of f is $C(f) = \gcd(a_0, a_1, \dots, a_m) \in R$. As usual this is defined only up to associates.

For example, if $f = 12x^2 + 15x - 6 \in \mathbb{Z}[x]$, then $C(f) = 3$ (or -3).

Since a lot of things will hold "up to associates" in this section, we use the notation $a \sim b$ to indicate that elements a, b are associates in the ring R. If we need to emphasize in which ring R the elements are associates, we write $a \sim_R b$.

Lemma 11.4. *Let R be a UFD and let $f, g \in R[x]$. Let $a \in R$.*

(1) $C(af) \sim aC(f)$.

(2) *If* $C(f) \sim 1$ *and* $C(g) \sim 1$ *then* $C(fg) \sim 1$.

(3) $C(fg) \sim C(f)C(g)$.

Proof. (1) It is easy to verify fact that for $a_1, \dots, a_n, b \in R$, $\gcd(ba_1, ba_2, \dots, ba_n) = b \gcd(a_1, \dots, a_n)$. The formula in (1) is an immediate consequence.

(2) To show $C(fg) = 1$, it is enough to prove that for every irreducible element $p \in R$, p does not divide $C(fg)$; in other words, fg has some coefficient not divisible by p. Now let $\phi : R \to R/(p)$ be the natural homomorphism. For $r \in R$ write $\overline{r} = \phi(r) = r + (p)$. We can extend this to a map $\widetilde{\phi} : R[x] \to R/(p)[x]$ defined by $\widetilde{\phi}(f) = \overline{f} = \widetilde{\phi}(a_0 + a_1x + \cdots + a_mx^m) = \overline{a_0} + \overline{a_1}x + \cdots + \overline{a_m}x^m$. It is easy exercise using the definition of the ring operations in a polynomial ring to prove that $\widetilde{\phi}$ is also a homomorphism of rings. Now since $C(f) \sim 1$, p does not divide every a_i, and thus some $\overline{a_i} \neq 0$ in $R/(p)$. It follows that $\overline{f} \neq 0$ in $R/(p)[x]$. Similarly, since $C(g) \sim 1$, $\overline{g} \neq 0$ in $R/(p)[x]$. But now note that since p is irreducible, it is a prime element by Lemma 10.32 and so (p) is a prime ideal. Thus $R/(p)$ is a domain. Then $R/(p)[x]$ is also a domain. Thus $\overline{fg} = \overline{f}\,\overline{g} \neq 0$. It follows that some coefficient of fg is not divisible by p. Since p was arbitrary, $C(fg) \sim 1$ as desired.

(3) We may assume that $f \neq 0$ and $g \neq 0$; otherwise the statement is trivial. Write $f = a_0 + a_1x + \cdots + a_mx^m$ and $g = b_0 + b_1x + \cdots + b_nx^n$. Since $C(f) = \gcd(a_0, a_1, \dots, a_m)$ divides every coefficient a_i, we can write $f = C(f)\widetilde{f}$ where $\widetilde{f} \in R[x]$ has content $C(\widetilde{f}) \sim 1$. Similarly, $g = C(g)\widetilde{g}$

for $\widetilde{g} \in R[x]$ with $C(\widetilde{g}) \sim 1$. Now $fg = C(f)C(g)\widetilde{f}\widetilde{g}$ and so using (1), $C(fg) \sim C(f)C(g)C(\widetilde{f}\widetilde{g})$. But by (2) we have $C(\widetilde{f}\widetilde{g}) \sim 1$. □

We are now ready to prove Gauss's Lemma.

Lemma 11.5 (Gauss)**.** *Let R be a UFD with field of fractions F. Consider $R[x]$ as a subring of $F[x]$. Suppose that $f \in R[x]$ and that $f = gh$ for $g, h \in F[x]$. Then there are is a scalar $0 \neq \lambda \in F$ such that $g' = \lambda g$ and $h' = \lambda^{-1}h$ satisfy $g', h' \in R[x]$ (and of course, $f = g'h'$).*

Proof. Notice that for any $f \in F[x]$, there is $a \in R$ such that $af \in R[x]$. (If $f = (s_1/t_1)+(s_2/t_2)x+\cdots+(s_m/t_m)x^m$ with $s_i, t_i \in R$, then $a = t_1t_2\ldots t_m$ suffices.)

Applying this to both g and h we have $a, b \in R$ such that $ag \in R[x]$ and $bh \in R[x]$. Now $ag = C(ag)\widetilde{g}$ for some $\widetilde{g} \in R[x]$ with $C(\widetilde{g}) \sim 1$. similarly, $bh = C(bh)\widetilde{h}$ for $\widetilde{h} \in R[x]$ with $C(\widetilde{h}) \sim 1$, and $f = C(f)\widetilde{f}$ with $C(\widetilde{f}) \sim 1$. We now have $abC(f)\widetilde{f} = (ag)(bh) = C(ag)C(bh)\widetilde{g}\widetilde{h}$. Taking the content of both sides and using that $C(\widetilde{g}\widetilde{h}) \sim 1$ by Lemma 11.4(2), we get $abC(f) \sim C(ag)C(bh)$. Cancelling gives a unit $u \in R$ such that $\widetilde{f} = u\widetilde{g}\widetilde{h}$ or $f = C(f)\widetilde{g}u\widetilde{h}$. Let $g' = C(f)\widetilde{g} \in R[x]$ and $h' = u\widetilde{h} \in R[x]$. We now get $f = g'h'$ with $g', h' \in R[x]$. Tracking through the proof we see that we only ever adjusted polynomials by scalars in F, so $g' = \lambda_1 g$ and $h' = \lambda_2 h$ with $\lambda_1, \lambda_2 \in F$. Since $f = gh = g'h'$, $\lambda_1\lambda_2 = 1$ so we can take $\lambda_1 = \lambda$, $\lambda_2 = \lambda^{-1}$ for some $\lambda \in F$. □

11.2. **Factorization in $R[x]$.** Gauss's Lemma allows us to understand the irreducibles in $R[x]$ in terms of those of $F[x]$.

Corollary 11.6. *Let R be a UFD with field of fractions F.*

(1) *Let $f \in R[x]$ be a polynomial with $\deg f \geq 1$. Then f is irreducible in $R[x]$ if and only if f is irreducible in $F[x]$ and $C(f) \sim 1$.*

(2) *Let $f, g \in R[x]$ be irreducibles in $R[x]$ of positive degree. Then f and g are associates in $R[x]$ if only if they are associates in $F[x]$.*

Proof. (1) Suppose that f is irreducible in $R[x]$. We can write $f = C(f)f'$ with $f' \in R[x]$. Then $\deg f' = \deg f \geq 1$, so f' is not a unit in $R[x]$. This forces $C(f)$ to be a unit, i.e. $C(f) \sim 1$. Next, suppose we write $f = gh$ for $g, h \in F[x]$. By Gauss's Lemma, we have $f = g'h'$ with $g', h' \in R[x]$, where $g' = \lambda g$ and $h' = \lambda^{-1}h$, some $\lambda \in F$. Since f is irreducible in $R[x]$, either g' or h' is a unit in $R[x]$, which means either $\deg g' = 0$ or $\deg h' = 0$. Then $\deg g = 0$ or $\deg h = 0$. But nonzero constant polynomials are units in $F[x]$, so either g or h is a unit in $F[x]$. Hence f is irreducible over $F[x]$.

Conversely, suppose that $C(f) \sim 1$ and f is irreducible in $F[x]$. Suppose that $f = gh$ with $g, h \in R[x]$. This is a factorization in $F[x]$ as well, so either g or h is a unit in $F[x]$, and hence either $\deg g = 0$ or $\deg h = 0$. Without loss of generality we may suppose that $\deg(g) = 0$, so $g = a \in R$ is a constant polynomial. Then a divides f, so a divides every coefficient of f. Since $C(f) \sim 1$, a is a unit in R. Thus f is irreducible in $R[x]$.

(2) Suppose that f and g are associates in $F[x]$. Then $f = \lambda g$ where $0 \neq \lambda \in F$. Write $\lambda = r/s$ with $r, s \in R$, so $sf = rg$. Now taking contents we have $sC(f) = C(sf) = C(rg) = rC(g)$ but since f and g are irreducible in $R[x]$, $C(f) \sim 1$ and $C(g) \sim 1$ by part (1). Thus $s \sim r$ and hence λ is a unit in R. So f and g are associates in $R[x]$. The converse is trivial. □

We are now ready to prove the main theorem.

Theorem 11.7. *Let R be a UFD. Then $R[x]$ is also a UFD.*

Proof. Let $f \in R[x]$ where f is nonzero and not a unit. We first need to show that f is a product of irreducibles in $R[x]$. We prove this by induction on $\deg f$. If $\deg f = 0$, then $f = r \in R$ for some nonzero nonunit $r \in R$, so $r = p_1 p_2 \dots p_m$ for some irreducibles p_i in R, some $m \geq 1$, since R is a UFD. Clearly each p_i is also irreducible in $R[x]$, so this case is done.

Now assume that $\deg f > 0$. Let $r = C(f)$; so we can write $f = rf'$ with $f' \in R[x]$ where $C(f') \sim 1$. Either r is a unit or else we can factor $r = p_1 p_2 \dots p_m$ as above. So we just need to prove that f' is a product of irreducibles in $R[x]$. If f' is irreducible in $R[x]$ we are done. If f' is reducible in $R[x]$, since $C(f') \sim 1$, by Corollary 11.6, f' is also reducible over $F[x]$, so $f' = gh$ for $g, h \in F[x]$ with $\deg g < \deg f$ and $\deg g < \deg f$. By Gauss's Lemma, we can adjust g and h by nonzero scalars in F to get a factorization $f' = g'h'$ with $g', h' \in R[x]$ and still $\deg g' < \deg f$, $\deg h' < \deg f$. By induction on degree, each of g' and h' is a product of finitely many irreducibles in $R[x]$, so f' is as well.

Next we need to prove uniqueness. Suppose that $p_1 p_2 \dots p_m g_1 g_2 \dots g_n = q_1 q_2 \dots q_s h_1 h_2 \dots h_t$, where p_i, q_i are irreducibles in $R[x]$ of degree 0 (i.e. irreducibles in R) and g_i, h_i are irreducibles in $R[x]$ of degree ≥ 1. Each g_i and h_i must have content 1, by Corollary 11.6. Taking contents of both sides we thus get $p_1 p_2 \dots p_m \sim_R q_1 q_2 \dots q_s$. By unique factorization in the UFD R, we conclude that $m = s$ and p_i is an associate of q_i after relabeling. We can now cancel the degree zero parts to get $g_1 g_2 \dots g_n \sim_{R[x]} h_1 h_2 \dots h_t$. Each g_i and h_i is also irreducible in $F[x]$, by Corollary 11.6. Since $F[x]$ is a UFD, we have $n = t$ and after relabeling g_i is an associate of h_i in $F[x]$ for all i. But then by Corollary 11.6(2), g_i is an associate of h_i in $R[x]$ for all i as well, so we are done. □

The main result of this section implies that there are many examples of rings that are UFDs and not PIDs.

Lemma 11.8. *Let R be a UFD which is not a field. Then $R[x]$ is a UFD and not a PID.*

Proof. The ring $R[x]$ is a UFD by Theorem 11.7. Since R is not a field, it has some irreducible element p. Then we claim that the ideal $I = (p, x)$ is a non-principal ideal of $R[x]$. If $I = (d)$, then $d|p$ and $d|x$. If $p = gd$ then $\deg(p) = 0 = \deg(g) + \deg(d)$ which forces $\deg(d) = 0$, in other words $d \in R$. But now $d|x$ means $x = df$ would force $\deg(f) = 1$, say $f = ax + b$ with $a, b \in R$, and $x = dax + db$. This means $da = 1$ and so d is a unit in R and hence also in $R[x]$. Now $(d) = R$. However, I is not the unit ideal, for $R[x]/(p, x) \cong R/(p)$ is a nonzero integral domain, as p is irreducible and hence not a unit. □

Example 11.9. Given a ring R, we can define inductively a ring of polynomials in n variables over R by $R[x_1, \ldots x_n] = (R[x_1, \ldots, x_{n-1}])[x_n]$. If R is a UFD, then our main theorem gives that $R[x_1, \ldots, x_n]$ is also a UFD for any n. In particular, if F is a field then $F[x_1, \ldots, x_n]$ is a UFD. These rings play an important role in commutative algebra.

Rather than an inductive definition, one can also define $S = R[x_1, \ldots, x_n]$ directly as follows. Let S be the set of all sums of the form $\sum_{(i_1,i_2,\ldots,i_n)\in\mathbb{N}^n} r_{(i_1,i_2,\ldots,i_n)} x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n}$, where $r_{(i_1,\ldots,i_n)} \in R$ is 0 except for finitely many n-tuples $(i_1, \ldots, i_n)$. (Recall that by our convention $0 \in \mathbb{N}$.) In other words, S consists of finite R-linear combinations of *monomials* $x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n}$. Monomials are multiplied in the obvious way, and this extends linearly to a product on S. It is straightforward to see that this ring is isomorphic to the one given by the inductive construction.

11.3. **Irreducible Polynomials.** In this section, we study some results that help one to understand whether or not a particular polynomial is irreducible.

Let F be a field. We know that $R = F[x]$ is a Euclidean domain, so it is a PID and UFD and every nonzero nonunit polynomial is a product of irreducible polynomials. But how do we determine which polynomials are irreducible? This is a hard problem in general that depends sensitively on the properties of the field F. Here we will state some of the most basic results which we will need when we study field theory in more detail later.

The following result is elementary from the point of view of our earlier study of Euclidean domains.

Lemma 11.10. *Let $f \in F[x]$ where F is a field. Given $a \in F$, we have $f = q(x-a) + r$ where $q \in F[x]$ and $r = f(a) \in F$. In other words $f(a)$ is the remainder when f is divided by $(x-a)$. In particular, $f(a) = 0$ if and only if $(x-a)|f$ in $F[x]$.*

Proof. We know that $F[x]$ is a Euclidean domain with respect to the function $d : F[x] \to \mathbb{N}$ given by $d(0) = 0$, $d(f) = \deg(f)$ for $f \neq 0$. Since $g = (x-a)$ has degree 1, we have $f = qg + r$ with $d(r) < d(g) = 1$ or $r = 0$. Thus $d(r) = 0$ and hence r is a constant. Now since evaluation at a is a homomorphism, we must have $f(a) = r(a) = r$. The last statement follows since q and r are unique. □

The fact that the remainder when we divide f by $(x-a)$ is equal to $f(a)$ is often called the "remainder theorem", and the fact that $(x-a)|f$ if and only if $f(a) = 0$ is often called the "factor theorem". We say that $a \in F$ is a *root* of $f \in F[x]$ if $f(a) = 0$.

Corollary 11.11. *A polynomial $f \in F[x]$ with $\deg(f) = n$ has at most n distinct roots in F.*

Proof. If $a \in F$ is a root of f then $f = (x-a)g$ with $g \in F[x]$ of $\deg g = n - 1$, by the factor theorem. If $b \neq a$ is also a root of f then $0 = f(b) = (b-a)g(b)$ forces $g(b) = 0$. But g has at most $n-1$ roots in F by induction. □

There are a few fields F for which we can say exactly what the irreducible polynomials in $F[x]$ look like.

Example 11.12. Let $F = \mathbb{C}$. By the fundamental theorem of algebra, which we will prove later in the course, every $f \in F[x]$ with $\deg f \geq 1$ factors as $f = c(x-a_1)\dots(x-a_n)$ for some $c, a_1, \dots, a_n \in \mathbb{C}$. It follows that the only irreducible elements in $\mathbb{C}[x]$ are the linear polynomials $\{x - a | a \in \mathbb{C}\}$ (up to associates).

Similarly, if $F = \mathbb{R}$ all irreducibles in $\mathbb{R}[x]$ can be described. Up to associates, they are the linear polynomials $x - a$ with $a \in \mathbb{R}$ and the quadratic polynomials $x^2 + ax + b$ with $a, b \in \mathbb{R}$ that have non-real roots. We leave this to the reader to check (use the fact that any polynomial factors into linear factors over $\mathbb{C}$, and that for a polynomial with real coefficients the complex roots come in conjugate pairs.)

Corollary 11.13. *Let $f \in F[x]$ where F is a field, with $\deg f \geq 2$.*

(1) *If f has a root in F then f is reducible in $F[x]$.*
(2) *If $\deg f \in \{2, 3\}$, then f is reducible in $F[x]$ if and only if f has a root in F.*

Proof. (1) If $f(a) = 0$ for $a \in F$ then $(x - a)$ divides f by the factor theorem, so $f = (x - a)g$ for some $g \in F[x]$. Since $\deg f \geq 2$, $\deg g \geq 1$. Thus f is reducible since the units in $F[x]$ are just the nonzero constant polynomials.

(2) Let f have degree 2 or 3. If f is reducible, it must be a product of polynomials of strictly smaller degree, so one of those polynomials has degree 1. Thus $(tx - s)$ divides f for some $s, t \in F$ with $t \neq 0$, and so the associate $(x - a)$ divides f, where $a = s/t \in F$. Thus a is a roof of f. The converse is part (1). □

A method for proving that a polynomial over a field is or is not irreducible is called an *irreducibility test*. We know that nonzero degree 0 polynomials in $F[x]$ are units; degree 1 polynomials are always irreducible, and for polynomials of degree 2 and 3, there is a simple test: it is irreducible if and only if it has no roots in F. Note however that a reducible polynomial of degree 4 could be a product of 2 irreducible polynomials of degree 2, and so needn't have a root in F.

To use this test for irreducibility of polynomials of degree 2 or 3 we need ways to tell if a polynomial has roots in the field or not. Here is a useful result in that regard.

Lemma 11.14. *Let R be a UFD with field of fractions F. Let $f = a_0 + a_1 x + \cdots + a_m x^m \in R[x]$. If $r \in F$ is a root of f, where $r = s/t$ with $s, t \in R$, $t \neq 0$ and $\gcd(s, t) = 1$, we must have $s | a_0$ and $t | a_m$ in R.*

Proof. If $f(r) = 0$ we have $0 = f(r) = a_0 + a_1(s/t) + \cdots + a_m(s/t)^m$. Multiplying by t^m we have $0 = a_0 t^m + a_1 s t^{m-1} + \cdots + a_{m-1} s^{m-1} t + a_m s^m$. This equation implies $s | a_0 t^m$. Since $\gcd(s, t) = 1$, we get $s | a_0$. Similarly, the equation implies $t | a_m s^m$ and since $\gcd(s, t) = 1$ we have $t | a_m$. □

The preceding result is often called the "rational root theorem", since it is frequently used to decide if $f \in \mathbb{Q}[x]$ has a root by taking $F = \mathbb{Q}$, $R = \mathbb{Z}$. Note that we can first clear denominators in f to assume that $f \in \mathbb{Z}[x]$, without affecting the roots of f.

Example 11.15. Let $f(x) = (3/2)x^3 + x - 5 \in \mathbb{Q}[x]$. Then f has the same roots as the polynomial $3x^3 + 2x - 10 \in \mathbb{Z}[x]$. By the rational root theorem, if $s/t \in \mathbb{Q}$ is a fraction in lowest terms which is a root of f, then $s | 10$ and $t | 3$. This gives a finite number of possible solutions $s = \pm 1, \pm 2, \pm 5, \pm 10$ and $t = \pm 1, \pm 3$. Checking all of them, no such fraction s/t is a root of f. Thus f has no roots in $\mathbb{Q}$ and hence f is irreducible in $\mathbb{Q}[x]$ because $\deg f = 3$.

Example 11.16. If F is a finite field, for example $F = \mathbb{F}_p$ for a prime p, then we can check if a polynomial of degree 2 or 3 in $F[x]$ has a root in F just by evaluating at all the finitely many elements of F. This allows one to find irreducible polynomials of higher degree inductively; for

example, once one finds all irreducible polynomials of degree 2 and 3, then we know all products of two degree 2 irreducibles and we can also find all degree 4 polynomials with a root. The degree 4 irreducibles are the remaining degree 4 polynomials. Similarly, we could find all degree 5 irreducibles by eliminating those with a root and the products of a degree 2 and a degree 3 irreducible. This method is quite easy if F is small and we are interested in polynomials of low degree.

For example, let $F = \mathbb{F}_2 = \{0, 1\}$. There are 4 polynomials of degree 2, and only $x^2 + x + 1$ does not have 0 or 1 as a root. So this is the only irreducible of degree 2. Similarly, the only degree 3 polynomials without a root are x^3+x+1 and x^3+x^2+1, so these are the degree 3 irreducibles. The degree 4 polynomials without a root are x^4+x^3+1, x^4+x^2+1, x^4+x+1, and $x^4+x^3+x^2+x+1$. The only product of 2 degree 2 irreducibles is $(x^2+x+1)^2 = x^4+x^2+1$; so x^4+x^3+1, x^4+x+1, and $x^4 + x^3 + x^2 + x + 1$ are the degree 4 irreducibles.

For polynomials of degree bigger than 3 over a general field, the methods above may not help. The following criterion due to Eisenstein only applies to polynomials of a fairly special form, but it does allow one to write down a lot of irreducible polynomials of arbitrarily high degree.

Proposition 11.17 (Eisenstein Criterion)**.** *The R be a UFD with field of fractions F. Suppose that $f = a_m x^m + \cdots + a_1 x + a_0 \in R[x]$ is a polynomial of degree ≥ 1. If there is an irreducible element $p \in R$ such that $p \not| a_m$; $p | a_i$ for $0 \leq i \leq m-1$; and $p^2 \not| a_0$, then f is irreducible in $F[x]$.*

Proof. Suppose that f is reducible in $F[x]$. Then $f = gh$ where $g, h \in F[x]$ both have degree ≥ 1. By Gauss's lemma (Lemma 11.5), adjusting by scalars if necessary, we can assume that $g, h \in R[x]$. Let $\overline{R} = R/(p)$ and consider the homomorphism $\phi : R[x] \to \overline{R}[x]$ given by $f = \sum b_i x^i \mapsto \overline{f} = \sum \overline{b_i} x^i$, where $\overline{b_i} = b_i + (p)$. Then $\overline{f} = \overline{g}\overline{h}$. Now by assumption every coefficient of f except a_m is a multiple of p, so $\overline{f} = \overline{a_m} x^m$ with $\overline{a_m} \neq 0$. Let $g = \sum b_i x^i$ and $h = \sum c_i x^i$ and suppose that $\deg g = k$, $\deg h = l$, where $k + l = m = \deg f$. Let i be minimal such that $\overline{b_i} \neq 0$ and let j be minimal such that $\overline{c_j} \neq 0$. Then since $R/(p)$ is a domain, $\overline{b_i}\overline{c_j} x^{i+j}$ is the smallest degree term with nonzero coefficient in $\overline{g}\overline{h} = \overline{f}$. But $\overline{f}$ has no nonzero coefficients except the coefficient of x^m, and this forces $i = k$ and $j = l$, so that $\overline{g} = \overline{b_k} x^k$ and $\overline{h} = \overline{c_l} x^l$. In particular, since $k > 0$ and $l > 0$, $\overline{b_0} = \overline{c_0} = 0$. But then $p | b_0$ and $p | c_0$ in R, and the constant term of f is $a_0 = b_0 c_0$, so $p^2 | a_0$. This contradicts the assumption. □

Example 11.18. $f(x) = 5x^7 + 3x^6 - 9x^3 + 6$ is irreducible in $\mathbb{Q}[x]$, by applying the Eisenstein criterion with $R = \mathbb{Z}$ and $p = 3$. While we are primarily interested in irreducbility over a field here, we can also say that f is irreducible in $\mathbb{Z}[x]$, since f has content $\gcd(5, 3, -9, 6) = 1$ (see Corollary 11.6).

Note that it was trivial to choose the polynomial in the previous example—we just had to make sure the leading coefficient was not a multiple of 3, the other coefficients were multiples of 3, and the constant term was not a multiple of 9. The other prime factors of the coefficients could be anything at all, so one immediately gets an infinite collection of irreducible polynomials this way.

It is quite useful that the ring R can be any UFD at all in the Eisenstein criterion. Here is an application to polynomials in two variables.

Example 11.19. Let $f = x + x^2y^{n-1} + y^n \in F[x,y] = (F[x])[y]$, where F is a field. We claim that f is an irreducible element in $F[x,y]$. To see this we embed $R = F[x]$ in its field of fractions $K = F(x)$, and consider $f \in K[y]$. Now we can consider f as a polynomial in y over the field $K = F(x)$. The element x is irreducible in $R = F[x]$. Writing $f = (1)y^n + (x^2)y^{n-1} + (x)y^0$ we see that x does not divide the leading coefficient in R, it divides the other coefficients, and x^2 does not divide the constant term. Thus Eisenstein's criterion applies and shows that f is an irreducible polynomial in $F(x)[y]$. Then f is also irreducible in $F[x][y] = F[x,y]$ by Corollary 11.6 since $\gcd(x, x^2, 1) = 1$.

There is a particularly useful polynomial which can be proved irreducible using a tricky application of the Eisenstein criterion.

Example 11.20. Let p be a prime. Then $f = x^{p-1} + x^{p-2} + \cdots + x + 1$ is irreducible in $\mathbb{Q}[x]$.

Proof. The trick is to make a substitution. Note that $f = (x^p - 1)/(x - 1)$. Substitute $z + 1$ for x where z is another variable. We obtain

$$g(z) = f(z+1) = ((z+1)^p - 1)/z = (z^p + \tbinom{p}{p-1} z^{p-1} + \cdots + \tbinom{p}{1} z + 1 - 1)/z = z^{p-1} + \tbinom{p}{p-1} z^{p-2} + \cdots + \tbinom{p}{1},$$

by the binomial theorem. The binomial coefficient $\binom{p}{i}$ is a multiple of p whenever $0 < i < p$, and $\binom{p}{1} = p$ is not a multiple of p^2. The Eisenstein criterion applies to $g(z)$ for the prime p, so $g(z)$ is irreducible in $\mathbb{Q}[z]$. But clearly then $f(x)$ is irreducible in $\mathbb{Q}[x]$. □

The substitution method above sometimes applies to other polynomials, but it is not easy to predict when a polynomial might satisfy the Eisenstein criterion after a substitution.

We mention one more method for proving irreducibility, though we may not need to use it much. It involves a similar idea as the Eisenstein criterion, but simpler.

Proposition 11.21 (Reduction mod p)**.** *Let R be a UFD with field of fractions F. Let $f = a_n x^n + \cdots + a_1 x + a_0 \in R[x]$. Suppose that p is prime in R and that $p \not| a_n$; let $\overline{R} = R/(p)$. Let $\phi : R[x] \to \overline{R}[x]$ be the homomorphism $g \to \overline{g}$ which reduces coefficients mod p.*

If $\overline{f}$ is irreducible in $\overline{R}[x]$, then f is irreducible in $F[x]$.

Proof. If f is reducible in $F[x]$, then using Gauss's Lemma (as in the proof of Proposition 11.17), we have $f = gh$ with $g, h \in R[x]$ and $\deg g, \deg h \geq 1$. Thus $\overline{f} = \overline{g}\overline{h}$ in $\overline{R}[x]$. Since $p \not| a_n$, $\overline{f}$ still has degree n. Since $n = \deg f = \deg g + \deg h = \deg \overline{g} + \deg \overline{h}$ and $\deg \overline{g} \leq \deg g$, $\deg \overline{h} \leq \deg h$, this forces $\deg \overline{g} = \deg g \geq 1$, $\deg \overline{h} = \deg h \geq 1$. But then $\overline{f} = \overline{g}\overline{h}$ contradicts that $\overline{f}$ is irreducible in $\overline{R}[x]$. □

Example 11.22. Let $f = x^4 + x + 2 \in \mathbb{Z}[x]$. We use reduction mod p to prove that f is irreducible in $\mathbb{Q}[x]$. We need to choose a p such that reducing mod p gives an irreducible polynomial in $\mathbb{F}_p[x]$. Obviously $p = 2$ won't work as the constant term will die, so we try $p = 3$. Consider $\overline{f} = x^4 + x + 2 \in \mathbb{F}_3[x]$. Clearly this polynomial has no root in $\mathbb{F}_3 = \{0, 1, 2\}$. Following the method of Example 11.16, one may find all degree 2 irreducibles in $\mathbb{F}_3[x]$ and show that $\overline{f}$ is not a product of 2 degree 2 irreducibles. Thus $\overline{f}$ is irreducible in $\mathbb{F}_3[x]$ and hence f is irreducible in $\mathbb{Q}[x]$ by Proposition 11.21.

Remark 11.23. There exist polynomials $f \in \mathbb{Z}[x]$ which are irreducible but for which the reduction mod p method fails for all primes p, as $\overline{f} \in \mathbb{F}_p[x]$ is always reducible. A simple example is $f(x) = x^4 + 1$.

12. Modules

12.1. **Definition and first examples.** Let R be a ring. In our initial study of modules, we will not assume that R is commutative. The concept of a left R-module is a "linearization" of the concept of a left group action on a set. We saw that studying group actions had a lot of consequences for the structure of the groups themselves. Similarly, to get a deeper understanding of rings, modules are essential.

Definition 12.1. Let R be a ring. A *left R-module* is an abelian group $(M, +)$ together with a left action of R on M, that is, a function $f : R \times M \to M$ where we write $f(r, m) = r \cdot m$, such that for all $r, s \in R$ and $m, n \in M$,

(i) $r \cdot (s \cdot m) = (rs) \cdot m$;

(ii) $1 \cdot m = m$;

(iii) $r \cdot (m + n) = r \cdot m + r \cdot n$;

(iv) $(r + s) \cdot m = r \cdot m + s \cdot m$.

Notice that axioms (i) and (ii) are the same as for the action of a group on a set (although R is just a monoid, not a group, under multiplication). However, the set being acted on in this case is

assume to be an abelian group, and the other two axioms are kind of generalized distributive laws. Namely, (iii) shows that each element of R acts linearly on M, and (iv) shows that the additive structure of the ring R is compatible with the action.

It is easy to check that the module axioms also force $0 \cdot m = 0$ and $-1 \cdot m = -m$ for all $m \in M$. When the module under discussion is clear, we often just write rm instead of $r \cdot m$ for the action of $r \in R$ on $m \in M$.

Just as it was possible to define a right action of a group on a set, there is also a notion of a right R-module, defined using a function $f' : M \times R \to M$ given by $f'(m, r) = m \cdot r$ and with the obvious right-sided versions of axioms $(i) - (iv)$. In particular, axiom (i) becomes $(i)' : (m \cdot s) \cdot r = m \cdot (sr)$.

In group theory, recall that left and right actions on a set are essentially equivalent concepts, and there is a natural way to turn any left action into a right action; namely, if G acts on X on the left, then there is a right action $*$ with $x * g = g^{-1}x$. In module theory, on the other hand, a left module cannot easily be turned into a right module over the same ring, in general. There is something we can do, however.

Definition 12.2. Given a ring R, its *opposite ring* R^{op} is the ring with the same underlying abelian group as R, but with new product $*$ defined by $r * s = sr$.

It is easy to check that the opposite ring is a ring. Now if M is a left R-module, then we can define a right R^{op}-module structure on the same abelian group M by $m \cdot r = rm$ (where we identify the underlying sets of R and R^{op}). The main thing to observe is that for axiom $(i)'$, we get $(m \cdot s) \cdot r = (sm) \cdot r = r(sm) = (rs)m = m \cdot (rs) = m \cdot (s * r)$ as required. Note that this would not work if we did not use the opposite multiplication $*$.

The rings R and R^{op} are not isomorphic for a general ring, and so left and right modules are distinct concepts that might behave quite differently. When R is commutative, however, of course $R = R^{op}$. In this case, given a left R-module M, we can freely turn it into a right R-module by acting on the other side, i.e. $(m \cdot r = rm)$.

We now give some important examples of modules.

Example 12.3. Let F be a field. A left F-module consists of an abelian group V together with an action of F on V which we call scalar multiplication in this case. Examining the module axioms, we see that the F-module V is exactly the same as a vector space over the field F.

Example 12.4. For any ring R, R is a left module over itself by left multiplication, i.e. with $r \cdot s = rs$. In this case module axiom (i) is the actual associativity of multiplication in R, and module axioms (iii) and (iv) are the actual distributive laws in the ring.

Similarly, R is a right module over itself by right multiplication.

Example 12.5. Let F be a field and let $V = F^n$ be the set of column vectors $\left\{ \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} | a_i \in F \right\}$. Then V is a left module over the $n \times n$-matrix ring $R = M_n(F)$, where $A \cdot v = Av$.

Similarly, the set of length n row vectors with entries in F is a right $M_n(F)$-module by right matrix multiplication.

Example 12.6. Let M be any abelian group. Given $n \in \mathbb{Z}$, recall that for $m \in M$ we have defined the nth multple of m by

$$nm = \begin{cases} \overbrace{m + m + \cdots + m}^{n} & n > 0 \\ 0 & n = 0 \\ \overbrace{(-m) + (-m) + \cdots + (-m)}^{|n|} & n < 0 \end{cases} \tag{12.7}$$

where these integer multiples of m are the additive analogs of the powers of an elements in a group. This defines a natural action of $\mathbb{Z}$ on M, and one can easily check that the module axioms hold. Conversely, given a $\mathbb{Z}$-module M, of course the underlying set of M is an abelian group, and the module axioms easily imply that the action of $n \in \mathbb{Z}$ on m must be given by (12.7).

In conclusion, a $\mathbb{Z}$-module is nothing more than an abelian group; the $\mathbb{Z}$-action comes for free and is uniquely determined. This is very useful because the theory of abelian groups will be subsumed into module theory, and our theorems about modules will have interesting applications to abelian groups.

Example 12.8. Let $\phi : R \to S$ be any ring homomorphism. Suppose that M is a left S-module. Then we can make M into an R-module by "restriction of scalars": for $m \in M$, $r \in R$, define $r \cdot m = \phi(r)m$. The module axioms are immediate. This is called restriction of scalars since in the common special case where ϕ is the inclusion homomomorphism of a subring R into a ring S, then we really are just restricting the action to a smaller ring.

This raises the question of whether given a left R-module, there is a natural way to make it into an S-module using the homomorphism ϕ. The answer is yes, but this is not nearly so simple to define—it will require the theory of tensor products we develop later.

12.2. **Basic module technology.** As with any new algebraic structure, we want to have a basic theory of functions that preserve the structure, definitions of substructures and factor structures, and so on. We make these definitions for left modules, but there are obvious counterparts for right modules.

Definition 12.9. Let M and N be left R-modules. A function $f : M \to N$ is a *homomorphism of modules* if f is a homomorphism of abelian groups, and $f(rm) = rf(m)$ for all $r \in R$ and $m \in M$. If a homomorphism f is bijective it is called an *isomorphism*.

Example 12.10. Let $R = F$ be a field. If V and W are F-modules, that is vector spaces over F, then a function $f : V \to W$ is a homomorphism of F-modules if and only if it is a linear transformation of vector spaces over F.

Example 12.11. Let R be a left module over itself by left multiplication. For any fixed $x \in R$, the function $\phi_x : R \to R$ given by $\phi_x(r) = rx$ is a homomorphism of left R-modules. It is a homomorphism of abelian groups by one of the distributive laws in R, and for $s \in R$, $\phi_x(sr) = (sr)x = s(rx) = s\phi_x(r)$, so ϕ_x preserves the left R-action.

The map ϕ_x is called "right multiplication by x" for obvious reasons. Note that left multiplication by x will not be a left module homomorphism in general (unless R is commutative, or more generally if x is in the center of the ring R).

Example 12.12. We saw above that a $\mathbb{Z}$-module is just an abelian group with its canonical $\mathbb{Z}$-action. If $f : M \to P$ is a homomorphism of abelian groups, it is automatically a homomorphism of $\mathbb{Z}$-modules. For $n \in \mathbb{Z}$ and $m \in M$, the fact that $f(nm) = nf(m)$ follows from the properties of homomorphisms of groups.

Definition 12.13. Let M be a left R-module. A subset $N \subseteq M$ is a *submodule* of M is N is a subgroup of M under $+$, and for all $r \in R, x \in N$, we have $rx \in N$.

Thus a submodule of M is closed under $+$ and under the left R-action. Clearly a submodule N of M is an R-module in its own right under the same R-action restricted to N. Also, the inclusion map $i : N \to M$ is an R-module homomorphism.

Example 12.14. Let M be a left R-module. Then both $\{0\}$ and M are submodules of M. $\{0\}$ is called the *trivial submodule* and may be written as 0 for simplicity.

Example 12.15. Let $f : M \to P$ be a homomorphism of left R-modules. Then it is a homomorphism of groups and so we have the kernel defined as $\ker f = \{m \in M | f(m) = 0\}$ like usual. Then $\ker f$ is a submodule of M: it is an additive subgroup of M by group theory, and if $m \in \ker f$ and $r \in R$, then $f(rm) = rf(m) = r0 = 0$, so $rm \in \ker f$.

The image of f, $f(M) = \{x \in P | x = f(m) \text{ for some } m \in M\}$ is a submodule of P, since if $x = f(m)$, then $rx = rf(m) = f(rm)$.

Example 12.16. Let R be a left module over itself by left multiplication. Then a submodule of R is an additive subgroup I of R such that $rx \in I$ for all $r \in R$ and $x \in I$. This is what we called a *left ideal* when we studied rings. So left ideals of R are the same as submodules of R as a left module over itself.

Example 12.17. If F is a field and V is an F-module, in other words a vector space over F, then a submodule of V is the same as a subspace of V as defined in linear algebra.

Example 12.18. Let $V = F^n$ be the set of length n column vectors over the field F, considered as a left module over the ring $R = M_n(F)$ by left matrix multiplication. We claim that the only R-submodules of V are 0 and V. To see this, suppose that $0 \neq v \in V$. It is a standard result of linear algebra that given any vector $w \in V$, there is some matrix $A \in M_n(F)$ such that $Av = w$. Indeed, if the ith entry of v is nonzero we can find such an A which is 0 except along its ith column. It follows that any submodule of V which contains v will be all of V. Since v was an arbitary nonzero vector, this proves the claim.

Definition 12.19. A left R-module M is called *simple* or *irreducible* if its only submodules are 0 and M.

We just saw an example of a simple module over $M_n(F)$, where F is a field.

Definition 12.20. Let N be a submodule of a left R-module M. Then the quotient of abelian groups M/N is an R-module via the action $r \cdot (m+N) = rm+N$. M/N is called a *quotient module* or *factor module*.

As usual, one must check that the definition of the R-action on M/N makes sense. We know that every subgroup of an abelian group is normal, so M/N is certainly a well-defined additive abelian group. Now if $m + N = m' + N$, then $(m - m') \in N$, so $r(m - m') \in N$ since N is a submodule. But then $rm - rm' \in N$ and so $rm + N = rm' + N$. Thus the R-action is well-defined. Once one has well-definedness the module axioms follow routinely.

We also note that the quotient map $\pi : M \to M/N$ given by $\pi(m) = m + N$ is a homomorphism of modules with kernel N.

One nice aspect of module theory is that the substructures of a modules which are modules in their own right, submodules, are also the same things that you can mod out by to get a factor module.

Example 12.21. Let R be a ring with a left ideal I. Then I is a submodule of R, considered as a left module via multiplication. Thus we have a factor module R/I with action $r \cdot (s + I) = rs + I$.

Note if I is just a left ideal (not an ideal), then R/I is not in general a ring, it is only a left R-module.

There are versions for modules of all of the basic homomorphism theorems. Here is the 1st isomorphism theorem.

Theorem 12.22. *Let $f : M \to N$ be a homomorphism of left R-modules, and let $P = \ker f$. Then there is an isomorphism of R-modules $\overline{f} : M/P \to f(M)$ given by $\overline{f}(m + P) = f(m)$.*

Proof. The 1st isomorphism theorem for groups tells us that $\overline{f}$ is well-defined and an isomorphism of abelian groups. We just need to check that $\overline{f}$ is a homomorphism of modules. But this is easy, since $\overline{f}(r(m + P)) = \overline{f}(rm + P) = f(rm) = rf(m) = r\overline{f}(m + P)$. □

Similarly, there are versions of the 2nd, 3rd, and 4th isomorphism theorems. For each one, one takes the corresponding isomorphism theorem for abelian groups and simply notes that everything works at the level of R-modules. We omit the statements here but will freely use these results when we need them.

12.3. **Additional structures on Hom.**

Definition 12.23. Let R be a ring and let M and N be left R-modules. We define

$$\mathrm{Hom}_R(M, N) = \{f : M \to N | f \text{ is a homomorphism of modules over } R\}.$$

A priori, $\mathrm{Hom}_R(M, N)$ is just a set of functions. However, it naturally has additional structure, and this is very useful.

First, we note that $\mathrm{Hom}_R(M, N)$ is always again an abelian group. For this, given $f, g \in \mathrm{Hom}_R(M, N)$ we define a function $f + g \in \mathrm{Hom}_R(M, N)$ by $[f + g](m) = f(m) + g(m)$. This is sometimes called *pointwise* addition of functions, since for each element $m \in M$ (a "point") we simply define the sum of functions at that point by summing the images of that point under the two functions, using that N is an abelian group. Note that $f + g$ is again an R-module homomorphism, since

$$[f + g](m_1 + m_2) = f(m_1 + m_2) + g(m_1 + m_2) = f(m_1) + f(m_2) + g(m_1) + g(m_2)$$
$$= f(m_1) + g(m_1) + f(m_2) + g(m_2) = [f + g](m_1) + [f + g](m_2)$$

and

$$[f + g](rm) = f(rm) + g(rm) = rf(m) + rg(m) = r(f(m) + g(m)) = r[f + g](m).$$

The identity element of $\mathrm{Hom}_R(M,N)$ is the identically zero function 0, and $-f$ is the function with $[-f](m) = -f(m)$. The group axioms for $\mathrm{Hom}_R(M,N)$ are immediate, and $\mathrm{Hom}_R(M,N)$ is again abelian since N is.

Now suppose that R is commutative. We claim that in this case $\mathrm{Hom}_R(M,N)$ even has an R-module structure. It is an abelian group as above, and we define for $r \in R$, $f \in \mathrm{Hom}_R(M,N)$ the function $rf \in \mathrm{Hom}_R(M,N)$ by $[rf](m) = rf(m)$, using the R-module structure of N. It is routine to check that rf respects addition, and note that $[rf](sm) = rf(sm) = rsf(m) = srf(m) = s[rf](m)$, so $rf \in \mathrm{Hom}_R(M,N)$. We have used here that R is commutative. The module axioms for $\mathrm{Hom}_R(M,N)$ are routine to check.

When R is not commutative, in general $\mathrm{Hom}_R(M,N)$ has no natural additional structure beyond being an abelian group. There are ways to give it a module structure when M or N is a *bimodule*; we will explore this in a homework exercise.

Suppose now that R is an arbitrary ring again. When $M = N$ we call a homomorphism of R-modules $f : M \to M$ an *endomorphism*. We may write $\mathrm{Hom}_R(M,M)$ as $\mathrm{End}_R(M)$. In this special case $\mathrm{End}_R(M)$ again has additional structure besides its abelian group structure: it is naturally a ring, called the *endomorphism ring* of M. The product is defined by composition: for $f, g \in \mathrm{End}_R(M)$ we let $fg = f \circ g$. It is obvious that a composition of two module endomorphisms of M is again an endomorphism. The ring axioms follow routinely. For the sake of example, let's check one of the distributive laws $(f+g)h = fh + gh$ for $f, g, h \in \mathrm{End}_R(M)$. Since it is two functions that are being claimed equal, we check by applying them to an arbitrary element of M. We have

$$[(f+g)h](m) = [(f+g) \circ h](m) = (f+g)(h(m)) = f(h(m)) + g(h(m))$$
$$= (f \circ h)(m) + (g \circ h)(m) = [f \circ h + g \circ h](m) = [fh + gh](m)$$

and so $(f+g)h = fh + gh$. The reader should check the other ring axioms to convince themselves that nothing complicated is going on.

If R is commutative and M is an R-module, then $\mathrm{End}_R(M)$ is both an R-module and a ring, by the constructions above. This is a structure called an *R-algebra* which will be defined later.

Example 12.24. Let F be a field and let V be an n-dimensional vector space over F. Then $\mathrm{End}_F(V)$ consists of all F-linear transformations from V to itself. As we saw above, $\mathrm{End}_F(V)$ is a ring. Suppose we fix a basis $v_1, v_2, \ldots, v_n$ for V. Given any $f \in \mathrm{End}_F(V)$, we have scalars $a_{ij}^f \in F$ defined by $f(v_j) = \sum_{i=1}^n a_{ij}^f v_i$. These form a matrix $(a_{ij}^f) \in M_n(F)$. This gives a map $\psi : \mathrm{End}_F(V) \to M_n(F)$, where $\psi(f) = (a_{ij}^f)$. One may check that ψ is an isomorphism of rings.

This isomorphism does depend on the choice of fixed basis; there is no canonical or preferred isomorphism between the two rings.

Example 12.25. Let R be a left module over itself by left multiplication. We will show that $\mathrm{End}_R(R)$ is isomorphic as a ring to the ring R^{op}.

Define a map $\phi : \mathrm{End}_R(R) \to R^{op}$ by $\phi(f) = f(1)$, where we identify the underlying sets of R and R^{op}, so that $f(1) \in R = R^{op}$.

Then we claim that ϕ is a homomorphism of rings. The map ϕ is clearly additive, since $\phi(f+g) = [f+g](1) = f(1) + g(1) = \phi(f) + \phi(g)$. Now $\phi(fg) = [f \circ g](1) = f(g(1))$. Since f is a module homomorphism, $f(r) = f(r \cdot 1) = rf(1)$ for all r. Thus $f(g(1)) = g(1)f(1) = \phi(g)\phi(f) = \phi(f) * \phi(g)$, where $*$ is the multiplication in R^{op}. This proves the claim.

Finally, if $\phi(f) = 0$, then $f(1) = 0$ and so $f(r) = rf(1) = 0$ for all r, and $f = 0$, so ϕ is injective. If $r \in R$, then we have the "right multiplication by r" map $f : s \mapsto sr$ which we have seen is an element of $\mathrm{End}_R(R)$; and $\phi(f) = f(1) = r$. So ϕ is surjective. Thus ϕ is an isomorphism of rings.

12.4. **Modules as maps to endomorphism rings.** Recall from our study of groups that there were two ways of thinking about a (left) action of a group on a set X. In the definition, one focuses on the action of $g \in G$ on one element $x \in X$ at a time. The other point of view thinks of how g acts on all of X at once as a permutation of X, and puts the whole action together into a homomorphism of groups $G \to \mathrm{Sym}(X)$.

A module is like a linearization of a group action, and in fact a module can also be thought of in terms of a single homomorphism (of rings, in this case).

Theorem 12.26. *Let R be a ring and let M be a fixed abelian group. There is a bijective correspondence*

$$\{\textit{left } R-\textit{module structures on } M\} \xrightarrow{\Phi} \{\textit{(unital) ring homomorphisms } \theta : R \to \mathrm{End}_{\mathbb{Z}}(M)\}$$

where given an R-module structure on M, Φ sends it to the map $\theta : R \to \mathrm{End}_{\mathbb{Z}}(M)$ where $[\theta(r)](m) = r \cdot m$.

We hope the reader sees the similarity between this result and the corresponding result for group actions. The main difference is that the ring R has an underlying abelian group structure, the set M to be acted on is assumed to already be an abelian group, and the action is assumed to be compatible with these linear structures. Note that because an abelian group is automatically a $\mathbb{Z}$-module, it does make sense to consider the endomorphism ring $\mathrm{End}_{\mathbb{Z}}(M)$.

Proof. There are many details to check in this result, but they are all routine steps that follow from definitions. We will check that the function Φ makes sense, but leave the rest for the reader to verify.

Assume that M is an R-module, where the action of r on m is written as $r \cdot m$. Then as in the statement we define $\theta : R \to \mathrm{End}_{\mathbb{Z}}(M)$ where $[\theta(r)](m) = r \cdot m$.

First, why is does θ land in $\mathrm{End}_{\mathbb{Z}}(M)$? For this we need $\theta(r)$ to be a $\mathbb{Z}$-module homomorphism from M to itself, in other words a homomorphism of abelian groups. But $[\theta(r)](m_1 + m_2) = r \cdot (m_1 + m_2) = r \cdot m_1 + r \cdot m_2 = [\theta(r)](m_1) + [\theta(r)](m_2)$ by module axiom (iii), so this is fine.

Next, why is θ a unital homomorphism of rings? First, to see that θ respects addition, we want $\theta(r + s) = \theta(r) + \theta(s)$. We check this by applying to an arbitrary $m \in M$. So $[\theta(r + s)](m) = (r + s) \cdot m = r \cdot m + s \cdot m = [\theta(r)](m) + [\theta(s)](m) = [\theta(r) + \theta(s)](m)$, where we have used module axiom (iv) and the pointwise definition of addition in the endomorphism ring.

To see that θ respects multiplication, we want $\theta(rs) = \theta(r) \circ \theta(s)$, since the multiplication in $\mathrm{End}_{\mathbb{Z}}(M)$ is composition. Using module axiom (i), we check that $[\theta(rs)](m) = (rs) \cdot m = r \cdot (s \cdot m) = r \cdot [\theta(s)](m) = [\theta(r)]([\theta(s)](m)) = [\theta(r) \circ \theta(s)](m)$, as required.

Finally, to see that θ is unital, we want $\theta(1)$ to be the identity element of $\mathrm{End}_{\mathbb{Z}}(M)$, which is the identity function. We have $[\theta(1)](m) = 1 \cdot m = m$ for all $m \in M$, by module axiom (ii).

We have checked that Φ makes sense; i.e. given a module M there is a homomorphism of rings $\theta : R \to \mathrm{End}_{\mathbb{Z}}(M)$ defined by $[\theta(r)](m) = r \cdot m$. Notice that we used all of the module axioms (i)-(iv).

To show that Φ is a bijection, one may directly construct an inverse function Ψ. Given a homomorphism of rings $\theta : R \to \mathrm{End}_{\mathbb{Z}}(M)$, we let $\Psi(\theta)$ be the R-module struture on M, where $r \cdot m = [\theta(r)](m)$. We leave it to the reader to check the axioms of a module; the argument is basically already contained in the work above, which related each module axiom to some aspect of the homomorphism θ.

The fact that Ψ and Φ are inverse functions is then clear from their definitions. $\square$

Next, we show how one may describe the structure of a module over a polynomial ring over a field. Fix a field F, and let $R = F[x]$ be the ring of polynomials in one variable over F. Suppose that V is a left R-module. Since we can identify F with the subring of $F[x]$ given by constant polynomials, by restricting scalars the $F[x]$-module module V is also an F-module. Now define $\phi : V \to V$ by $\phi(v) = x \cdot v$, where $\cdot$ is the action of R on V. Clearly ϕ respects sums. Note that $ax = xa$ in $F[x]$ for all scalars $a \in F$. Thus $\phi(av) = x \cdot (a \cdot v) = xa \cdot v = ax \cdot v = a \cdot (x \cdot v) = a\phi(v)$. In other words, $\phi : V \to V$ is a linear transformation, or alternatively an element of $\mathrm{End}_F(V)$.

The argument above shows that an $F[x]$-module V leads to an F-vector space and a choice of linear transformation of V. In fact, conversely, a vector space together with a choice of linear transformation uniquely determines an $F[x]$-module. We formalize this as follows.

Proposition 12.27. *Let F be a field. An $F[x]$-module is the same thing as an F-vector space V together with a choice of F-linear transformation $\phi : V \to V$.*

Proof. Let V be an $F[x]$-module with action $\cdot$. We saw above that V is an F-vector space by restriction of scalars, and that $\phi : V \to V$ defined by $\phi(v) = x \cdot v$ is a linear transformation of V. Now by the module axioms, $x^2 \cdot v = x \cdot (x \cdot v) = \phi(\phi(v)) = \phi^2(v)$. By induction we get $x^n \cdot v = \phi^n(v)$ for all $n \geq 0$ (where we define $\phi^0 = 1_V$.) It follows that the action of $F[x]$ on V can be described by the formula

$$(\sum_{i \geq 0} a_i x^i) \cdot v = \sum_{i \geq 0} a_i \phi^i(v). \tag{12.28}$$

Conversely, suppose we are given a vector space V and a linear transformation $\phi : V \to V$. Then we define an action of $F[x]$ on V by (12.28). It is routine to check the module axioms, so that this does make V into an $F[x]$-module.

If we start with an $F[x]$-module V, it determines a vector space structure on V and a linear transformation ϕ. If we use this data to define an $F[x]$ action on V using (12.28), we have already seen that the original $F[x]$-module action must be given by this formula. Conversely, if we start with a V and a ϕ and use it to determine an $F[x]$ action on V via (12.28), clearly restricting the action to F gives the original V, and (12.28) gives $x \cdot v = \phi(v)$, so we recover V and ϕ. Thus we have proved there is a bijection between $F[x]$ modules and choices of (V, ϕ) where V is an F-vector space and $\phi : V \to V$ is linear. □

There is another point of view on Proposition 12.27 that uses Theorem 12.26, which we would like to describe. We first note the following result about homomorphisms from a polynomial ring. We leave the proof as an exercise.

Lemma 12.29. *Let $\phi : R \to T$ be a homomorphism of rings. Suppose that $t \in T$ commutes with every element of $\phi(R)$. Then there exists a unique homomorphism of rings $\widetilde{\phi} : R[x] \to T$ such that $\widetilde{\phi}(r) = \phi(r)$ for $r \in R$ and $\widetilde{\phi}(x) = t$. Conversely, given any homomorphism $\psi : R[x] \to T$, $\psi(x)$ commutes with every element of $\psi(R)$.*

The result shows that to define a homomorphism from $R[x]$ to another ring T, it is equivalent to define a homomorphism from R to T, and choose any element in T which commutes with the

image of R to send x to. This is a freeness property of the polynomial ring with respect to ring homomorphisms.

Now let us consider $F[x]$-modules again, where F is a field. We know that an $F[x]$-module action on an abelian group V can be described by a ring homomorphism $\widetilde{\theta} : F[x] \to \mathrm{End}_{\mathbb{Z}}(V)$, using Theorem 12.26. By Lemma 12.29, such a homomorphism is equivalent to a choice of ring homomorphism $\theta : F \to \mathrm{End}_{\mathbb{Z}}(V)$ and a choice of $\phi \in \mathrm{End}_{\mathbb{Z}}(V)$ which commutes with $\theta(F)$. By Theorem 12.26 again, θ corresponds to an F-module action, i.e. vector space structure, on V. The fact that ϕ commutes with $\theta(F)$ means that $[\phi \circ \theta(a)](v) = \phi(av)$ is the same as $[\theta(a) \circ \phi](v) = a\phi(v)$, in other words that ϕ respects the scalar multiplication. We see from this that an $F[x]$-module amounts to a choice of F-vector space V and an F-linear transformation $\phi : V \to V$, recovering Proposition 12.27.

Because an $F[x]$-module encodes a choice of linear transformation of a vector space, we are going to derive applications to linear algebra by proving later that modules over PIDs such as $F[x]$ have a tightly restricted struture.

Example 12.30. Let F be a field. Define an $F[x]$-module structure on the vector space $V = F^2$ of length-2 column vectors, where x acts by the linear transformation ϕ given by left multiplication by the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Explicitly, we have $x \cdot \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right] = \left[\begin{smallmatrix} a+b \\ b \end{smallmatrix}\right]$. What are the $F[x]$-submodules of V?

To answer this question, a little thought shows that an $F[x]$-submodule of V is a subset closed under the action of both F and x, in other words an F-subspace W such that $x \cdot W \subseteq W$, or $\phi(W) \subseteq W$. So $F[x]$-submodules are subspaces which are stable under the action of the linear transformation ϕ. In this case, it is not hard to work out that the only such subspaces are 0, V, and the 1-dimensional F-subspace $W = \{\left[\begin{smallmatrix} a \\ 0 \end{smallmatrix}\right] | a \in F\}$.

12.5. **Generation of modules and cyclic modules.**

Definition 12.31. Let M be an R-module. Given a subset $X \subseteq M$, the R-submodule *generated by* X is the unique smallest R-submodule of M containing X. Since the intersection of an arbitrary collection of submodules is again a submodule, it can be described as the intersection of all R-submodules of M containing X. We say that M is *finitely generated* if there is some finite subset of M which generates M; otherwise we say that M is *infinitely generated*. M is *cyclic* if it is generated by a subset with one element.

More explicitly, if we define $RX = \{r_1x_1 + \ldots r_nx_n | n \geq 0, r_i \in R, x_i \in X\}$ to be the set of all finite sums of elements of R acting on elements in X, then it is easy to see that RX is the

submodule of M generated by X. If $X = \{x_1, \ldots, x_n\}$ is finite, we also write this submodule as $Rx_1 + \cdots + Rx_n$. So M is cyclic if $M = Rx$ for some $x \in M$. We also call any submodule of M of the form Rx a cyclic submodule, as it is cyclic considered as a module in its own right.

Example 12.32. If M is a $\mathbb{Z}$-module, we have seen this is the just the canonical $\mathbb{Z}$-action on the abelian group M. Then $\mathbb{Z}$-submodules are the same as subgroups, so the submodule generated by a subset X is the same as the subgroup generated by the subset. Thus a $\mathbb{Z}$-module is cyclic if and only if it is cyclic as a group, i.e. either isomorphic to $\mathbb{Z}$ or to $\mathbb{Z}_n$ for some $n \geq 1$.

Example 12.33. Let R be a commutative ring, and let R be a left R-module by left multiplication. The R-submodule generated by $x \in R$ is Rx, in other words the principal ideal generated by x. If R is a PID, then we know that every ideal is principal, and so every submodule of R is a cyclic submodule of R. The ring R is a noetherian ring if and only if every ideal is finitely generated as an ideal, i.e. if and only if every submodule of R is finitely generated as an R-module.

Example 12.34. If $R = F$ is a field, then the submodule of an F-module (i.e. vector space) V generated by a subset X is just the span of X. So V is finitely generated as an F-module if and only if it is spanned by a finite subset, i.e. if and only if V is finite dimensional as a vector space. The cyclic submodules of V are the 1-dimensional subspaces (and the 0-subspace).

Example 12.35. Let R be an arbitrary ring, and let M be a cyclic left R-module, which is generated by $x \in M$. Then we can define a homomorphism $f : R \to M$ by $f(r) = rx$, where R is the usual left R-module structure on R. It is easy to see that f is an R-module homomorphism. The image of f is Rx, which is M by assumption, so f is surjective. Let $I = \ker f$. Then I is a left ideal of R, since kernels of homomorphisms are submodules. The 1st isomorphism theorem now tells us that $R/I \cong M$ as R-modules.

Conversely, for any left ideal I of R, we can form the factor module R/I, and this is a cyclic module generated by the element $(1 + I)$, since $r + I = r(1 + I)$ for all $r \in R$.

We see that the cyclic left R-modules are exactly the factor modules R/I for left ideals I, up to isomorphism.

Example 12.36. $(\mathbb{Q}, +)$ is an example of an infinitely generated $\mathbb{Z}$-module (i.e, abelian group). We will see in a few lectures that finitely generated abelian groups are easy to completely describe and classify. Infinite abelian groups are much more complicated, and there are still many open questions about their structure, many of which involve sensitive set-theoretic issues.

12.6. **Free modules.**

Definition 12.37. Given an indexed family $\{M_\alpha | \alpha \in I\}$ of left R-modules, the *direct product* is the cartesian product $\prod_{\alpha \in I} M_\alpha$, which is again an R-module under the coordinate-wise operations; in other words it is the direct product of abelian groups, with R-action $r \cdot (m_\alpha) = (r \cdot m_\alpha)$.

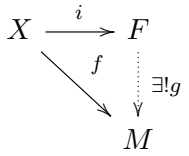
The *direct sum* of the family is the submodule of the direct product given by $\bigoplus_{\alpha \in I} M_\alpha = \{(m_\alpha) \in \prod M_\alpha | m_\alpha = 0 \text{ for all but finitely many } \alpha\}$. As an abelian group, we called this the *restricted product* of the groups M_α earlier, but the term direct sum is the standard one in the context of modules.

Note that when we have a finite family $M_1, M_2, \ldots, M_n$ of R-modules, the direct product and direct sum of this family are the same. We usually preference the direct sum notation and write this as $M_1 \oplus M_2 \cdots \oplus M_n$.

Free modules are the modules over a ring which are the most like vector spaces over a field. In general, a structure that is called "free" on a subset satisfies a certain universal property; we already saw the example of the free group when we studied group theory. For this reason we will take the universal property as our definition of a free module, and then show what they look like more explicitly.

Definition 12.38. Let F be a left R-module. Let X be a subset of F, and let $i : X \to F$ be the inclusion function. The module F is called *free* on a subset X if given any R-module M and a function $f : X \to M$, there is a unique R-module homomorphism $g : F \to M$ such that $g \circ i = f$. The *rank* of the free module F is the cardinality $|X|$ of the set X.

This universal property can be represented by the following commutative diagram:



In other words, given the inclusion function i and the homomorphism f, there exists a unique homomorphism g that completes the diagram to a commutative diagram (so $g \circ i = f$). Here the dashed arrow indicates that that homomorphism exists as a consequence of the property, and the ! indicates uniqueness.

The term "free" is used for properties like this because we may freely choose any elements of M whatsoever to send the elements in X to; then there is a unique homomorphism from F to M which does this to the elements in X.

Free objects are always determined uniquely up to isomorphism as a consequence of their universal properties, and the argument is always basically the same. Here is the result for reference, but it is really no different from our proof in group theory that the free group on a set is uniquely determined up to isomorphism by the cardinality of the set.

Proposition 12.39. *Let F be a free R-module on a subset X, and let G be a free R-module on a subset Y. If $|X| = |Y|$, then $F \cong G$ as R-modules.*

Proof. Let $i : X \to F$ and $j : Y \to G$ be the inclusion functions. Choose a set bijection $h : X \to Y$. Then the function $j \circ h : X \to G$ extends uniquely to a homomorphism $f : F \to G$, i.e. a homomorphism f such that $f|_X = h$. Similarly, $i \circ h^{-1} : Y \to F$ extends uniquely to a homomorphism $g : G \to F$. Now $g \circ f : F \to F$ is a homomorphism which restricts on X to the identity function; the identity homomorphism $1_F : F \to F$ is also such a homomorphism, so by the uniqueness property of the free module F, $g \circ f = 1_F$. Similarly, $f \circ g = 1_G$ since G is free. Thus f is an isomorphism with inverse g. □

Example 12.40. Let R be an R-module by left multiplication. Then R is free on the set $\{1\}$. To see this, let M be any R-module. Given $m \in M$, then $f : R \to M$ defined by $f(r) = rm$ is an R-module homomorphism such that $f(1) = m$. Moreover, it is the unique homomorphism sending 1 to m because $f(1) = m$ forces $f(r) = f(r1) = rf(1) = rm$.

We show now that it is easy to construct free modules explicitly, by extending Example 12.40 to a direct sum of copies of the module R.

Theorem 12.41. *Let I be any index set. Then $F = \bigoplus_{\alpha \in I} R$ is a free left R-module on the subset $X = \{e_\beta | \beta \in I\}$, where $e_\beta = (r_\alpha)_{\alpha \in I}$ with $r_\alpha = 0$ for $\alpha \neq \beta$ and $r_\alpha = 1$.*

You can think of the elements e_β as like "standard basis vectors" in a Euclidean space—they are 1 in exactly one coordinate and 0 elsewhere.

Proof. Let M be a module and let $f : X \to M$ be a function. We define $g : F \to M$ by $g((r_\alpha)) = \sum_\alpha r_\alpha f(e_\alpha)$. Note that this sum makes sense because $r_\alpha = 0$ for all but finitely many α. It is easy to see that g is an R-module homomorphism. Moreover, $g(e_\alpha) = f(e_\alpha)$ for all α, so g extends f. To see that g is unique, note that $(r_\alpha) = \sum_\alpha r_\alpha e_\alpha$, and so since g is an R-module homomorphism the formula $g((r_\alpha)) = \sum_\alpha r_\alpha f(e_\alpha)$ is forced. □

The theorem shows that for any set I, there is a free module on a subset X with cardinality $|I|$; and the free module is uniquely determined up to isomorphism by the cardinality of that set

by Proposition 12.39. So up to isomorphism, there is exactly one free module of any given rank, namely a direct sum of copies of the module R, over an index set of that rank.

There is another important way of thinking about free modules in terms of the concept of a basis.

Definition 12.42. Let F be an R-module. a subset X of F is called a *basis* of F if (i) X generates F as an R-module, so every $m \in F$ has an expression $m = r_1x_1 + \cdots + r_nx_n$ with $r_i \in R$ and $x_i \in X$; and (ii) whenever $r_1x_1 + \cdots + r_nx_n = 0$ for $r_i \in R$ and distinct $x_1, x_2, \ldots, x_n \in X$, then $r_i = 0$ for all i.

Example 12.43. Let V be a K-module, where K is a field, in other words a vector space over K. Then a subset X of V is a basis in the sense of the definition above if and only if X is a basis in the usual sense in linear algebra—(i) is the property that X spans V, and (ii) is the property that X is linearly independent.

We see that the idea of a basis of an R-module is modelled on the basis concept from linear algebra. We now show that it is precisely free modules that have a basis. So free modules are the objects in general module theory that behave most like vector spaces over a field.

Theorem 12.44. *An R-module F is free on a subset X if and only if X is a basis for F.*

Proof. Suppose that X is a basis for F. Property (i) of the definition of basis shows that an arbitrary $m \in F$ is an R-linear combination of elements in X. Thus m has an expression $m = \sum_{x\in X} r_x x$ with $r_x \in R$, where of course all but finitely many of the r_x are zero. But this expression is uniquely determined, for if also $m = \sum_{x\in X} r'_x x$, then $\sum_{x\in X}(r_x - r'_x)x = 0$, which shows that a finite R-linear combination of distinct elements in X is 0; by property (ii) of the definition of basis we get $r_x - r'_x = 0$ for all x, so $r_x = r'_x$ for all x.

We have seen that $\bigoplus_{x\in X} R$ is free on the set $\{e_x\}$ of standard basis vectos, where e_x is the element of the direct sum which is 1 in coordinate x and 0 elsewhere. By the universal property of this module there is a unique R-module homomorphism $g : \sum_{x\in X} R \to F$ such that $g(e_x) = x$. Explicitly, $g(\sum_{x\in X} r_x e_x) = \sum_{x\in X} r_x x$. The fact that every $m \in F$ has a unique expression of the latter form shows that g is bijective. Thus g is an isomorphism of modules. Since $\sum_{x\in X} R$ is free on the subset $\{e_x | x \in X\}$ by Theorem 12.41, F is free on the subset $\{g(e_x)|x \in X\} = X$.

Conversely, if F is free on a subset X, then by Proposition 12.39 there is an R-module isomorphism $F \to \sum_{x\in X} R$ which sends $x \in X$ to the standard basis vector e_x of the direct sum. Since this is an R-module isomorphism, it is easy to see that X is a basis of F if and only if $\{e_x|x \in X\}$ is

a basis of $\sum_{x \in X} R$. But it is trivial to see that the latter subset satisfies (i) and (ii) of the definition of a basis. □

It is well-known that every vector space has a basis, so one consequence of the preceding result is that all F-vector spaces are free as modules over F. This is just saying something that should already be familiar to you about vector spaces, which is that to define a linear transformation from one vector space V to another W, it suffices to choose arbitrary destinations in W for the elements in a basis of V.

The reader may well have never seen a completely general proof of the fact that every vector space has a basis, however. Since we have Zorn's Lemma in our toolbox, this is not difficult.

Theorem 12.45. *Let V be a vector space over a field K. Then V has a basis X. Moreover V is a free K-module on X.*

Proof. If $V = \{0\}$ is a vector space of dimension 0, then by convention we consider the empty set as a basis. So this case is fine, and from now on we assume that V is a nonzero vector space.

Let S be the collection of all K-linearly independent subsets of V. S is nonempty since any single nonzero vector is an independent set, and we are assuming that $V \neq 0$. Order S by inclusion. Consider a chain $\{X_\alpha\}$ of elements of S. Let X be the union of all of the sets in the chain. We claim that $X \in S$. To see this, take distinct $v_1, \ldots, v_n \in X$, and suppose that $a_1 v_1 + \cdots + a_n v_n = 0$. Now each v_i belongs to some set in the chain. Since it is a chain, there is a single α such that $v_1, \ldots, v_n \in X_\alpha$. By definition then the set $\{v_1, \ldots, v_n\}$ is linearly independent. This forces $a_i = 0$ for all i. Thus X is also linearly independent. So $X \in S$ as claimed, and clearly X is an upper bound for the chain.

Now by Zorn's Lemma, S has a maximal element X, which is by definition a linearly independent set. Suppose that X does not span V, and let W be the span of X. Then we can pick some $v \in V \setminus W$. Now consider $Y = X \cup \{v\}$. We claim that Y is again linearly independent. Suppose that $a_1 x_1 + \cdots + a_n x_n = 0$ where $x_i \in Y$ are all distinct. If $x_i \in X$ for all i, then $a_i = 0$ for all i since X is independent. Otherwise we can assume that $x_n = v$, with $a_n \neq 0$, and $x_i \in X$ for $i < n$. Then $v = -(a_n)^{-1}(a_1 x_1 + \cdots + a_{n-1} x_{n-1}) \in W$, a contradiction. so Y is independent as claimed, but this contradicts the maximality of X. Thus X must span V, and so X is a basis of V. □

The key step in the proof above is the ability to invert the coefficient $a_n \neq 0$, since K is a field. The same proof would show that for an arbitrary ring R, given a module M, there exists a subset (possibly empty) which is maximal among subsets which are R-linearly independent in the sense

of condition (ii) in the definition of basis. The submodule generated by this subset will be a free R-module, but there is no reason for it to equal M.

Free R-modules are certainly useful, but from the point of view of module theory perhaps not the most interesting. For each cardinality, there is a uniquely determined free module of that rank, which is a direct sum of copies of R over an index set of that cardinality. This is a very simple kind of classification result, since we know what they all look like up to isomorphism. We will work starting in the next section on the classification of finitely generated modules over PIDs, and that classification will be harder-earned and will have deeper consequences.

On the other hand, free modules over arbitrary rings R can behave in curious ways that defy the intuition we have from vector spaces. While there exists a unique free module up to isomorphism of each rank, there is no obvious reason that free modules of different ranks cannot be isomorphic. In fact, one can find a ring R such that $R \cong R \oplus R$ as left R-modules, and thus the free modules of rank 1 and rank 2 are isomorphic! For many rings R, however, it is true that two free modules are isomorphic if and only if they have the same rank. This is true for all commutative rings R, for example, which we leave as an exercise.

12.7. **Internal direct sums.** In our study of group theory, we gave conditions for a group G to be an internal direct product of a finite set of subgroups $H_1, \ldots, H_n$. This result, applied in the case of abelian groups, extends immediately to modules, as we see in the next theorem.

Definition 12.46. Let M be an R-module and let $\{N_\alpha | \alpha \in I\}$ be an arbitrary collection of submodules of M. The *sum* of these submodules, $N = \sum_{\alpha \in I} N_\alpha$, is the submodule of M generated by all of the elements in the submodules N_α. More explicitly, N consists of all elements of the form $\sum_{\alpha \in I} n_\alpha$ such that $n_\alpha \in N_\alpha$ and $n_\alpha = 0$ for all but finitely many α.

Note that a direct sum $\bigoplus_{\alpha \in I} N_\alpha$ is the sum of its submodules N_α (identifying N_α with its image under the αeth injection i_α). But in general the sum of some collection of submodules of a module is not direct. The case where this does happen is called an internal direct sum.

Because it is the main case we will be concerned with below, we state the theorem on internal direct sums for finite sums only, just as we did for groups. The general case is not really more difficult, it is just notationally more awkward.

Theorem 12.47. *Let M be an R-module. Suppose that M has R-submodules $N_1, \ldots, N_m$ with the properties that (i) $N_1 + N_2 + \cdots + N_m = M$ and (ii) $N_i \cap (N_1 + N_2 + \cdots + N_{i-1} + N_{i+1} + \cdots + N_m) = 0$ for all $0 \leq i \leq m$. Then $M \cong N_1 \oplus N_2 \oplus \cdots \oplus N_m$ as R-modules.*

Proof. Conditions (i) and (ii) are precisely the conditions for M to be an internal direct product as groups of the subgroups N_i (when written in additive form). Thus by our earlier study of such internal direct products, conditions (i) and (ii) force the natural map $\phi : N_1 \oplus N_2 \oplus \dots N_m \to M$ given by $\phi(n_1, n_2, \dots, n_m) = n_1 + n_2 + \dots + n_m$ to be an isomorphism of abelian groups. Now one just notices that ϕ also preserves the R-action and so is an isomorphism of R-modules. □

Here is an application of internal direct sums.

Definition 12.48. Let $f : M \to N$ be a homomorphism of left R-modules. Then f is called a *split surjection* if there is a homomorphism $g : N \to M$ such that $f \circ g = 1_N$.

Lemma 12.49. *Suppose that $f : M \to N$ is a split surjection, where $g : N \to M$ is a homomorphism with $f \circ g = 1_N$. Then f is surjective, g is injective, and $M \cong N \oplus K$ as R-modules, where $K = \ker(f)$.*

Proof. The fact that $f \circ g = 1_N$ immediately forces f to be surjective and g to be injective. We claim that M is the internal direct sum of its submodules $N' = g(N)$ and $K = \ker(f)$. Since g is injective, $N' \cong N$ as R-modules, so this will imply the result.

By Theorem 12.47 applied to two submodules, we just have to show that $N' + K = M$ and $N' \cap K = 0$. If $m \in N' \cap K$, then since $m \in N'$, $m = g(x)$ for some $x \in N$, so $f(m) = f(g(x)) = x$. But $m \in K$, so $x = f(m) = 0$. thus $m = g(x) = 0$. So $N' \cap K = 0$. For any $m \in M$, consider $y = m - g(f(m))$. Now $f(y) = f(m) - f(g(f(m))) = f(m) - f(m) = 0$ since $f \circ g = 1_N$. So $y \in \ker(f) = K$. But certainly $g(f(m)) \in g(N) = N'$. Thus $m = g(f(m)) + y \in N' + K$. So $N' + K = M$. □

There is a dual notion of split injection which also leads to a direct sum decomposition; we don't need it at the moment, so we postpone it until a later section where we examine results like this in the context of exact sequences.

One very useful consequence of this result is that surjections onto free modules are split.

Corollary 12.50. *Let $f : M \to F$ be a surjective homomorphism of R-modules, where F is a free R-module. Then f is a split surjection and hence $M \cong F \oplus K$ where $K = \ker(f)$.*

Proof. We just need to find $g : F \to M$ such that $f \circ g = 1_F$. Let F be free on the basis $\{x_\alpha | \alpha \in I\}$. Since f is surjective, for each α we can find an element $m_\alpha \in M$ such that $f(m_\alpha) = x_\alpha$. Now since F is free, there is a unique module homomorphism $g : F \to M$ such that $g(x_\alpha) = m_\alpha$. We have $f(g(x_\alpha)) = x_\alpha$ by definition, for all α. But since F is generated by the elements x_α, and $f \circ g$ is the identity on this subset, $f \circ g = 1_F$. □

13. Classification of modules over PIDs

13.1. **Torsion.** In this section, for simplicity we will only consider modules over commutative rings R.

We have seen that when K is a field, then K-modules are vector spaces V. If V is a finitely generated K-module, this is just a finite dimensional vector space. This is a free K-module, and is very easy to describe and understand using a basis.

After fields, the commutative rings which are simplest in some sense are the principal ideal domains (PIDs). The goal of this section is to show that we can completely understand finitely generated modules over a PID R.

Let us first study some definitions that are useful for modules over general integral domains.

Definition 13.1. Let R be an integral domain, and let M be an R-module. An element $m \in M$ is called *torsion* if there is $0 \neq r$ such that $rm = 0$. The subset $\mathrm{Tors}(M) = \{m \in M | m \text{ is torsion}\}$ is called the *torsion submodule* of M. The module M is called a *torsion module* if $M = \mathrm{Tors}(M)$, and M is *torsionfree* if $\mathrm{Tors}(M) = 0$.

Lemma 13.2. *Let R be an integral domain and let M be an R-module.*

(1) $\mathrm{Tors}(M)$ *is an R-submodule of M.*

(2) $M/\mathrm{Tors}(M)$ *is a torsionfree module.*

Proof. (1) If $rm = 0$ and $sm' = 0$ with $0 \neq r, 0 \neq s$ and $m, m' \in \mathrm{Tors}(M)$, then $rs(m - m') = s(rm) - r(sm') = 0$ and $rs \neq 0$ since R is a domain, so $m - m' \in \mathrm{Tors}(M)$. Also, for any $t \in R$, $r(tm) = t(rm) = 0$, so $tm \in \mathrm{Tors}(M)$.

(2) Let $N = \mathrm{Tors}(M)$. Suppose that $m + N \in M/N$ is a torsion element of M/N. Then there is $r \neq 0$ such that $r(m + N) = 0$. This means that $rm \in N$. Then rm is torsion, so there is $s \neq 0$ with $s(rm) = 0$. Since $sr \neq 0$, m is torsion and so $m \in N$. Thus $m + N = 0$ in M/N and so M/N is torsionfree. □

Notice that the proof above works only because R is a domain. One could define torsion elements and modules in the same way over an arbitrary commutative ring, but one typically does not because Lemma 13.2 is what makes these definitions useful.

If M is an R-module, where R is commutative, then for any $m \in M$ we define the *annihilator of m* to be $\mathrm{ann}_R(m) = \{r \in R | rm = 0\}$. It is easy to see that $\mathrm{ann}_R(m)$ is an ideal of R. If R is an integral domain, then clearly m is torsion if and only if $\mathrm{ann}_R(m) \neq 0$. We can also define

the *annihilator of* M to be $\operatorname{ann}_R(M) = \{r \in R | rm = 0 \text{ for all } m \in M\}$ which is also equal to $\bigcap_{m \in M} \operatorname{ann}_R(m)$.

Lemma 13.3. *Let M be a finitely generated module over an integral domain R. Then M is torsion if and only if* $\operatorname{ann}_R(M) \neq 0$.

Proof. Suppose that M is generated by $m_1, \ldots, m_n$, so $M = Rm_1 + \cdots + Rm_n$. Suppose that M is torsion, and choose $0 \neq r_i$ such that $r_i m_i = 0$. Then $(r_1 r_2 \ldots r_n)(s_1 m_1 + \cdots + s_n m_n) = 0$ for all $s_i \in R$ (since R is commutative). So $0 \neq r_1 r_2 \ldots r_n \in \operatorname{ann}_R(M)$. Conversely, if $I = \operatorname{ann}_R(M) \neq 0$, then for any $0 \neq r \in I$ we have $rm = 0$ for all $m \in M$, so M is torsion. □

Lemma 13.3 does not necessarily hold for infinitely generated modules. In this case such a module M can be torsion and yet have $\operatorname{ann}_R(M) = 0$.

Example 13.4. Let R be an integral domain. Any free R-module is torsionfree. A cylic R-module is of the form R/I for some ideal I; we have $\operatorname{ann}_R(R/I) = I$ and so R/I is torsionfree if $I = 0$ and otherwise is torsion.

Example 13.5. Let $R = \mathbb{Z}$. A $\mathbb{Z}$-module is torsion if and only if as a group of all of its elements have finite order. A $\mathbb{Z}$-module is torsionfree if and only if every nonzero element of the abelian group has infinite order. A finitely generated torsion $\mathbb{Z}$-module is just a finite abelian group.

13.2. **Classification of modules over PIDs.** Our main goal is prove the following classification theorem.

Theorem 13.6. *Let R be a PID. Let M be a finitely generated R-module. Then*

(1)
$$M \cong \overbrace{R \oplus R \oplus \cdots \oplus R}^{r} \oplus R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \cdots \oplus R/(p_m^{e_m})$$
as R-modules, where the p_i are (not necessarily distinct) primes and $e_i \geq 1$. The number r is called the rank *of M and the prime powers $p_1^{e_1}, \ldots, p_m^{e_m}$ are called the* elementary divisors *of M.*

(2) *The rank and elementary divisors are uniquely determined by M (up to reordering the elementary divisors or replacing the primes by associates). Two modules M and N are isomorphic if and only if they have the same rank and elementary divisors (up to order and associates).*

The theorem shows that every finitely generated module over a PID is a direct sum of finitely many cyclic modules, where the torsion cyclic modules appearing have annihilators which are ideals generated by prime powers. Moreover, this decomposition is unique. It is a very strong structure theorem. If we determine the rank and elementary divisors of a module we essentially know everything we need to know about it.

Let us make some more comments about the theorem before working towards the proof. First, the hypothesis that the module M is finitely generated is essential to Theorem 13.6.

Example 13.7. Consider $\mathbb{Q}$ as a $\mathbb{Z}$-module. It is easy to show that given nonzero $p, q \in \mathbb{Q}$, there are $a, b \in \mathbb{Z}$ such that $ap = bq$. It follows that any two nonzero subgroups of $\mathbb{Q}$ have nonzero intersection. Because of this $\mathbb{Q}$ cannot be an internal direct sum of two $\mathbb{Z}$-submodules. Moreover, $\mathbb{Q}$ is clearly not a cyclic $\mathbb{Z}$-module itself. Thus $\mathbb{Q}$ cannot be expressed as a direct sum of cyclic $\mathbb{Z}$-modules. In particular, $\mathbb{Q}$ is not a free $\mathbb{Z}$-module.

Second, the classification theorem should not be expected to hold for integral domains that are not PIDs.

Example 13.8. Let K be a field and let $R = K[x, y] = (K[x])[y]$ be polynomials in two variables over K. We have seen that R is a UFD but not a PID. Consider the ideal $I = Rx + Ry$ of R as a module over R by left multiplication. As a module, I cannot be written as an internal direct sum of two nonzero modules; if $I = J \oplus L$ for nonzero ideals J and L, then in particular $J \cap L = 0$, but if $0 \neq a \in J$ and $0 \neq b \in L$ then $0 \neq ab \in J \cap L$, a contradiction. Moreover, I is not a cyclic module itself, as it is not a principal ideal. (If $I = (z)$ then $z|x$ and $z|y$, but x and y are non-associate irreducibles in R so this forces z to be a unit, and $I = (z) = R$, but this is absurd.)

We see that I cannot be expressed as a direct sum of cyclic modules, even though I is finitely generated as a module. In particular, I is not a free module.

13.3. **The torsionfree case.** Now we start to work towards the proof of the theorem, which will be accomplished through a series of subsidiary results.

The first step is to handle the torsionfree case.

Proposition 13.9. *Let R be a PID. Let M be a torsionfree R-module which is finitely generated by n elements. Then M is free of finite rank $\leq n$.*

Proof. Suppose that M is generated by n elements as an R-module, say $M = Rm_1 + \ldots Rm_n$ for $m_i \in M$. The prove the result we induct on n. The base case is where M has 0 generators, in

which case $M = 0$ and the result is trivial. Now assume that $n \geq 1$ and that the result is true for fewer than n generators.

Consider the factor module M/Rm_1. Let $\text{Tors}(M/Rm_1)$ be its torsion submodule. By the correspondence theorem for modules, $\text{Tors}(M/Rm_1) = K/Rm_1$ where K is a submodule of M containing Rm_1. Explicitly, $K = \{m \in M | rm \in Rm_1 \text{ for some nonzero } r \in R\}$. Now by Lemma 13.2, $(M/Rm_1)/(K/Rm_1)$ is torsionfree, but also by the 4th isomorphism theorem for modules, this module is isomorphic to M/K.

Now M/K is torsionfree and since $m_1 \in K$, M/K is generated by the $n-1$ elements $m_2 + K, \ldots, m_n + K$. By the induction hypothesis, M/K is free of rank at most $n-1$. This implies that the natural surjection $\pi : M \to M/K$ is a split surjection, by Corollary 12.50. Then by Lemma 12.49, $M \cong M/K \oplus \ker(\pi) \cong M/K \bigoplus K$. To complete the proof, it now suffices to show that K is free of rank at most 1.

Since K is isomorphic to a summand of M, K is a surjective image of M and so K is also finitely generated. Then K/Rm_1 is a finitely generated torsion module, so by Lemma 13.3, it has nonzero annhilator. Say $0 \neq x \in \text{ann}_R(K/Rm_1)$. Then $xK \subseteq Rm_1$. Now $f : K \to xK$ given by $f(k) = xk$ is an isomorphism of modules, since M and hence K is torsionfree. Similarly, there is an isomorphism of modules $R \to Rm_1$ given by $r \mapsto rm_1$. Now we just need to show that every submodule of R is free of rank ≤ 1. But this is obvious, since a submodule is a principal ideal Ry, which is either 0 or free of rank 1. □

The hypothesis of finite generation is essential in Proposition 13.9. As we saw earlier in Example 13.7, $\mathbb{Q}$ is not a free $\mathbb{Z}$-module, but it is clearly a torsionfree $\mathbb{Z}$-module. The preceding result also certainly need not be true for integral domains that are not PIDs; Example 13.8 already gave the example of the finitely generated submodule $xR + yR$ of $R = K[x, y]$ which is not free.

Corollary 13.10. *Let R be a PID. If F is a free R-module of finite rank n, then every submodule of F is free of rank at most n.*

Proof. In particular, F is torsionfree and n-generated, so the result is imemdiate from Proposition 13.9. □

Unlike Proposition 13.9, it is possible to remove the finite rank assumption from Corollary 13.10; it is true that submodules of arbitrary free modules are free, for modules over a PID. We omit the proof of this result, which is not relevant for the classification of finitely generated modules over PIDs. On the other hand, the example $I = xR + yR \subseteq R = K[x, y]$ for a field K is a non-free

submodule of the rank one free module R itself, so again for non-PIDs we don't have a result like Corollary 13.10.

13.4. **The torsion case.** The remaining work is to analyze the torsion part in more detail. Recall that a PID is a UFD, and the prime and irreducible elements are the same. We will use the term prime below.

Definition 13.11. Let R be a PID and let $p \in R$ be prime. An R-module M is called *p-primary* if for all $m \in M$, there exists $n \geq 1$ such that $p^n m = 0$.

If M is a p-primary module, then for every $m \in M$, $(p^n) \subseteq \operatorname{ann}_R(m)$ for some n, and so $\operatorname{ann}_R(M) = (p^i)$ for some i since every ideal containing (p^n) is generated by a divisor of p^n and hence a power of p. Similarly, if M is a finitely generated p-primary module, then $\operatorname{ann}_R(M) = (p^i)$ for some i.

The first step in understanding finitely generated torsion modules over a PID is to show that such a module decomposes as a direct sum of p-primary submodules.

Definition 13.12. Let R be a PID and let M be an R-module. If p is a prime element of R, the *p-primary component* of M is $M_p = \{m \in M \mid p^n m = 0 \text{ for some } n \geq 1\}$.

It is easy to see that M_p is the unique largest p-primary R-submodule of M.

Proposition 13.13. *let M be a finitely generated torsion module over a PID R. Then there are pairwise non-associate primes $p_1, \ldots, p_k$ of R such that $M \cong M_{p_1} \oplus \cdots \oplus M_{p_k}$.*

Proof. Since M is finitely generated torsion, $\operatorname{ann}_R(M) \neq 0$, say $\operatorname{ann}_R(M) = (a)$ with $a \neq 0$. By unique factorization we can write $a = p_1^{e_1} p_2^{e_2} \ldots p_k^{e_k}$ for some pairwise non-associate primes p_i and integers $e_i \geq 1$. We claim that M is the internal direct sum of the submodules $M_{p_1}, \ldots M_{p_k}$, where M_{p_i} is the p_i-th primary component of M.

First, define $q_i = p_1^{e_1} \ldots p_{i-1}^{e_{i-1}} p_{i+1}^{e_{i+1}} \ldots p_k^{e_k}$, i.e. q_i is the prime factorization of a with the $p_i^{e_i}$ term removed. It is clear that $\gcd(q_1, \ldots, q_k) = 1$, since the only primes (up to associates) that divide any q_j are the primes p_i, but p_i does not divide q_i. Since R is a PID, this means that $1 = b_1 q_1 + \cdots + b_k q_k$ for some $b_i \in R$. Now if $m \in M$, then $m = 1m = b_1 q_1 m + \cdots + b_k q_k m$. By definition $p_i^{e_i} b_i q_i m = b_i a m = 0$. Thus $b_i q_i m \in M_{p_i}$ for all i. It follows that $M = M_{p_1} + \cdots + M_{p_k}$.

Next, suppose that $m \in M_{p_1} \cap (M_{p_2} + \cdots + M_{p_k})$, then p_1^s kills m for some s since $m \in M_{p_1}$, and $p_2^{n_1} \ldots p_k^{n_k}$ kills m for some $n_i \geq 1$, since $m \in M_{p_2} + \cdots + M_{p_k}$. But $\gcd(p_1^s, p_2^{n_1} \ldots p_k^{n_k}) = 1$. Since every R-linear combination of these elements will also kill m, we have $1m = 0$ and so $m = 0$. By

relabeling the primes, the same argument shows that $M_{p_i} \cap (M_{p_1} + \cdots + M_{p_{i-1}} + M_{p_{i+1}} + \cdots + M_{p_k}) = 0$ for all i. We have checked both conditions for an internal direct sum, so we see that M is an internal direct sum $M = M_{p_1} \oplus \cdots \oplus M_{p_k}$ as claimed. $\square$

The last and perhaps most sensitive step is to show that a p-primary module is a direct sum of cyclic modules. We first make an observation about modules that are killed by an actual prime (not just a prime power).

Example 13.14. Let R be a PID and let $p \in R$ be a prime element. Then we claim that an R-module M such that $(p) \subseteq \mathrm{ann}_R(M)$ is the same thing as a vector space over the field $K = R/(p)$.

In fact this is just a special case of a general phenomenon. If I is an ideal of a commutative ring R, and M is an R-module such that $IM = 0$, i.e. $I \subseteq \mathrm{ann}_R(M)$, then M is naturally an R/I-module defined by $(r + I) \cdot m = rm$; the fact that $IM = 0$ is used to show that this action is well-defined. Conversely, any R/I-module N is also an R-module, by pulling back along the ring homomorphism $\phi : R \to R/I$, in other words defining $r \cdot x = (r+I)x$, and the resulting R-module is certainly killed by I. It is easy to see that in this way R/I-modules are in bijective correspondence with R-modules that are annihilated by I.

Apply this to R and $R/(p)$, noting that $R/(p) = K$ is a field since in a PID a prime element generates a maximal ideal, and that a K-module is the same as a vector space over K.

Lemma 13.15. *Let M be a finitely generated p-primary module. Suppose that we have elements $0 \neq g_i \in M$ such that the sum $\sum_{i=1}^n Rg_i$ is the internal direct sum of its cyclic submodules $Rg_1, \ldots, Rg_n$. Assume that there are $h_1, \ldots, h_n \in M$ such that $g_i = ph_i$ for all i. Then $\sum_{i=1}^n Rh_i$ is also the internal direct sum of its cyclic submodules $Rh_1, \ldots, Rh_n$.*

Proof. The reader may easily check that condition (ii) of Theorem 12.47 for the submodules $Rh_1, \ldots, Rh_n$ is equivalent to the following statement: if $x_1 + \cdots + x_n = 0$ for $x_i \in Rh_i$, then $x_i = 0$ for all i.

Suppose that $r_1h_1 + \cdots + r_nh_n = 0$. Acting by p, we have $r_1g_1 + \cdots + r_ng_n = 0$. Since $\sum_{i=1}^n Rg_i$ is the internal direct sum of its submodules $Rg_1, \ldots, Rg_n$, we must have $r_ig_i = 0$ for all i. Since M is p-primary, $\mathrm{ann}_R(g_i) = p^{m_i}$ for some $m_i \geq 1$ (since $g_1 \neq 0$). So each r_i is a multiple of p. But then $r_ih_i \in Rg_i$ for all i. Again since $\sum_{i=1}^n Rg_i$ is the internal direct sum of its submodules $Rg_1, \ldots, Rg_n$, this forces $r_ih_i = 0$ for all i. $\square$

Proposition 13.16. *Let R be a PID with prime p, and let M be a finitely generated p-primary R-module. Then $M \cong R/(p^{s_1}) \oplus R/(p^{s_2}) \oplus \cdots \oplus R/(p^{s_k})$ as R-modules, for some list of positive integers $s_1, s_2, \ldots, s_k$.*

Proof. We have $\mathrm{ann}_R(M) = p^n$ for some $n \geq 0$. The proof is by induction on n. We take $n = 0$ as the base case; this is when $M = 0$ and the result holds trivially with the empty list of integers.

Now assume that $n \geq 1$ and that the result holds for all smaller n. Let $N = pM = \{pm | m \in M\}$. Then N is a submodule of M for which $p^{n-1}N = p^nM = 0$. By the induction hypothesis, we have an internal direct sum $N \cong \bigoplus_{i=1}^{l} C_i$ for some cyclic submodules C_i, where $C_i \cong R/(p^{j_1})$ for some $j_1, \ldots, j_l$ (it could be that $N = 0$ and the list is empty). Let g_i be a generator of C_i, so $\mathrm{ann}_R(g_i) = \mathrm{ann}_R(C_i) = (p^{j_i})$. Then $g_i = ph_i$ for some $h_i \in M$, since $N = pM$. It follows that $\mathrm{ann}_R(h_i) = (p^{j_i+1})$. By Lemma 13.15, the submodule $H = \sum_{i=1}^{l} Rh_i$ is also the direct sum of its cyclic submodules $Rh_1, \ldots, Rh_l$.

Now consider the submodule $M[p] = \{m \in M | pm = 0\}$ of M. As discussed in Example 13.14, this can be thought of as a vector space over $K = R/(p)$. In particular, $(M[p] \cap H)$ is a K-subpace of $M[p]$ and so we can choose a complement in $M[p]$, that is, a K-subspace V of $M[p]$ such that $M[p] = (M[p] \cap H) \oplus V$ as K-modules (it is a standard result for vector spaces that any subspace has a complement; or use that all K-modules are free and apply Corollary 12.50).

We now claim that we have an internal direct sum $M = H \oplus V$. First, by the choice of $V \subseteq M[p]$ we have $V \cap H \subseteq V \cap (M[p] \cap H) = 0$. Next, if $m \in M$ then $pm \in N$. By the definition of H, $pH = N = pM$, so there is $h \in H$ such that $ph = pm$ and hence $p(h - m) = 0$. Then $h - m \in M[p] = (M[p] \cap H) + V \subseteq H + V$. So $m \in H + V$. Hence $M = H \oplus V$ as claimed.

Finally, now V is a summand of M and so is also finitely generated as an R-module (and so also as a K-space). If $v_1, \ldots, v_s$ is a K-basis of V, then $V = Rv_1 \oplus \ldots Rv_s$ where $Rv_i \cong R/(p)$ as an R-module, for all i. Since $H = Rh_1 \oplus \cdots \oplus Rh_l$ where $Rh_i \cong R/(p^{j_i+1})$ for all i, we have $M \cong R/(p^{s_1}) \oplus \cdots \oplus R/(p^{s_k})$ for some positive integers $s_1, \ldots, s_k$. □

13.5. **Proof of the classification theorem.** We now put together all of the results we have proved to give the proof of the classification of finitely generated modules over PIDs.

Proof of Theorem 13.6. (1) Let M be a finitely generated module over the PID R. Let $T = \mathrm{Tors}(M)$. Of course T is torsion, and we know that M/T is torsionfree, by Lemma 13.2. Now M/T is finitely generated, since M is. It follows that M/T is free of finite rank, by Proposition 13.9. But then the quotient homomorphism $\pi : M \to M/T$ is a split surjection, and hence there is an internal direct sum decomposition $M \cong T \bigoplus F$, where $F \cong M/T$ is free of finite rank, say r. This

implies also that $M/F \cong T$, so T is also isomorphic to a factor module of M and thus T is also finitely generated. Now by Proposition 13.13, $T \cong T_{p_1} \oplus \cdots \oplus T_{p_k}$ for some pairwise nonassociate primes p_i, where T_{p_i} is p_i-primary and is finitely generated since it is a summand of T. Finally, each T_{p_i} is isomorphic to a direct sum $T_{p_i} \cong R/(p_i^{s_{i,1}}) \oplus \cdots \oplus R/(p_i^{s_{i,k_i}})$, by Proposition 13.16.

This proves that M is a direct sum of a rank r free module and a finite number of cyclic modules with annihilators generated by prime powers.

(2). Note that it is obvious that if two modules have the same rank and the same elementary divisors, then the modules are isomorphic. To prove the converse, we take two modules

$$M = \overbrace{R \oplus R \oplus \cdots \oplus R}^{r} \oplus R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \cdots \oplus R/(p_m^{e_m})$$

and

$$M' = \overbrace{R \oplus R \oplus \cdots \oplus R}^{s} \oplus R/(q_1^{f_1}) \oplus R/(q_2^{f_2}) \oplus \cdots \oplus R/(q_n^{f_n})$$

where p_i, q_i are primes and $e_i, f_i \geq 1$. It suffices to prove that if $M \cong M'$, then $r = s$, $m = n$, and after renumbering one of the sequences of prime powers we have $e_i = f_i$ and p_i and q_i are associates for all i.

Let $f : M \to M'$ be an isomorphism. It is clear that f restricts to an isomorphism of the torsion submodules $f : T = \text{Tors}(M) \to T' = \text{Tors}(M')$. Then f induces an isomorphism of the factor modules $F = M/T \to F' = M'/T'$. Thus M and M' have isomorphic free and torsion parts. It is clear that $T = R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \cdots \oplus R/(p_m^{e_m})$ and $T' = R/(q_1^{f_1}) \oplus R/(q_2^{f_2}) \oplus \cdots \oplus R/(q_n^{f_n})$. Thus $F \cong R^r$ and $F' \cong R^s$. A free module over a commutative ring has a uniquely determined rank by a homework exercise. Thus $r = s$ and the ranks are the same.

We have the isomorphism $f : T \to T'$ which we still call f. For any isomorphism of torsion modules and for any prime p, it restricts to an isomorphism $f : T_p \to T'_p$ between the p-primary components, by the definition of these components. Now notice that T_p is the direct sum of all of the summands $R/(p_i^{e_i})$ (if any) such that $(p_i) = (p)$, i.e. such that p_i is an associate of p. A similar comment holds for T'.

It now suffices to work one primary component at a time and show that if f is an isomorphism from $T_p = R/(p^{s_1}) \oplus \cdots \oplus R/(p^{s_k})$ to $T'_p = R/(p^{t_1}) \oplus \cdots \oplus R/(p^{t_l})$, then $k = l$ and after renumbering the t_i we have $s_i = t_i$ for all i. Equivalently, we just need that for each positive integer b, the number of s_i which is equal to b is the same as the number of t_i which is equal to b.

For each $b \geq 1$, if N is a p-primary module we can define $N[b] = \{x \in N | p^b x = 0\}$. By convention we put $N[0] = 0$. These are submodules of N with $0 = N[1] \subseteq N[1] \subseteq N[2] \subseteq \ldots$. Also, each factor module $N[b]/N[b-1]$ is killed by p and so is a vector space over $K = R/(p)$. In particular, a

short calculation shows that $T_p[b]/T_p[b-1]$ is a K-vector space of dimension equal to the number of s_j which are greater than or equal to b.

Now the isomorphism $f : T_p \to T'_p$ restricts to an isomorphism $T_p[b] \to T'_p[b]$ for all b, and hence induces an isomorphism $T_p[b]/T_p[b-1] \to T'_p[b]/T'_p[b-1]$ for all b as R-modules, hence also as K-vector spaces. It follows that the number of s_i which are greater than or equal to b is the same as the number of t_i which are greater than or equal to b, for all b. But this implies that the number of s_i which are equal to b is the same as the number of t_i which are equal to b. □

13.6. **The invariant factor form.** There is another form of the classification theorem in which the torsion part is written as a direct sum of cyclic modules in a different way. For completeness we restate the theorem in its entirety in this version.

Theorem 13.17. *Let R be a PID. Let M be a finitely generated R-module. Then*

(1)
$$M \cong \overbrace{R \oplus R \oplus \cdots \oplus R}^{r} \oplus R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_n)$$
as R-modules, where the $a_i \in R$ are nonzero, nonunit elements such that $a_i | a_{i+1}$ in R for all i. The number r is called the rank *of M and the elements $a_1, \ldots, a_n$ are called the* invariant factors *of M.*

(2) *The rank and invariant factors are uniquely determined by M (up to replacing the a_i by associates). Two modules M and N are isomorphic if and only if they have the same rank and invariant factors (up to associates).*

There are a few reasons why sometimes one might prefer the version of the classification in terms of invariant factors. In this version the torsion part is typically given as a direct sum of fewer cyclic modules. Also, the invariant factors occur in a specific order, unlike the ambiguity of the order in which the elementary divisors appear. We will see later how this leads to the uniqueness of the rational canonical form of a matrix, which has important applications.

In practice, if one is given the torsion part of a finitely generated module over a PID in elementary divisor form or in invariant factor form, it is routine to change to the other form. The reason is the following application of the Chinese remainder theorem.

Lemma 13.18. *Let R be a PID and let $a = p_1^{e_1} p_2^{e_2} \ldots p_n^{e_n}$, where the p_i are pairwise non-associate primes. Then*
$$R/(a) \cong R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \ldots R/(p_n^{e_n})$$
as both rings and as R-modules.

Proof. The fact that the primes are pairwise non-associate implies that $\gcd(p_i^{e_i}, p_j^{e_j}) = 1$ for $i \neq j$, and so the ideals $(p_i^{e_i})$ and $(p_j^{e_j})$ are comaximal. The Chinese remainder theorem now gives that the natural map $\phi : R/(a) \to R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \dots R/(p_n^{e_n})$ defined by the formula $\phi(r + (a)) = (r + (p_1^{e_1}), \dots, r + (p_n^{e_n}))$ is an isomorphism of rings. But it is clear that ϕ is also a homomorphism of R-modules. □

We give a proof of the invariant factor version of the fundamental theorem, but really seeing some examples in action may make this as clear as the rather technical proof. The reader might just skip this proof and move on to the examples we present afterwards.

Proof of Theorem 13.17. We only need to show that a finitely generated torsion module M over a PID can be expressed in invariant factor form and that the invariant factors are uniquely determined up to associates. The other parts of the theorem are the same as for Theorem 13.6.

We apply Theorem 13.6 to express M in terms of cyclic modules whose annihilators are the elementary divisors. Group together those elementary divisors which are associates of the same prime and change them if necessary so they are powers of exactly the same prime (which doesn't change the ideals they generate). We see that there are pairwise non-associate primes $p_1, \dots, p_m$ and exponents $e_{i,1}, \dots, e_{i,s_i}$ for each i such that M is a direct sum $M \cong \bigoplus_{i=1}^{m} \bigoplus_{j=1}^{s_i} R/(p_i^{e_{i,j}})$. We can order the exponents so that $e_{i,1} \geq e_{i,2} \geq \dots \geq e_{i,s_i}$. We also define $e_{i,j} = 0$ for $j > s_i$.

Now define $b_j = p_1^{e_{1,j}} \dots p_m^{e_{1,j}}$ for each $j \geq 1$, where as usual $p_i^0 = 1$ by convention. Then $b_i | b_{i-1}$ for all $i \geq 1$, with $b_j = 1$ for $j > n = \max\{s_i | 1 \leq i \leq m\}$. Define $a_i = b_{n+1-i}$ for $1 \leq i \leq n$. Then it is clear that $a_i | a_{i+1}$ for all $1 \leq i \leq n$. By Lemma 13.18,

$$\bigoplus_{j=1}^{n} R/(a_j) = \bigoplus_{j=1}^{n} R/(b_j) = \bigoplus_{j=1}^{n} \bigoplus_{i=1}^{m} R/(p_i^{e_{i,j}})$$

which is equal to the elementary divisor decomposition of M we started with if we ignore any zero summands. This proves that an invariant factor decomposition $M \cong R/(a_1) \oplus \dots \oplus R/(a_n)$ with $a_i | a_{i+1}$ for all i exists.

Conversely, suppose that $M \cong R/(c_1) \oplus \dots \oplus R/(c_t)$ is an invariant factor decomposition, with $c_i | c_{i+1}$ for all i. By using Lemma 13.18, we can break up each $R/(c_i)$ as a direct sum of cyclic modules with prime power annihilators. In this way we get an elementary factor decomposition. By the uniqueness of the elementary factor decomposition, the list of all of the prime powers occurring in the prime power decompositions of the c_i is the same as the list of all of the prime powers occuring in the prime power decompositions of the a_i. Now using the divisibility conditions, we see that the power of p_j occurring in c_i (possibly 0) is less than or equal to the power of p_j occurring

in c_{i+1} for all i. The same is true of the a_i. There is only one way to arrange a sequence of prime powers in nondecreasing order of exponents. It is straightforward to see now that $(c_i) = (a_i)$ for all i and $t = n$. □

13.7. **Examples.** Applying the fundamental classification theorem in the case $R = \mathbb{Z}$ immediately gives us a classification theorem for finitely generated abelian groups. We did not discuss this when we studied group theory, since it is more convenient to obtain it as a consequence of module theory. There is no proof for Abelian groups that doesn't have to do more or less the same steps as a proof for modules over general PID's.

We restate the fundamental theorem for the case $R = \mathbb{Z}$ for convenience.

Theorem 13.19. *Let G be a finitely generated abelian group. Then $G \cong \mathbb{Z}^r \oplus H$ for some uniquely determined free abelian group $\mathbb{Z}^r$ of rank r and finite abelian group H. The group H is isomorphic to $\bigoplus_{i=1}^m \mathbb{Z}/(p_i^{e_i})$ for some prime powers $p_i^{e_i}$ (elementary divisors) uniquely determined up to their order. H is also isomorphic to $\bigoplus_{i=1}^n \mathbb{Z}/(a_i)$ for some uniquely determined integers $a_i \geq 2$ (invariant factors) satisfying $a_i | a_{i+1}$ for all i.*

Example 13.20. Here are some explicit examples of going between invariant factor form and elementary divisor form for a finite abelian group.

Suppose that $G = \mathbb{Z}/(3) \oplus \mathbb{Z}/(12) \oplus \mathbb{Z}/(60) \oplus \mathbb{Z}/(360)$ is a group given in invariant factor form. To find the elementary divisor form, we simply factor each invariant factor into prime powers, and take the list of all of those prime powers. We have $3 = 3^1$, $12 = 2^2 3^1$, $60 = 2^2 3^1 5^1$, and $360 = 2^3 3^2 5^1$. Thus the elementary divisors are $2^2, 2^2, 2^3, 3, 3, 3, 3^2, 5, 5$ and

$$G \cong \mathbb{Z}/(2^2) \oplus \mathbb{Z}/(2^2) \oplus \mathbb{Z}/(2^3) \oplus \mathbb{Z}/(3) \oplus \mathbb{Z}/(3) \oplus \mathbb{Z}/(3) \oplus \mathbb{Z}/(3^2) \oplus \mathbb{Z}/(5) \oplus \mathbb{Z}/(5)$$

as $\mathbb{Z}$-modules and hence abelian groups. This is justified by Lemma 13.18.

For an example of the reverse process, consider the abelian group given in elementary divisor form by

$$G \cong \mathbb{Z}/(5) \oplus \mathbb{Z}/(5^2) \oplus \mathbb{Z}/(5^2) \oplus \mathbb{Z}/(7^2) \oplus \mathbb{Z}(7^3) \oplus \mathbb{Z}/(11).$$

The elementary divisors are $5, 5^2, 5^2, 7^2, 7^3, 11$.

To find the invariant factors, it is easiest to find them in reverse order, as in the proof of Theorem 13.17. Take the product of the largest powers of each prime among the elementary divisors, then the product of the largest powers of the primes among the remaining elementary divisors, etc. In this case we have $b_1 = (5^2)(7^3)(11)$, $b_2 = (5^2)(7^2)$, $b_3 = 5$. The invariant factors

are these integers in the reverse order: $a_1 = 5$, $a_2 = (5^2)(7^2) = 1715$, $a_3 = (5^2)(7^3)(11) = 94325$. So $G \cong \mathbb{Z}/(5) \oplus \mathbb{Z}/(1715) \oplus \mathbb{Z}/(94325)$.

In the next section we will explore the consequences of the fundamental theorem when applied to modules over a polynomial ring $K[x]$ for a field K. Here is an example of moving between invariant factors and elementary divisors in that context.

Example 13.21. Let $R = \mathbb{Q}[x]$. Consider the module $M = \mathbb{Q}[x]/(x^3 - 1) \oplus \mathbb{Q}[x]/(x^6 - 1)$, which is in invariant factor form, since $x^3 - 1 | x^6 - 1$ (as $x^6 - 1 = (x^3 - 1)(x^3 + 1)$).

To put this in elementary divisor form requires factorizing $x^3 - 1$ and $x^6 - 1$ as products of powers of prime (i.e. irreducible) polynomials in $\mathbb{Q}[x]$. We will discuss irreducibility for polynomials in more detail later (we ran out of time in the ring theory part last quarter, so we will do it in the field theory part this quarter). The only thing we use here is that a degree 2 polynomial over $\mathbb{Q}$ is irreducible if and only if it does not have a root in $\mathbb{Q}$.

By the standard formula for a difference of cubes, $x^3 - 1 = (x - 1)(x^2 + x + 1)$. It is easy to see that $x^2 + x + 1$ has no root in $\mathbb{Q}$, so it is irreducible over $\mathbb{Q}$. Similarly, $x^6 - 1 = (x^3 - 1)(x^3 + 1) = (x - 1)(x^2 + x + 1)(x + 1)(x^2 - x + 1)$, and $(x^2 - x + 1)$ is irreducible over $\mathbb{Q}$. We conclude that the elementary divisor form of M is

$$M \cong \mathbb{Q}[x]/(x-1) \oplus \mathbb{Q}[x]/(x-1) \oplus \mathbb{Q}[x]/(x+1) \oplus \mathbb{Q}[x]/(x^2+x+1) \oplus \mathbb{Q}[x]/(x^2+x+1) \oplus \mathbb{Q}[x]/(x^2-x+1).$$

It happens that all primes occur to the first power in this case.

14. Canonical Forms

In this section we will use the classification theorem of finitely generated modules over PIDs to develop the theory of canonical forms for linear transformations. These forms have many theoretical as well as practical uses in linear algebra.

14.1. **Linear algebra review.** Let V be a f.d. vector space over a field F. Suppose that $\phi : V \to V$ is an F-linear transformation. Fix an F-basis $\{v_1, \ldots, v_n\} = \mathcal{B}$ for V. Although we just use set bracket notation for the basis, we always assume that the order of the basis vectors is fixed as well. We can define the matrix of ϕ relative to $\mathcal{B}$ to be $M_{\mathcal{B}}^{\mathcal{B}}(\phi) = (a_{ij}) \in M_n(F)$ where $\phi(v_j) = \sum_i a_{ij} v_i$.

Then if we identify $v \in V$ with the column vector $v_{\mathcal{B}} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \in F^n$, where $v = \sum_j b_j v_j$, then

$$\phi(v) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

that is, ϕ is given by left multiplication by the matrix $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$. For each fixed basis $\mathcal{B}$, this association of matrices to linear transformations gives a ring isomorphism $\Psi : \mathrm{End}_F(V) \to M_n(F)$ defined by $\Psi(\phi) = M_{\mathcal{B}}^{\mathcal{B}}(\phi)$, such that $\phi(v)_{\mathcal{B}} = M_{\mathcal{B}}^{\mathcal{B}}(\phi) v_{\mathcal{B}}$ for all v.

Let us recall what happens to the matrix when we change the basis. If $\mathcal{B}' = \{w_1, \dots, w_n\}$ is also a basis, then we let $P = (p_{ij})$ be the change of basis matrix whose coordinates are defined by $w_j = \sum_i p_{ij} v_i$. Similarly, we can define $Q = (q_{ij})$ to be the change of basis matrix defined by $v_j = \sum_i q_{ij} w_i$. Then $v_j = \sum_i q_{ij} \sum_k p_{ki} v_k$ and so $\sum_k (\sum_i p_{ki} q_{ij}) v_k = v_j$ forces $\sum_i p_{ki} q_{ij} = \delta_{kj}$, where this is the Kronecker δ symbol defined by $\delta_{kj} = \begin{cases} 0 & k \neq j \\ 1 & k = j \end{cases}$. This implies that $PQ = I$ where I is the $n \times n$ identity matrix, so P is invertible with $Q = P^{-1}$.

Now we calculate

$$\phi(w_j) = \sum_i p_{ij} \phi(v_i) = \sum_k \sum_i p_{ij} a_{ki} v_k = \sum_k \sum_i \sum_l q_{lk} p_{ij} a_{ki} w_l$$

which implies that

$$M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)_{lj} = \sum_k \sum_i q_{lk} a_{ki} p_{ij} = [P^{-1} M_{\mathcal{B}}^{\mathcal{B}}(\phi) P]_{lj}.$$

Thus the matrix with respect to the new basis, $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi) = P^{-1} M_{\mathcal{B}}^{\mathcal{B}}(\phi) P$, is a conjugate of the matrix associated to the old basis.

Definition 14.1. Matrices $A, B \in M_n(F)$ are *similar* if there is an invertible matrix $P \in \mathrm{GL}_n(F)$ such that $P^{-1} A P = B$.

Similarity is obviously an equivalence relation on the set of all $n \times n$ matrices. The above calculations showed that if two matrices represent the same linear transformation with respect to two different bases, then the matrices are similar. It is easy to see that the converse also holds.

Given that similarity is an equivalence relation, we can consider equivalence classes of matrices in $M_n(F)$ with respect to this relation, which we call *similarity classes*. The idea of canonical forms is to choose a representative of each similarity class with a particularly nice form. Then in proofs

or calculations involving properties which are independent of similarity, one can reduce to the case of these canonical forms.

Definition 14.2. A matrix $A \in M_n(F)$ is *diagonal* if $A = (a_{ij})$ with $a_{ij} = 0$ for all $i \neq j$. Then $B \in M_n(F)$ is *diagonalizable* if B is similar to a diagonal matrix. A linear transformation $\phi \in \mathrm{End}_F(V)$ is called diagonalizable if there is a basis $\mathcal{B}$ of V such that $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is diagonal.

Diagonalizable matrices are usually the simplest ones to deal with when it comes to calculations. The first canonical form we study, the Jordan canonical form, will give a matrix in each similarity class which is as close to diagonal as possible in some sense. The Jordan form is also closely related to the theory of eignenvectors.

Definition 14.3. If V is a vector space over F and $\phi \in \mathrm{End}_F(V)$, then a nonzero vector $v \in V$ is an *eigenvector* of ϕ with *eigenvalue* λ if $\phi(v) = \lambda v$. Similarly, if $A \in M_n(F)$ and $0 \neq w \in F^n$ where the elements of F^n are written as column vectors, then w is an eigenvector of A with eigenvector λ if $Aw = \lambda w$.

It should be clear that $0 \neq v$ is an eigenvector of $\phi \in \mathrm{End}_F(V)$ if and only if $v_{\mathcal{B}}$ is an eigenvector of the matrix $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ for all choices of basis $\mathcal{B}$.

The following familiar result follows immediately from the definitions reviewed above.

Lemma 14.4. *A linear transformation $\phi \in \mathrm{End}_F(V)$ is diagonalizable if and only if V has a basis $\mathcal{B}$ consisting of eigenvectors of ϕ (in which case $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is diagonal). Similarly, a matrix $A \in M_n(F)$ is diagonalizable if and only if F^n has a basis of eigenvectors for A.*

Definition 14.5. Let $A \in M_n(F)$. The *characteristic polynomial* of A is $\mathrm{charpoly}(A) = \det(xI - A) \in F[x]$. If V is a vector space over F of dimension n and $\phi \in \mathrm{End}_F(V)$, then the *characteristic polynomial* of ϕ is $\mathrm{charpoly}(\phi) = \det(xI - A)$ where $A = M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ for any basis $\mathcal{B}$ of V.

Note that the choice of basis $\mathcal{B}$ in the definition above doesn't matter, because if matrices A and B are similar, they have the same characteristic polynomial, as

$$\det(xI - P^{-1}AP) = \det(P^{-1}(xI)P - P^{-1}AP) = \det(P^{-1}(xI - A)P)$$
$$= \det(P^{-1})\det(xI - A)\det(P) = \det(xI - A).$$

If v is an eigenvector of A, then $Av = \lambda v$ implies $(A - \lambda I)v = 0$ and so v is a nonzero vector in the nullspace of $(A - \lambda I)$. Thus $(A - \lambda I)$ is singular. Conversely if $(A - \lambda I)$ is singular, then a nonzero element in its nullspace will be an eigenvector of A with eigenvalue λ. It follows that the eigenvalues

of A are precisely the scalars λ such that $A-\lambda I$ is singular, or equivalently $\det(A-\lambda I)=0$. These are the roots of the characteristic polynomial $\det(A-xI)$. We have proved that the eigenvalues of A are the roots of the characteristic polynomial of A. Similarly, for an endomorphism ϕ of an n-dimensional vector space V, the eigenvalues of ϕ are the roots of the characteristic polynomial of ϕ.

In the next section we will want to focus on the case of matrices whose elementary divisors are powers of degree one primes. The next definitions are useful for this purpose.

Definition 14.6. Let $0 \neq f(x) \in F[x]$ for a field F with $d = \deg f$. We say that f *splits* over F if f factors as $f = c(x-r_1)(x-r_2)\dots(x-r_d)$ in $F[x]$ (i.e., with $c, r_1, \dots, r_d \in F$).

If a polynomial f of degree d splits over F, then it has d roots $r_1, \dots, r_d$ in F (counted with multiplicity).

Definition 14.7. A field F is *algebraically closed* if every nonzero polynomial $f \in F[x]$ splits over F.

The *Fundamental Theorem of Algebra* is the statement that the field $\mathbb{C}$ of complex numbers is algebraically closed. We will give a proof of this at the end of our study of field theory, but we simply assume it now for convenience. We will also prove later that any field is contained in an algebraically closed one.

Note that if a field F is algebraically closed, then every nonzero polynomial in $F[x]$ factors as a product of degree 1 factors in $F[x]$ and a unit. This implies that the irreducible polynomials in $F[x]$ are precisely the polynomials of degree 1. Recall that a polynomial is *monic* if its leading coefficient is 1. The monic irreducibles in $F[x]$ are just the polynomals $(x-r)$ with $r \in F$.

14.2. **Jordan canonical form.** Let F be a field. Consider a finite dimensional vector space V over F, and an F-linear transformation $\phi \in \mathrm{End}_F(V)$. Recall that given any vector space with a choice of linear endomorphism, we can encode this information by making V into a module over the ring $F[x]$, where the constant polynomials in F act by the existing scalar multiplication and x acts by $x \cdot v = \phi(v)$. The action by a general element of $F[x]$ then follows the rule $(\sum_{i=0}^n a_i x^i) \cdot v = \sum_{i=0}^n a_i \phi^n(v)$, where we take $\phi^0 = 1_V$.

Now since V is finite-dimensional over F, it is finitely generated over F (by a basis) and so it is certainly finitely generated as a module over the larger ring $F[x]$. Since $F[x]$ is a PID, we can apply the classification of finitely generated modules over a PID to the module V. Here we apply the

elementary divisor version; in the next section we show how to get somewhat different information by applying the invariant factor version.

Note that a nonzero free $F[x]$-module has infinite dimension as an F-vector space. Thus applying the classification to our module V, we see that the free part of V is zero and V is a torsion $F[x]$-module. The theorem tells us that there is an $F[x]$-module isomorphism

$$V \cong F[x]/(f_1^{e_1}) \oplus F[x]/(f_2^{e_2}) \oplus \cdots \oplus F[x]/(f_s^{e_s})$$

where the $f_1, \ldots, f_s$ are prime, i.e. irreducible, polynomials in $F[x]$ and the $e_i \geq 1$. The prime powers $f_i^{e_i}$ are unique up to the order in which they appear, and possibly replacing f_i with associates. In this case by multiplying each by a nonzero scalar we can insist that the f_i be monic and then there is no ambiguity up to associates.

The Jordan canonical form we want to develop exists only under an additional condition: we assume now that the all of the irreducible polynomials f_i appearing in the elementary divisors have degree 1. By the comments in the previous section, this is always the case if we assume that F is algebraically closed. When we study fields we will see that every field is contained in an algebraically closed one, so this is not a huge restriction.

With our new assumption, we have that we can write $f_i = (x - \lambda_i)$ for some $\lambda_i \in F$. So as $F[x]$-modules we have an isomorphism

$$V \cong F[x]/((x-\lambda_1)^{e_1}) \oplus F[x]/((x-\lambda_2)^{e_2}) \oplus \cdots \oplus F[x]/((x-\lambda_s)^{e_s}).$$

The λ_i are not necessarily distinct.

For the moment, consider the case where $s = 1$, that is where V has only one elementary divisor. For convenience, drop the indexing and write $V \cong F[x]/(x-\lambda)^e$ as $F[x]$-modules. Now we choose a F-basis of $F[x]/(x-\lambda)^e$ for which the multiplication by x map will have a simple form. Let $I = ((x-\lambda)^e)$ be the ideal of $R = F[x]$ generated by $(x-\lambda)^e$, so $V \cong R/I$ as R-modules.

Now notice that $\{w_1 = (x-\lambda)^{e-1} + I, w_2 = (x-\lambda)^{e-2} + I, \ldots, w_{e-1} = (x-\lambda) + I, w_e = 1 + I\}$ is an F-basis of R/I. This follows just since I is generated by a polynomial of degree e, and $(x-\lambda)^i$ has degree i, so we have coset representatives of degrees $1, 2, \ldots, e-1$.

In the $R = F[x]$-module action on R/I, we have

$$(x-\lambda) \cdot w_i = (x-\lambda)[(x-\lambda)^{e-i} + I] = (x-\lambda)^{e-i+1} + I = \begin{cases} w_{i-1} & 2 \leq i \leq e \\ 0 & i = 1. \end{cases}$$

We can rewrite $(x-\lambda)\cdot w_i = w_{i-1}$ as $x\cdot w_i = \lambda w_i + w_{i-1}$ for $2 \le i \le e$, while $(x-\lambda)\cdot w_1 = 0$ becomes $x\cdot w_1 = \lambda w_1$. In other words, w_1 is an eigenvector for the linear transformation of $F[x]/I$ given by action by x.

Now let $\theta : V \to F[x]/I$ be a $F[x]$-module isomorphism. Then define $v_i = \theta^{-1}(w_i)$ for all i. Since θ is an $F[x]$-module isomorphism, it is also an F-vector space isomorphism. Thus $\mathcal{B} = \{v_1, \ldots, v_e\}$ is an F-basis of V. Moreover, we have $\theta(x\cdot v_i) = x\cdot\theta(v_i) = x\cdot w_i$ for all i. But by definition $x\cdot v_i = \phi(v_i)$ for all i. Given the rules above for how x acts on the basis $\{w_i\}$ of R/I, we have

$$\phi(v_i) = \begin{cases} \lambda v_i + v_{i-1} & 2 \le i \le e \\ \lambda v_1 & i = 1. \end{cases}$$

This shows that the matrix of ϕ with respect to the basis $\mathcal{B}$ has an especially simple form:

$$M_{\mathcal{B}}^{\mathcal{B}}(\phi) = \begin{pmatrix} \lambda & 1 & & & & \mathbf{0} \\ & \lambda & 1 & & & \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ & & & & \lambda & 1 \\ \mathbf{0} & & & & & \lambda \end{pmatrix}$$

More precisely,

$$[M_{\mathcal{B}}^{\mathcal{B}}(\phi)]_{ij} = \begin{cases} \lambda & i = j \\ 1 & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

This matrix is called the *Jordan block* of size e associated to the eigenvalue λ. Its only nonzero entries are along the main diagonal (all λs) and the diagonal just above it (all 1's). We also write this $e \times e$ matrix as $J_{\lambda,e}$.

Now we pass to the general case where there is more than one elementary divisor and an isomorphism

$$\theta : V \cong F[x]/((x-\lambda_1)^{e_1}) \oplus F[x]/((x-\lambda_2)^{e_2}) \oplus \cdots \oplus F[x]/((x-\lambda_s)^{e_s}).$$

We can choose a special basis $\{w_{i,1}, \ldots, w_{i,e_i}\}$ of each summand $F[x]/((x-\lambda_i)^{e_i})$ of the right hand side, as above, where $w_{i,j} = (x-\lambda_i)^{e_i-j} + ((x-\lambda_i)^{e_i})$. Then stringing these together gives as a basis $\{w_{1,1}, \ldots, w_{1,e_1}, w_{2,1}, \ldots, w_{2,e_2}, \ldots, w_{s,1}, \ldots w_{s,e_s}\}$ of the right hand side. Applying θ^{-1} gives

us a basis $\mathcal{B}$ of V. The matrix of ϕ with respect to $\mathcal{B}$ is then a block matrix

$$M_{\mathcal{B}}^{\mathcal{B}}(\phi) = \left(\begin{array}{c|c|c|c} J_{\lambda_1,e_1} & & & \mathbf{0} \\ \hline & J_{\lambda_2,e_2} & & \\ \hline & & \ddots & \\ & & & \ddots \\ \hline \mathbf{0} & & & J_{\lambda_s,e_s} \end{array}\right)$$

Here, this is a matrix of size $n \times n$ where $n = e_1 + \cdots + e_s$. The diagonal blocks are Jordan blocks, and all other blocks are 0. A matrix of this type is said to be in *Jordan canonical form.*

Theorem 14.8. *Let $\phi : V \to V$ be a linear transformation of the n-dimensional vector space V over F. Make V into an $F[x]$-module via ϕ, and assume the that elementary divisors of the $F[x]$-module V are of the form $(x-\lambda_1)^{e_1}, \ldots, (x-\lambda_s)^{e_s}$. Then there is a basis $\mathcal{B}$ of V s.t. $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is in Jordan canonical form, with Jordan blocks $J_{\lambda_1,e_1}, \ldots, J_{\lambda_s,e_s}$.*

If there is another basis $\mathcal{B}'$ such that $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)$ is also in Jordan canonical form, then this matrix is the same as $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ after possibly reordering the Jordan blocks.

Proof. The existence of the basis $\mathcal{B}$ was proved in the preceding discussion.

Conversely, suppose that $\mathcal{B}'$ is another basis for which $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)$ is in Jordan form. The fact that this is a block matrix with only the blocks along the main diagonal nonzero, say with blocks of size $f_1, f_2, \ldots, f_t$, means that $V \cong V_1 \oplus V_2 \oplus \cdots \oplus V_t$ with $\phi(V_i) \subseteq V_i$, where V_i is spanned by f_i elements of the basis, and with $f_1 + \cdots + f_t = n$. Since the matrix of $\phi|_{V_i}$ with respect to the corresponding f_i basis elements is a Jordan block J_{μ_i,f_i}, as an $F[x]$-module x acts on the basis $v_1, \ldots, v_{f_i}$ of V_i via the rules

$$\phi(v_i) = \begin{cases} \mu_i v_i + v_{i-1} & 2 \leq i \leq f_i \\ \mu_i v_1 & i = 1. \end{cases}$$

This easily implies that $V_i \cong F[x]/(x-\mu_i)^{f_i}$ as an $F[x]$-module, by reversing the steps in the earlier argument. We conclude that

$$V \cong F[x]/((x-\mu_1)^{f_1}) \oplus F[x]/((x-\mu_2)^{f_2}) \oplus \cdots \oplus F[x]/((x-\mu_t)^{f_t}).$$

as $F[x]$-modules. But now $(x-\mu_1)^{f_1}, \ldots, (x-\mu_t)^{f_t}$ are elementary divisors for the module V. By the uniqueness of elementary divisors, $s = t$, and possibly after renumbering, we have $\mu_i = \lambda_i$ and $e_i = f_i$ for all i. In other words, the Jordan form $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)$ is the same as $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ after reordering the Jordan blocks. $\square$

Corollary 14.9. *Let F be an algebraically closed field. Then every matrix $A \in M_n(F)$ is similar to a matrix in Jordan canonical form. The Jordan form is uniquely determined up to rearrangement of the Jordan blocks.*

One reason that the Jordan form is useful is that calculation of the powers of a matrix in this form is especially simple.

Example 14.10. Let $J = J_{\lambda,3}$ be a Jordan block of size 3. Then for all $n \geq 1$ we have

$$J^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1} & \binom{n}{2}\lambda^{n-2} \\ 0 & \lambda^n & n\lambda^{n-1} \\ 0 & 0 & \lambda^n \end{pmatrix},$$

by an easy inductive proof.

Similarly, a Jordan block of any size has an explicit formula for its powers involving binomial coefficients. Then the powers of any Jordan form may also be explicitly determined, simply by taking powers of the blocks. Finally, if A is an arbitrary matrix which is similar to a Jordan form J, if we calculate explicitly the matrix P such that $A = P^{-1}JP$, then the powers of A may be explicitly determined as $A^n = P^{-1}J^nP$.

If we are trying to understand all elements of $M_n(F)$ with a certain property that is invariant under similarity, then if F is algebraically closed we can reduce to the case of a Jordan form, where the calculation is usually much easier.

Example 14.11. Suppose we would like to find all elements of $\mathrm{GL}_2(\mathbb{C})$ which have order dividing 3 in this group. In other words we want all matrices A such that $A^3 = I$.

Since we are working over the algebraically closed field $\mathbb{C}$, all matrices have a Jordan form. Note that $A^3 = I$ if and only if $B^3 = I$, for any matrix B similar to A. Thus if we find all nonsingular matrices A in Jordan form which have $A^3 = I$, then the answer will simply be the union of all similarity classes containing those Jordan forms.

Jordan forms of 2×2 matrices are either diagonal, say $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ for $\lambda_1, \lambda_2 \in \mathbb{C}$, or else a Jordan block of size 2, $\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$. The formula for powers of a 2×2 Jordan block, which is even easier than Example 14.10, is

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1} \\ 0 & \lambda^n \end{pmatrix}.$$

In particular no positive power of such a Jordan block can be the identity matrix I. We conclude that the Jordan form of an invertible matrix of order dividing 3 is a diagonal matrix. Since $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^3 = \begin{pmatrix} \lambda_1^3 & 0 \\ 0 & \lambda_2^3 \end{pmatrix}$, The Jordan forms with multiplicative order dividing 3 are

$$S = \left\{ \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \middle| \lambda_1^3 = \lambda_2^3 = 1 \right\}.$$

$\mathbb{C}$ has three cube roots of 1, $\{1, e^{2\pi i/3}, e^{4\pi i/3}\}$. There are thus 9 distinct diagonal matrices in S, and the set of matrices with multiplicative order dividing 3 is equal to the set of all conjugates of S, i.e. the union of the similarity classes containing elements of S.

Suppose we want to know how many distinct similarity classes there are of such matrices. Then we need to determine whether any of the elements in S are similar. We know the Jordan form is determined only up to rearranging the Jordan blocks; thus two diagonal matrices with the same diagonal elements in some order are similar. It is now easy to see that there are only 6 distinct similarity classes containing the invertible matrices A such that $A^3 = I$.

14.3. **Rational canonical form.** The rational canonical form is developed in a very similar way to the Jordan canonical form, except using the invariant factor form instead of the elementary divisor form of the fundamental theorem.

Again let F be any field, let V be a finite dimensional F-space, and choose an F-linear transformation $\phi \in \mathrm{End}_F(V)$. Make V into an $F[x]$-module where x acts by ϕ. Then V is a torsion $F[x]$-module, as we have already argued earlier. Theorem 13.17 tells us that there is an $F[x]$-module isomorphism

$$V \cong F[x]/(f_1) \oplus F[x]/(f_2) \oplus \cdots \oplus F[x]/(f_m)$$

where the invariant factors $f_1, \ldots, f_m \in F[x]$ satisfy $f_i | f_{i+1}$ for all $1 \leq i \leq m-1$. The invariant factors are uniquely determined up to associates. We will insist that the f_i are monic polynomials, and then they are completely unique.

Consider first the case where $n = 1$, so there is one invariant factor. Then $V \cong F[x]/(f)$ for some monic polyomial $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 \in F[x]$, some $n \geq 1$. In other words, V is essentially an arbitrary nonzero torsion cyclic $F[x]$-module.

Let $I = (f)$ for convenience of notation. Now $\{w_1 = 1 + I, w_2 = x + I, \ldots, w_n = x^{n-1} + I\}$ is an F-basis of $F[x]/I$. In the $F[x]$-module action on R/I, we have

$$x \cdot w_i = x(x^{i-1} + I) = x^i + I = \begin{cases} w_{i+1} & 1 \leq i \leq n-1 \\ -b_{n-1}w_n - b_{n-2}w_{n-1} - \cdots - b_1 w_2 - b_0 w_1 & i = n; \end{cases}$$

the second case follows since $f \in I$ and so

$$x^n + I = -(b_{n-1}x^{n-1} + \cdots + b_1 x + b_0) + I = -b_{n-1}(x^{n-1} + I) - \cdots - b_1(x + I) - b_0(1 + I).$$

Now if we fix an $F[x]$-module isomorphism $\theta : V \to F[x]/I$ and define $v_i = \theta^{-1}(w_i)$, then $\mathcal{B} = \{v_1, \ldots, v_n\}$ is an F-basis of V for which

$$\phi(v_i) = \begin{cases} v_{i+1} & 1 \leq i \leq n-1 \\ -b_{n-1}v_n - b_{n-2}v_{n-1} - \cdots - b_1 v_2 - b_0 v_1 & i = n. \end{cases}$$

We conclude that

$$M_{\mathcal{B}}^{\mathcal{B}}(\phi) = \begin{pmatrix} 0 & 0 & & & & -b_0 \\ 1 & 0 & & & & -b_1 \\ 0 & 1 & & & & -b_2 \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & 1 & 0 & -b_{n-2} \\ & & & & & 1 & -b_{n-1} \end{pmatrix}.$$

More precisely,

$$[M_{\mathcal{B}}^{\mathcal{B}}(\phi)]_{ij} = \begin{cases} 1 & i = j + 1 \\ -b_{i-1} & j = n \\ 0 & \text{otherwise.} \end{cases}$$

This matrix is called the *companion matrix* of the monic polynomial $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$. We also write it as C_f.

In the general case where we have more than one invariant factor, we proceed exactly as we did with the Jordan form. Fixing an isomorphism

$$\theta : V \cong F[x]/(f_1) \oplus F[x]/(f_2) \oplus \cdots \oplus F[x]/(f_m)$$

we choose a basis $\mathcal{B}$ of V which is the preimage under θ of the basis of the right hand side obtained by stringing together the special bases of the modules $F[x]/(f_i)$ we picked above. The matrix of ϕ

with respect to $\mathcal{B}$ is then a block matrix

$$M_{\mathcal{B}}^{\mathcal{B}}(\phi) = \left(\begin{array}{c|c|cc|c} C_{f_1} & & \cdots & \cdots & \mathbf{0} \\ \hline & C_{f_2} & & & \\ \hline \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \hline \mathbf{0} & & & & C_{f_m} \end{array}\right)$$

A matrix of this type is said to be in *rational canonical form*. Note that there is no ambiguity in the order of the blocks of a matrix in rational canonical form. The blocks must be the companion matrices of polynomials each of which divides the next as we go down the diagonal from the top left to the bottom right.

Theorem 14.12. *Let $\phi : V \to V$ be a linear transformation of the n-dimensional vector space V over F. Make V into an $F[x]$-module via ϕ, and assume that the invariant factors of the $F[x]$-module V are $f_1, \ldots, f_m$. Then there is a basis $\mathcal{B}$ of V such that $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is in rational canonical form, with blocks $C_{f_1}, \ldots, C_{f_m}$.*

If $\mathcal{B}'$ is a basis such that $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)$ is also in rational canonical form, then $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi) = M_{\mathcal{B}}^{\mathcal{B}}(\phi)$.

Proof. The proof is completely analogous to the proof of Theorem 14.8, but using the uniqueness of the invariant factor decomposition of a torsion module instead. We leave the details to the reader. □

Corollary 14.13. *Let F be an arbitrary field. Then every matrix $A \in M_n(F)$ is similar to a unique matrix in $M_n(F)$ which is in rational canonical form.*

We have emphasized the Jordan canonical form since this is the form which is most useful in calculations and in applications. There is no nice formula for the powers of a companion matrix, in contrast. The rational canonical form is useful for theoretical reasons, however, because it is defined over an arbitrary field, and it is absolutely unique. The Jordan canonical form, by contrast, only exists over certain fields, and is unique only up to permutation of the blocks.

Here is a useful theorem whose proof becomes nearly trivial with the use of the rational canonical form. Recall that $F \subseteq K$ is called a *field extension* if F and K are fields, and F is a subring of K.

Theorem 14.14. *Let $F \subseteq K$ be a field extension. Let $A, B \in M_n(F)$. Then we can also consider $A, B \in M_n(K)$. If A and B are similar in the ring $M_n(K)$ (i.e. $A = P^{-1}BP$ for some $P \in \mathrm{GL}_n(K)$) then A and B are similar in $M_n(F)$.*

Proof. Let C be the rational canonical form of A over F, and let C' be the rational canonical form of B over F. Thus $C, C' \in M_n(F)$.

Now A and C are similar in $M_n(F)$, so A and C are certainly similar in $M_n(K)$ as well. But the matrix C is in rational canonical form, that is it is block diagonal with companion matrices C_{f_i} along the diagonal, for $f_i \in F[x]$ with $f_i | f_{i+1}$ for all i. Clearly then C is also in rational canonical form when considered as a matrix in $M_n(K)$. By the uniqueness of the rational canonical form, C is the rational canonical form of A in $M_n(K)$. We have in the same way that C' is the rational canonical form of B in $M_n(K)$. But by assumption A and B are similar in $M_n(K)$. Thus $C = C'$ by the uniqueness of the rational canonical form in $M_n(K)$. But then A is similar to B in $M_n(F)$ by the uniqueness of the rational canonical form in $M_n(F)$. □

The preceding result is highly non-obvious without introducing forms. If $A = P^{-1}BP$ for some $P \in \mathrm{GL}_n(K)$, there is no obvious way to adjust P to obtain a $Q \in \mathrm{GL}_n(F)$ such that $A = Q^{-1}BQ$ also.

14.4. **Characteristic and minimal polynomials.** We now relate canonical forms of a matrix to its characteristic and minimal polynomials (we will define the minimal polynomial shortly). Because the rational canonical form is defined over any field we use this form in our initial approach.

Lemma 14.15. *Let $C_f \in M_n(F)$ be a companion matrix for a monic polynomial $f \in F[x]$. Then* $\mathrm{charpoly}(C_f) = f[x]$.

Proof. Let $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$. We have

$$\mathrm{charpoly}(C_f) = \det \begin{pmatrix} x & 0 & & & & b_0 \\ -1 & x & & & & b_1 \\ 0 & -1 & x & & & b_2 \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & -1 & x & b_{n-2} \\ & & & & & -1 & x + b_{n-1} \end{pmatrix}.$$

Expanding by minors along the first row gives

$$x \det \begin{pmatrix} x & 0 & & & & b_1 \\ -1 & x & & & & b_2 \\ 0 & -1 & x & & & b_3 \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & -1 & x & b_{n-2} \\ & & & & -1 & x + b_{n-1} \end{pmatrix} + (-1)^{n-1} b_0 \det \begin{pmatrix} -1 & x & & & & \\ 0 & -1 & x & & & \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & -1 & x \\ & & & & & -1 \end{pmatrix}$$

$$= x \det(C_g) + (-1)^{n-1}(-1)^{n-1} b_0 = x \det(C_g) + b_0,$$

where $g(x) = x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1$.

By induction on the size of the companion matrix, we have $\det(C_g) = g$, and so we get $\det(C_f) = xg + b_0 = f$. □

Corollary 14.16. *Let $\phi : V \to V$ be a F-linear transformation of a finite dimensional F-space V. Suppose the corresponding $F[x]$-module structure on V has invariant factors $f_1, \ldots, f_n$. Then* $\operatorname{charpoly}(\phi) = f_1 f_2 \ldots f_n$.

Proof. $\operatorname{charpoly}(\phi)$ is the same as $\operatorname{charpoly}(C)$ for a rational canonical form C of ϕ. The result follows from Lemma 14.15 and the fact that the determinant of a block diagonal matrix is the product of the determinants of the blocks. □

Let us also now discuss the minimal polynomial of a matrix or a linear transformation.

Definition 14.17. Let $f = \sum_{i=0}^{n} a_i x^i \in F[x]$. Given a matrix $A \in M_n(F)$ we define the evaluation of f at A to be $f(A) = \sum_{i=0}^{n} a_i A^i \in M_n(F)$ (where $A^0 = I$).

Note that for a fixed matrix A, the "evaulation at A" map $\epsilon_A : F[x] \to M_n(F)$ given by $\epsilon_A(f) = f(A)$ is a ring homomorphism. Thus $\ker \epsilon_A$ is an ideal of $F[x]$ and $\ker \epsilon_A = (g)$ for some $g \in F[x]$ since $F[x]$ is a PID. The map ϵ_A is also a linear transformation over F; since $\dim_F F[x] = \infty$ while $\dim_F M_n(F) = n^2$, we have $\ker \epsilon_A \neq 0$ and so $g \neq 0$.

Definition 14.18. The unique monic polynomial $g \in F[x]$ such that $\ker \epsilon_A = (g)$ is called the *minimal polynomial* of A and denoted $\operatorname{minpoly}(A)$.

Recall that a monic generator of a nonzero ideal I of $F[x]$ is the monic polynomial of uniquely smallest degree among nonzero elements of I. Thus $\operatorname{minpoly}(A)$ is the monic polynomial of smallest degree which when evaluated at A gives 0. This justifies the terminology.

Of course we can make this definition for linear transformations as well. If $\dim_F V < \infty$ and $\phi \in \mathrm{End}_F(V)$, then we can define an evaluation map $\epsilon_\phi : F[x] \to \mathrm{End}_F(V)$ by $\epsilon_\phi(\sum_{i=0}^n a_i x^i) = \sum_{i=0}^n a_i \phi^i$ (where $\phi^0 = 1_V$) and define $\mathrm{minpoly}(\phi)$ to be the unique monic generator of $\ker \epsilon_\phi$. It is easy to see that $\mathrm{minpoly}(\phi) = \mathrm{minpoly}(M_{\mathcal{B}}^{\mathcal{B}}(\phi))$ for any basis $\mathcal{B}$ of V.

Proposition 14.19. *Let $\phi : V \to V$ be a F-linear transformation of a finite dimensional F-vector space V. Let $f_1, f_2, \ldots, f_n$ be the invariant factors of V considered as an $F[x]$-module where x acts as ϕ. Then $f_n = \mathrm{minpoly}(\phi)$.*

Proof. For any commutative ring R and ideal I, $\mathrm{ann}_R(R/I) = I$. Thus $\mathrm{ann}_{F[x]} F[x]/(f_i) = (f_i)$. Also, it is clear that $\mathrm{ann}_R(M_1 \oplus \cdots \oplus M_n) = \bigcap_{i=1}^n \mathrm{ann}_R(M_i)$ for any direct sum of R-modules M_i. Since we have $V \cong F[x]/(f_1) \oplus \cdots \oplus F[x]/(f_n)$, we conclude that $\mathrm{ann}_{F[x]} V = \bigcap_{i=1}^n (f_i) = (f_n)$ since $f_i | f_n$ for $1 \le i \le n$.

We also claim that for $h \in F[x]$, $h \in \mathrm{ann}_{F[x]} V$ if and only if $h(\phi) = 0$. Writing $h = \sum_{i=0}^n a_i x^n$, we have

$$\begin{aligned}
& h \cdot v = 0 \text{ for all } v \in V \\
\iff & (\sum_{i=0}^n a_i x^n) \cdot v = 0 \text{ for all } v \in V \\
\iff & \sum_{i=0}^n a_i \phi^n(v) = 0 \text{ for all } v \in V \\
\iff & [h(\phi)](v) = 0 \text{ for all } v \in V \\
\iff & h(\phi) = 0
\end{aligned}$$

as claimed.

Thus $\mathrm{ann}_{F[x]} V = \ker \epsilon_\phi$ for the evaluation map $\epsilon_\phi : F[x] \to \mathrm{End}_F(V)$ and we conclude that $\ker \epsilon_\phi = (f_n)$. By definition, $f_n = \mathrm{minpoly}(\phi)$. □

Proposition 14.20. *Let $\phi \in \mathrm{End}_F(V)$, where V is a finite dimensional F-vector space.*

(1) $\mathrm{minpoly}(\phi) | \mathrm{charpoly}(\phi)$.

(2) *If $p(x)$ is irreducible in $F[x]$ and $p | \mathrm{charpoly}(\phi)$, then $p | \mathrm{minpoly}(\phi)$.*

Proof. (1) Let V be an $F[x]$-module where x acts as ϕ, and let $f_1, \ldots, f_n$ be the invariant factors of this module. We have seen that $\mathrm{minpoly}(\phi) = f_n$ is the largest invariant factor by Proposition 14.19, and $\mathrm{charpoly}(\phi) = f_1 f_2 \ldots f_n$ is the product of the invariant factors by Corollary 14.16. The result follows.

(2) If p is irreducible and $p|f_1 f_2 \dots f_n$, then $p|f_i$ for some i since p is prime. Since $f_i|f_n$ for all i we get that $p|f_n$. □

An immediate consequence of the results above is the following pretty result known as the Cayley-Hamilton Theorem. There are various tricky proofs of this result that may be considered more elementary since they do not rely on the theory of forms, but it is striking how the result just falls out from the simple properties of the rational canonical form we have developed.

Theorem 14.21. *Let $A \in M_n(F)$ for a field F. If $f = \text{charpoly}(A)$, then $f(A) = 0$. In other words, any matrix satisfies its own characteristic polynomial.*

Proof. We have seen in Proposition 14.20 that $g = \text{minpoly}(A)$ divides $f = \text{charpoly}(A)$ in $F[x]$; say $f = gh$. But $g(A) = 0$ by definition. So $f(A) = g(A)h(A) = 0$ as well. □

Since the elementary divisors are related to the invariant factors in a simple way, we can also relate the characteristic polynomial and minimal polynomial to these.

Lemma 14.22. *Let $\phi : V \to V$ be a F-linear transformation of a finite dimensional F-vector space V. Consider V as an $F[x]$-module where x acts as ϕ. We can write the elementary divisors of V in the form $\{p_i^{e_{i,j}} | 1 \leq i \leq m, 1 \leq j \leq s_i\}$ where the p_i are monic and pairwise distinct primes in $F[x]$, and where $e_{i,1} \geq e_{i,2} \geq \cdots \geq e_{i,s_i}$ for each i. Then $\text{charpoly}(\phi)$ is the product of all of the elementary divisors, and $\text{minpoly}(\phi) = p_1^{e_{1,1}} p_2^{e_{2,1}} \dots p_m^{e_{m,1}}$ is the product of the largest powers of the distinct primes occurring among the elementary divisors.*

Proof. This follows from now the invariant factors are related to elementary divisors, as in the proof of Theorem 13.17. In particular, since the product of the elementary divisors is the product of the invariant factors, $\text{charpoly}(\phi)$ is the product of all elementary divisors. Since the largest invariant factor is the product of the largest powers of the distinct primes occurring among the elementary divisors, we get $\text{minpoly}(\phi) = p_1^{e_{1,1}} p_2^{e_{2,1}} \dots p_m^{e_{m,1}}$. □

We can use this to give an easy to understand condition for when a particular linear transformation has a Jordan canonical form over a field F.

Corollary 14.23. *Let $\phi : V \to V$ be an F-linear transformation of a finite dimensional F-vector space V. Then ϕ has a Jordan canonical form in $M_n(F)$ if and only if $\text{charpoly}(\phi)$ splits over F.*

Proof. Consider the elementary divisors of V as an $F[x]$-module via ϕ. Then ϕ has a Jordan form if and only if those elementary divisors are all powers of degree 1 irreducibles $(x - \lambda)$ in $F[x]$. Since

charpoly(ϕ) is the product of the elementary divisors, this is if and only if charpoly(ϕ) is a product of degree 1 irreducibles, i.e. it splits over F. □

For a small matrix, often knowledge of the characteristic and minimal polynomials is enough to determine what the Jordan and rational forms must be. We illustrate this in the following simple example.

Example 14.24. Consider $A = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

Since A is upper triangular, clearly charpoly$(A) = (x-2)^2(x-1)$. Recall from Proposition 14.20(b) that every prime dividing the characteristic polynomial divides the minimal polynomial, so minpoly(A) is either $(x-2)(x-1)$ or $(x-2)^2(x-1)$.

We can calculate

$$(A-2I)(A-I) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \neq 0$$

and thus we must have minpoly$(A) = (x-2)^2(x-1) =$ charpoly(A). This implies that A has a single invariant factor $f_1 = (x-2)^2(x-1) = x^3 - 5x^2 + 8x - 4$. The rational canonical form of A is

$$\begin{pmatrix} 0 & 0 & 4 \\ 1 & 0 & -8 \\ 0 & 1 & 5 \end{pmatrix}.$$

Because charpoly(A) is a product of linear factors in $F[x]$, A also has a Jordan form over F, no matter what the field F is. The elementary divisors are $(x-2)^2$ and $(x-1)$ (note that in any field F, $1 \neq 2$).

So the Jordan form is

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

14.5. **Generalized eigenspaces and the Jordan form.** For convenience, let us assume now that the field F is algebraically closed, so that every matrix in $M_n(F)$ has a Jordan canonical form. In applications of the Jordan form it is useful to relate it to the theory of generalized eigenvalues and eigenvectors.

Let V be a vector space with $\dim_F V = n$ and let $\phi \in \mathrm{End}_F(V)$.

Definition 14.25. Given $\lambda \in F$, we say that $v \in V$ is a *generalized eigenvector* with eigenvalue λ if $(\phi - \lambda 1_V)^n(v) = 0$ for some $n \geq 1$.

Thinking of V as an $F[x]$-module where x acts as ϕ, it is equivalent to define a generalized eigenvector to be $v \in V$ such that $(x - \lambda)^n \cdot v = 0$. It is easy to see that

$$V_\lambda = \{v \in V | v \text{ is a generalized eigenvector with eigenvalue } \lambda\}$$

is an $F[x]$-submodule of V; in other words it is an F-subspace such that $\phi(V_\lambda) = V_\lambda$. V_λ is called a *generalized eigenspace*. If $(x - \lambda)^n \cdot v = 0$ and n is the minimal exponent for which this holds, then $0 \neq w = (x - \lambda)^{n-1} \cdot v$ and $(x - \lambda) \cdot w = 0$, so w is a genuine eigenvector for ϕ with eigenvalue λ. Thus the λ such that the generalized eigenspace V_λ is nonzero are exactly the eigenvalues of ϕ.

Since F is algebraically closed, the monic irreducible polynomials in $F[x]$ are just the polynomials $(x - \lambda)$ for $\lambda \in F$. We see that by definition V_λ is precisely the $(x - \lambda)$-primary component of V as defined earlier. By Proposition 13.13, $V \cong V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_m}$ for some distinct λ_i, as $F[x]$-modules. In other words, V is the direct sum of finitely many generalized eigenspaces for ϕ.

Moreover, by Proposition 13.16, each primary component V_λ is a direct sum of cyclic modules, say $V_\lambda \cong F[x]/(x - \lambda)^{e_1} \cdots \oplus \ldots F[x]/(x - \lambda)^{e_s}$. In other words, $(x - \lambda)^{e_1}, \ldots, (x - \lambda)^{e_s}$ are those elementary divisors of ϕ which are powers of the prime $(x - \lambda)$. By the proof of Theorem 13.6, we can determine the list of positive integers e_i as follows: the number of $e_i \geq b$ is the dimension over F of $V_\lambda[b]/V_\lambda[b-1]$, where $V_\lambda[b] = \{v \in V_\lambda | (x - \lambda)^b \cdot v = 0\}$. Thus it suffices to find $\dim_F V_\lambda[b]$ for each b.

Note that $V_\lambda[1]$ is precisely the space of eigenvectors for ϕ with eigenvalue λ, and $\dim_F V_\lambda[1]$ is equal to the number of the e_i, which is all of them. $\dim_F V_\lambda[1]$ is the same as the number of elementary divisors which are powers of $(x - \lambda)$. Thus the number of independent eigenvectors with eigenvalue λ is the same as the number of Jordan blocks of the Jordan form which are associated to the eigenvalue λ.

This also makes sense because a Jordan block only has a 1-dimensional space of eigenvectors; it is useful for intuition to see what is going on directly in that case:

Example 14.26. Suppose that $\phi : V \to V$ has a basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ with respect to which $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is a single Jordan block $J_{\lambda,e}$. Then $\phi(v_i) = \lambda v_i + v_{i-1}$ for $i \geq 1$, and $\phi(v_1) = \lambda v_1$, as we have seen. By definition $(\phi - \lambda 1_V)(v_i) = v_{i-1}$ for $i \geq 1$, and $(\phi - \lambda 1_V)(v_1) = 0$. From this one sees that $(\phi - \lambda 1_V)^i(v_i) = 0$ while $(\phi - \lambda 1_V)^{i-1}(v_i) = v_1 \neq 0$, for $1 \leq i \leq n$. It easily follows that $v_i + V_\lambda[i-1]$ is a basis of $V_\lambda[i]/V_\lambda[i-1]$ and so all of these factor spaces are 1-dimensional, for $1 \leq i \leq n$. This is consistent with the fact that the corresponding $F[x]$-module structure on V

is isomorphic to $F[x]/(x-\lambda)^n$ and there is only one elementary divisor $(x-\lambda)^n$. The eigenspace of ϕ is 1-dimensional, so ϕ is highly defective in the sense that V is far from being spanned by eigenvectors (but every element of V is a generalized eignevector for the eigenvalue λ).

We can use the generalized eigenvector point of view to help determine the Jordan form of a matrix. We give a simple example next.

Example 14.27. Let $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & -2 & 0 & 1 \\ -2 & 0 & -1 & 2 \end{pmatrix}$.

We would like to find the Jordan form of A over $\mathbb{C}$. To put this explicitly in the context of the previous discussion, we can let $V = \mathbb{C}^4$ be a space column vectors and make it an $F[x]$-module where x acts by the linear transformation $\phi : V \to V$ determined by left multiplication by A.

The first step is to calculate the characteristic polynomial of A:

$$\det \begin{pmatrix} x-1 & 0 & 0 & 0 \\ 0 & x-1 & 0 & 0 \\ 2 & 2 & x & -1 \\ 2 & 0 & 1 & x-2 \end{pmatrix} = (x-1)^2 \det \begin{pmatrix} x & -1 \\ 1 & x-2 \end{pmatrix} = (x-1)^2(x^2-2x-1) = (x-1)^4.$$

We see that A has only one eigenvalue, $\lambda = 1$. Thus the elementary divisors of A must all be of the form $(x-1)^i$. Moreover, by Lemma 14.22, the product of the elementary divisors is $(x-1)^4$. Let the elementary divisors be $(x-1)^{e_1}, (x-1)^{e_2}, (x-1)^{e_3}, \ldots$ (we don't know how many there are yet, though there are at most 4).

We first calculate the dimension of the 1-eigenspace of A. Since

$$(A - I) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -2 & -2 & -1 & 1 \\ -2 & 0 & -1 & 1 \end{pmatrix}$$

is clearly a matrix of rank 2, its nullspace has dimension 2. This means we have two linearly independent eigenvectors and the Jordan form has two Jordan blocks.

We still don't know if those blocks have size 1 and 3 or size 2 and 2. For this we can easily calculate that

$$(A-I)^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix}.$$

This matrix as rank 1, so nullspace of dimension 3.

We have $V = V_1$, that is, V is the 1-generalized eigenspace. In the notation above, $\dim_F V_1[1] = 2$, and $\dim_F V_1[2] = 3$. This says that there are 2 e_i's with $e_i \geq 1$, and $\dim_F V_1[2]/V_1[1] = 3 - 2 = 1$ of the e_i's with $e_i \geq 2$. The only possibility is that $e_1 = 3$, $e_2 = 1$ and the elementary divisors are $(x-1)^3, (x-1)^1$. Thus the Jordan form of A is

$$J = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is also easy to determine the rational canonical form. Clearly the invariant factors are also $f_1 = (x-1)$, $f_2 = (x-1)^3$. The minimal polynomial is thus $f_2 = (x-1)^3 = x^3 - 3x^2 + 3x - 1$. The rational form is

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & 3 \end{pmatrix}.$$

We did an example above with only one eigenvalue to demonstrate the method; in general, one would need to calculate the dimensions of the nullspaces of $(A - \lambda_i I)^j$ for each of the eigenvalues λ_i and each $j \geq 1$ until one had enough information to determine the sizes of the Jordan blocks associated to λ_i.

Throughout this section, we have concentrated on the theoretical aspects of canonical forms rather than methods of calculation. The reader interested in computations can find more details in Chapter 12 of Dummit and Foote's book.

15. Tensor products

15.1. **Balanced maps.** Tensor products are very important gadgets, but they take a while to get comfortable with and see why they are so useful. It is natural the first time you learn about them

to understand the basics of how to manipulate them, but feel like you still don't quite "get" them. I think once they appear in your later classes and you see them in action they tend to sink in better.

For a little motivation we consider the case of vector spaces (i.e. modules over a field F) where tensor products are already interesting. Let V and W be F-vector spaces. As we know, we have the direct sum of vector spaces $V \oplus W$ which as a set is just the cartesian product $V \times W = \{(v, w) | v \in V, w \in W\}$. There may be natural functions from $V \times W$ to another vector space U which are not F-linear maps, but rather *F-linear in each coordinate separately.* Here is a very common example.

Example 15.1. Let V be a vector space over F and define $V^* = \mathrm{Hom}_F(V, F)$. Since F is a commutative ring, V^* is again an F-module, i.e. vector space, as we have seen. It is called the *dual* vector space, or the space of linear functionals on V. There is a very natural function $\theta : V^* \times V \to F$ defined by $\theta(\psi, v) = \psi(v)$. (Note that for notational convenience we write $\theta(\psi, v)$ instead of $\theta((\psi, v))$, which would be more technically correct.) This function satisfies $\theta(\psi+\phi, v) = [\psi+\phi](v) = \psi(v)+\phi(v) = \theta(\psi, v)+\theta(\phi, v)$ as well as $\theta(\psi, v+w) = \psi(v+w) = \psi(v)+\psi(w) = \theta(\psi, v)+\theta(\psi, w)$. Thus θ is linear in each coordinate and we say that θ is *bilinear*. However, θ is not an F-linear transformation from $V^* \oplus V$ to F. This would require $\theta((\psi, v) + (\phi, w)) = \theta(\psi, v) + \theta(\phi, w)$. But the left hand side of this equation is $\theta(\psi + \phi, v + w) = [\psi + \phi](v + w) = \psi(v) + \phi(v) + \psi(w) + \phi(w)$, while the right hand side is $\psi(v) + \phi(w)$. These are certainly not equal in general.

On the other hand, when doing linear algebra one would really like to work with linear maps. We are going to define a vector space $V^* \otimes_F V$, the tensor product of V^* and V over F, and an F-linear map $\widetilde{\theta} : V^* \otimes_F V \to F$ which contains the same information as the map θ. In some sense $V^* \otimes_F V$ is the vector space where bilinear functions like θ naturally live. We will see that to make this work $V^* \otimes_F V$ will have to be a bigger vector space than $V^* \oplus V$.

Let us begin the technical work to define tensor products. We are going to make the main definitions for modules over an arbitrary, possibly noncommutative ring R.

Definition 15.2. Let R be a ring. Let M be a right R-module and N a left R-module. Let P be any abelian group. A function $\phi : M \times N \to P$ is called *R-balanced* if

(1) $\phi(m_1 + m_2, n) = \phi(m_1, n) + \phi(m_2, n)$ for all $m_1, m_2 \in M$ and $n \in N$.

(2) $\phi(m, n_1 + n_2) = \phi(m, n_1) + \phi(m, n_2)$ for all $m \in M$ and $n_1, n_2 \in N$.

(3) $\phi(mr, n) = \phi(m, rn)$ for all $m \in M$, $r \in R$, and $n \in N$.

Note that the first two conditions say that an R-balanced map respects addition separately in each coordinate. It may seem awkward that this definition is made for one right module and one left

module; this will also appear in the definition of tensor product to come shortly and it is essential to make the theory work properly. In the applications to modules over commutative rings R, of course, we can identify left and right modules and one can stick to modules over one side. We will make more comments about this later.

Example 15.3. Consider again the map $\theta : V^* \times V \to F$ defined by $\theta(\phi, v) = \phi(v)$ that was defined in Example 15.1. Since F is commutative, we can think of V^* as a right F-module just as well, where $\phi \cdot a = a\phi$ for $a \in F$, $\phi \in V^*$. With this convention we claim that the map θ is F-balanced. We have already seen that it respects sums separately in each coordinate. Moreover

$$\theta(\phi \cdot a, v) = [\phi \cdot a](v) = [a\phi](v) = a\phi(v) = \phi(av) = \theta(\phi, av),$$

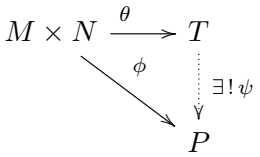
for all $\phi \in V^*$, $a \in F$, and $v \in V$.

Example 15.4. Let M be a left R-module. Note that R is naturally a right R-module by right multiplication. Then the map $\psi : R \times M \to M$ defined by $\psi(r, m) = rm$ is R-balanced.

We now define a tensor product as a universal object for R-balanced maps from $M \times N$ to any abelian group.

Definition 15.5. Let M be a right R-module, and N a left R-module. A *tensor product* for M and N (over R) is an abelian group T together with an R-balanced map $\theta : M \times N \to T$ with the following universal property: for any abelian group P and R-balanced map $\phi : M \times N \to P$, there exists a unique homomorphism of abelian groups $\psi : T \to P$ such that $\theta\psi = \phi$.

Here is the commutative diagram which represents the universal propery of the tensor product:



The important thing is that ϕ is not a homomorphism of abelian groups (it does not respect addition), while ψ *is*. Moreover, since θ is a given part of the structure of the tensor product, the homomorphism ψ contains all of the information in the map ϕ (as we can recover ϕ by $\phi = \theta\psi$). So the tensor product allows one to replace balanced maps by actual homomorphisms, without losing information.

As a warmup, let us give an example of a tensor product in a special case where we can check directly that the definition holds.

Example 15.6. Let R be a right R-module by right multiplication, and let M be a left R-module. Then the R-balanced map $\theta : R \times M \to M$ from Example 15.4, where $\theta(r, m) = rm$, is a tensor product of R and M over R.

Proof. Suppose we have an abelian group P and an R-balanced map $\phi : R \times M \to P$. We need to find a group homomorphism $\psi : M \to P$ such that $\psi\theta = \phi$, and show that ψ is unique with this property.

Define $\psi : M \to P$ by $\psi(m) = \phi(1, m)$. Then since ϕ is R-balanced, for all $r \in R$ and $m \in M$ we have

$$\phi(r, m) = \phi(1r, m) = \phi(1, rm) = \psi(rm) = \psi\theta(r, m).$$

Thus $\psi\theta = \phi$. ψ is certainly uniquely determined, since if $\psi'\theta = \phi$ then $\psi'(m) = \psi'\theta(1, m) = \phi(1, m)$ and so $\psi' = \psi$.

Finally, ψ is a group homomorphism since $\psi(m_1 + m_2) = \phi(1, m_1 + m_2) = \phi(1, m_1) + \phi(1, m_2) = \psi(m_1) + \psi(m_2)$, using that ϕ is R-balanced. □

15.2. **Existence and uniqueness.** Let us first state the uniqueness of tensor products up to isomorphism. This follows by exactly the same argument as we have already seen for other universal properties, and so we leave the proof to the reader.

Proposition 15.7. *Let M be a right R-module and N a left R-module. Suppose that $\theta_1 : M \times N \to T_1$ and $\theta_2 : M \times N \to T_2$ are both tensor products of M and N over R. Then there is a unique isomorphism of abelian groups $\theta : \psi : T_1 \to T_2$ such that $\psi\theta_1 = \theta_2$.*

The proposition shows that there is essentially only one tensor product of M and N over R (if there is any). Now let us show that the tensor product always exists.

Theorem 15.8. *Let M be a right R-module and N a left R-module. Then there is an abelian group T and an R-balanced map $\theta : M \times N \to T$ which is a tensor product of M and N over R.*

Proof. Consider $S = M \times N = \{(m, n) | m \in M, n \in N\}$ as a set. Construct a free $\mathbb{Z}$-module (i.e. abelian group) indexed by S, in other words $F = \bigoplus_{s \in S} \mathbb{Z}$. We introduce a new formal symbol $m \otimes n$ to represent the standard basis element which is 1 in the (m, n)-spot and 0 elsewhere. Then a general element of F looks like $a_1(m_1 \otimes n_1) + \cdots + a_k(m_k \otimes n_k)$ for $a_i \in \mathbb{Z}$, $m_i \in M$, $n_i \in N$.

Now let $T = F/I$ where I is the subgroup of F generated by all elements of the form

$$\begin{aligned} (m_1 + m_2) \otimes n - m_1 \otimes n - m_2 \otimes n, \qquad & m_1, m_2 \in M,\ n \in N; \\ m \otimes (n_1 + n_2) - m \otimes n_1 - m \otimes n_2 \qquad & m \in M,\ n_1, n_2 \in N; \\ mr \otimes n - m \otimes rn \qquad & m \in M,\ r \in R,\ n \in N. \end{aligned}$$

We claim that $\theta : M \times N \to T = F/I$ defined by $\theta(m, n) = (m \otimes n) + I$ is a tensor product of M and N over R.

First we need that θ is R-balanced. This is immediate from the "relations" we threw into the subgroup I. For example, let us check the third condition of the R-balanced property:

$$\theta(mr, n) = (mr \otimes n) + I = (m \otimes rn) + I = \theta(m, rn) \text{ for all } m \in M, r \in R, n \in N,$$

where we have used that the cosets $(mr \otimes n)+I$ and $(m \otimes rn)+I$ are equal because $mr \otimes n - m \otimes rn \in I$ by definition. The other two conditions of the balanced property, that sums are respected in each coordinate, follow immediately in the same way.

Next, we need that if $\phi : M \times N \to P$ is R-balanced, for some abelian group P, there there is a unique linear map $\psi : T \to P$ such that $\phi = \psi\theta$. This also follows very formally from the fact that F is free, and that to produce T we have modded out the subgroup generated exactly those relations that represent being R-balanced.

More precisely, we first get that there is a unique homomorphism of $\mathbb{Z}$-modules (i.e. abelian groups) $\widehat{\psi} : F \to P$ such that $\widehat{\psi}(m \otimes n) = \phi(m, n)$ for all $m \in M$, $n \in N$. This is just because F is a free $\mathbb{Z}$-module on the basis $\{m \otimes n | m \in M, n \in N\}$. Second, because ϕ is R-balanced, it is easy to check that every element in the generating set of I must be in the kernel of $\widehat{\psi}$. Since $\ker \widehat{\psi}$ is a subgroup of F, it must contain all of I. This implies that $\widehat{\psi}$ factors through I to give a group homomorphism $\psi : T = F/I \to P$ such that $\psi((m \otimes n) + I) = \phi(m, n)$. The fact that $\phi = \psi\theta$ is immediate.

Finally, if ψ' were another homomorphism such that $\phi = \psi'\theta$, we would have $\psi((m \otimes n) + I) = \phi(m, n) = \psi'\theta(m, n) = \psi'((m \otimes n)+I)$ for all $m \in M$, $n \in N$. Since the basis elements $\{(m \otimes n) | m \in M, n \in N\}$ generate F as an abelian group, their images $\{(m \otimes n) + I | m \in M, n \in N\}$ generate $T = F/I$ as an abelian group. Thus ψ' and ψ agree on a generating set of T. Since they are group homomorphisms, $\psi' = \psi$. □

Remark 15.9. From now on we use the following standard notation. Given a right R-module M and a left R-module N, we now know from Theorem 15.8 that there exists a tensor product of M and N over R, given by a group homomorphism $\theta : M \times N \to T$ for some abelian group T. By

Proposition 15.7, this tensor product is unique up to isomorphism. The standard notation for the abelian group T is $M \otimes_R N$, and we will adopt this notation from now on.

Technically the tensor product of M and N over R is the abelian group $M \otimes_R N$ together with a R-balanced map $\theta : M \times N \to M \otimes_R N$. In practice, we refer to the abelian group $M \otimes_R N$ as the tensor product and suppress the map θ. Instead θ is remembered by writing the element $\theta(m, n)$ as $m \otimes n$ for each $m \in M, n \in N$ (this notation was already suggested by the notation used in the proof of Theorem 15.8). These elements $m \otimes n$ in $M \otimes_R N$ are referred to as *pure tensors*. The fact that θ is R-balanced means that we have the following rules for manipulating pure tensors:

$$(m_1 + m_2) \otimes n = m_1 \otimes n + m_2 \otimes n \text{ for all } m_1, m_2 \in M, n \in N;$$

$$m \otimes (n_1 + n_2) = m \otimes n_1 + m \otimes n_2 \text{ for all } m \in M, n_1, n_2 \in N; \text{and}$$

$$mr \otimes n = m \otimes rn \text{ for all } m \in M, r \in R, n \in N.$$

As we will see shortly, θ is not surjective in general, and so it is important to realize that not all elements of $M \otimes_R N$ are equal to pure tensors. Rather, by the construction in Theorem 15.8, an arbitrary element of $M \otimes_R N$ has the form $\sum_{i=1}^{d} a_i(m_i \otimes n_i)$ for $a_i \in \mathbb{Z}$, $m_i \in M$, and $n_i \in N$. Using that the map θ is additive in the first coordinate, this is the same as $\sum_{i=1}^{d}(a_i m_i \otimes n_i) = \sum_{i=1}^{d}(m_i' \otimes n_i)$ for some $m_i' \in M$. We conclude that *every element of $M \otimes_R N$ is a finite sum of pure tensors*. Of course in general an element can be written as a sum of pure tensors in many different ways.

We have seen that the tensor product $M \otimes_R N$ always exists and is unique up to isomorphism. The proof of existence in Theorem 15.8 is very formal, and unfortunately it does not really give any intuition for what a particular tensor product looks like. For example, the tensor product of R and M over R is given by the natural multiplication map $R \times M \to M$, as we saw in Example 15.6. The proof of Theorem 15.8 also constructs a tensor product of R and M over R as a factor group of a massive free abelian group. This must be isomorphic to M as an abelian group in this case; but this is certainly not obvious. In practice, when working with tensor products, it is usually best to try to understand them using their defining universal property and to forget the formal construction as a factor group of a free abelian group which appeared in the proof of Theorem 15.8.

The tensor product can behave in ways that are quite unintuitive at first. For example, it can easily happen that the tensor product of two nonzero modules is 0.

Lemma 15.10. *In any tensor product $M \otimes_R N$, we have $0 \otimes n = 0 = m \otimes 0$ for all $m \in M, n \in N$.*

Proof. We have $(0 \otimes n) = (0 + 0) \otimes n = 0 \otimes n + 0 \otimes n$. Subtracting, we get $0 \otimes n = 0$. Similarly, $0 = m \otimes 0$. □

Example 15.11. Let G be a torsion abelian group, thought of as a right $\mathbb{Z}$-module. Consider $\mathbb{Q}$ as a left $\mathbb{Z}$-module as usual. Then we claim that $G \otimes_{\mathbb{Z}} \mathbb{Q} = 0$.

Consider a pure tensor in $G \otimes_{\mathbb{Z}} \mathbb{Q}$. It has the form $g \otimes a/b$ for $a, b \in \mathbb{Z}$ with $b \neq 0$. Since G is torsion, $g \cdot n = ng = 0$ for some $n \geq 1$. Then

$$g \otimes a/b = g \otimes an/bn = gn \otimes a/bn = 0 \otimes a/bn = 0,$$

using Lemma 15.10. Thus all pure tensors are equal to 0. Since every element of $G \otimes_{\mathbb{Z}} \mathbb{Q}$ is a finite sum of pure tensors, $G \otimes_{\mathbb{Z}} \mathbb{Q} = 0$ as claimed.

15.3. **Functoriality; module structure on the tensor product.** It is important that the formation of tensor products is *functorial*: this means that the operation $- \otimes_R N$ of tensoring modules with N respects homomorphisms (and similarly in the other coordinate). Here is what we mean precisely.

Lemma 15.12. *Let R be a ring, let M, M' be right R-modules, and let N, N' be left R-modules.*

(1) *Let $f : M \to M'$ be a homomorphism of right R-modules. Then there is a homomorphism of abelian groups $f \otimes 1 : M \otimes_R N \to M' \otimes_R N$ given by $[f \otimes 1](m \otimes n) = f(m) \otimes n$.*
(2) *Let $g : N \to N'$ be a homomorphism of left R-modules. Then there is a homomorphism of abelian groups $1 \otimes g : M \otimes_R N \to M \otimes_R N'$ given by $[1 \otimes g](m \otimes n) = m \otimes g(n)$.*

Before beginning the proof we make some comments about statements like this. It is not at all clear that a formula such as $[f \otimes 1](m \otimes n) = f(m) \otimes n$ defines a function. The problem is that it is not clear it is well-defined, as there are many relations among the pure tensors; for example, two pure tensors might well be equal. (Remember that $m \otimes n$ means $\theta(m, n)$ for the underlying map $\theta : M \times N \to M \otimes_R N$ of the tensor product, and there is no reason why θ should be injective.) So we need to make sure those relations are respected. The best way to do this is to use the universal property of the tensor product, as we will see in the proof.

Note also that the formula $[f \otimes 1](m \otimes n) = f(m) \otimes n$, even once we show it is well-defined, only gives the action of the function on pure tensors. But since we require $f \otimes 1$ to be a homomorphism of groups, the action on an arbitrary element, i.e. a sum of pure tensors, is uniquely determined. For this reason it is standard to use only pure tensors in the formulas for functions and actions, and often one only verifies these formulas for pure tensors in proofs. You should be careful not to let the appearance of those formulas seduce you into forgetting that not every element of the tensor product is a pure tensor.

Proof. (1) We define a function $\phi : M \times N \to M' \otimes N$ by $\phi(m, n) = f(m) \otimes n$. Using the rules for manipulating the tensor product symbol and the fact that f is a homomorphism of right modules, it is easy to check that ϕ is R-balanced. It follows from the universal property of the tensor product that there is a unique group homomorphism $f \otimes 1 : M \otimes_R N \to M' \otimes_R N$ such that $[f \otimes 1](m \otimes n) = f(m) \otimes n$ for $m \in M, n \in N$, as required.

(2) This is proved in a symmetric manner. □

Before studying some more examples we should discuss when $M \otimes_R N$ is actually a module and not just an abelian group.

Definition 15.13. Let M be an abelian group. Then M is called an (R, S)-bimodule if M is both a left R-module and a right S-module, and the two module structures are compatible in the sense that $(rm)s = r(ms)$ for all $r \in R, m \in M, s \in S$.

Proposition 15.14. *Let M be a right R-module and N a left R-module.*

(1) *If M is an (S, R)-bimodule, then $M \otimes_R N$ is a left S-module, where $s \cdot (m \otimes n) = sm \otimes n$.*

(2) *If N is an (R, T)-bimodule, then $M \otimes_R N$ is a right T-module, where $(m \otimes n) \cdot t = m \otimes nt$.*

(3) *If both (1) and (2) occur then $M \otimes_R N$ is an (S, T)-bimodule.*

Proof. (1) For any $s \in S$, define $\ell_s : M \to M$ by $\ell_s(m) = sm$. This "left multiplication by s" map is not a homomorphism of M as a left S-module in general (unless S is commutative) but it *is* always a right R-module map, since $\ell_s(mr) = s(mr) = (sm)r = \ell_s(m)r$ for $r \in R$.

By the functoriality of the tensor product given in Lemma 15.12, we get a homomorphism of abelian groups $\ell_s \otimes 1 : M \otimes_R N \to M \otimes_R N$ such that $[\ell_s \otimes 1](m \otimes n) = (sm \otimes n)$. Thus there is in fact a well-defined left action of S on $M \otimes_R N$ for which the action on pure tensors is given by $s \cdot (m \otimes n) = [\ell_s \otimes 1](m \otimes n) = sm \otimes n$, and for which left action by s is an abelian group homomorphism. This also implies one of the module axioms ($(s \cdot (x + y) = s \cdot x + s \cdot y)$ for $s \in S$, $x, y \in M \otimes_R N$), and the others are easy to check.

(2) This is proved by a completely symmetric proof to the proof of part (1).

(3) On pure tensors we have

$$[s \cdot (m \otimes n)] \cdot t = (sm \otimes n) \cdot t = (sm \otimes nt) = s \cdot (m \otimes nt) = s \cdot [(m \otimes n) \cdot t]$$

and this extends immediately to the action on a finite sum of pure tensors, i.e. a general element of $M \otimes_R N$. □

Proposition 15.14 is analogous to earlier observations we made about $\mathrm{Hom}_R(M, N)$ for two left R-modules M and N. In general this Hom-space is just an abelian group, but in an exercise on the

homework you verified that if either M or N is a bimodule then $\mathrm{Hom}_R(M,N)$ obtains a module structure as well, and it is a bimodule if both M and N are bimodules.

Example 15.15. Let R be a ring with ideal I. Thus R/I is naturally an (R,R)-bimodule. Let M be a left R-module. We claim that $(R/I)\otimes_R M \cong M/IM$ as left R-modules. (This generalizes Example 15.6.)

Since R/I is an (R,R)-bimodule, $R/I\otimes_R M$ is a left R-module by Proposition 15.14. Define a map $\phi : R/I\times M \to M/IM$ by $\phi(r+I,m) = rm+IM$. If $r+I = r'+I$, then $r-r'\in I$, so $(r-r')m\in IM$ and hence $rm+IM = r'm+IM$. Hence ϕ is well-defined. It is now easy to check that ϕ is R-balanced. Thus the universal property of the tensor product gives us a unique group homomorphism $\psi : R/I\otimes M\to M/IM$ such that $\psi((r+I)\otimes m) = rm+IM$ for all $r\in R, m\in M$. But from this formula, in fact we see that ψ is an R-module homomorphism, since for $x\in R$ we have $\psi(x(r+I\otimes m)) = \psi(xr+I\otimes m) = (xr)m+IM = x(rm)+IM = x(rm+IM) = x\psi(r+I\otimes m)$.

To show that ψ is an isomorphism, one can define an inverse map $\rho : M/IM\to (R/I)\otimes_R M$ by $\rho(m+IM) = (1+I)\otimes m$. To see that this is well defined, if $m+IM = m'+IM$, then $m-m'\in IM$, say $m-m' = \sum x_i m_i$ with $x_i\in I$, $m_i\in M$. Then

$$(1+I)\otimes m-(1+I)\otimes m' = (1+I)\otimes(m-m') = (1+I)\otimes\sum x_i m_i = \sum(1+I)x_i\otimes m_i$$
$$= \sum(x_i+I)\otimes m_i = \sum(0+I)\otimes m_i = 0.$$

Thus ρ is well defined. It is obvious that $\psi\rho = 1_{M/IM}$. On the other hand, $\rho\psi((r+I)\otimes m) = \rho(rm+I) = (1+I)\otimes rm = (1+I)r\otimes m = (r+I)\otimes m$, so $\rho\psi$ is also the identity and we are done.

15.4. **The commutative case.** The tensor product over a commutative ring R is often described in a slightly different way. Recall that any right R-module M is naturally also a left R-module with $r\cdot m = mr$; symmetrically, every left module is also a right module. So there is no need to use one left and one right module in the definition of a tensor product, and usually the definition is made in terms of left modules only.

Definition 15.16. Let R be a commutative ring and let M,N, and P be left R-modules. A function $\phi : M\times N\to P$ is *R-bilinear* if

(1) $\phi(r_1m_1+r_2m_2,n) = r_1\phi(m_1,n)+r_2\phi(m_2,n)$ for all $r_1,r_2\in R$, $m_1,m_2\in M$, $n\in N$; and
(2) $\phi(m,r_1n_1+r_2n_2) = r_1\phi(m,n_1)+r_2\phi(m,n_2)$ forall $r_1,r_2\in R$, $m\in M$, $n_1,n_2\in N$.

In the commutative case, the universal property for the tensor product can be (and usually is) phrased in terms of R-bilinear maps, as in part (3) of the next result.

Proposition 15.17. *Let R be a commutative ring and let M and N be left R-modules. Considering M as a right R-module, we can define $T = M \otimes_R N$. Then*

(1) *T is naturally a left R-module with $r \cdot (m \otimes n) = rm \otimes n = m \otimes rn$.*

(2) *The map $\theta : M \times N \to M \otimes_R N$ given by $\theta(m, n) = m \otimes n$ is R-bilinear.*

(3) *Given any R-module P and an R-bilinear map $\phi : M \times N \to P$, there is a unique R-module homomorphism $\psi : M \otimes_R N \to P$ such that $\psi\theta = \phi$.*

Proof. (1) Since R is commutative, we can actually think of M as an (R, R)-bimodule with the same left and right actions: $mr = rm$ for $r \in R$, $m \in M$. (This is a bimodule since $s(mr) = s(rm) = (sr)m = (rs)m = r(sm) = (sm)r$.) Then $T = M \otimes_R N$ obtains a left R-module structure by Proposition 15.14, where $r \cdot (m \otimes n) = rm \otimes n$. We also have $rm \otimes n = mr \otimes n = m \otimes rn$.

(2) By the R-module structure on T given in (1), we have

$$(r_1 m_1 + r_2 m_2) \otimes n = (r_1 m_1 \otimes n) + (r_2 m_2 + n) = r_1(m_1 \otimes n) + r_2(m_2 \otimes n)$$

and

$$m \otimes (r_1 n_1 + r_2 n_2) = m \otimes r_1 n_1 + m \otimes r_2 n_2 = r_1(m \otimes n_1) + r_2(m \otimes n_2).$$

(3) The map ϕ is also R-balanced, since $mr \otimes n = rm \otimes n = m \otimes rn$ by (1). Thus there is a unique group homomorphism $\psi : T \to P$ such that $\psi(m \otimes n) = \phi(m, n)$. Then ψ is an R-module homomorphism because $\psi(r(m \otimes n)) = \psi(rm \otimes n) = \phi(rm, n) = r\phi(m, n) = r\psi(m \otimes n)$. □

An important special case of the tensor product over a commutative ring R is the case where R is a field F. In this case one can get a very explicit description of the tensor product of two vector spaces over F.

Theorem 15.18. *Let V and W be vector spaces over the field F. Suppose that $\{v_i | i \in I\}$ is an F-basis for V and $\{w_j | j \in J\}$ is an F-basis for W.*

Then $V \otimes_F W$ is an F-vector space with basis $\{v_i \otimes w_j | i \in I, j \in J\}$. In particular, if $\dim_F V = m$ and $\dim_F W = n$ then $\dim_F V \otimes W = mn$.

Proof. Given a pure tensor $v \otimes w \in V \otimes_F W$, write $v = \sum a_i v_i$ and $w = \sum b_j w_j$, where all but finitely many of the a_i and b_j are nonzero. Then

$$v \otimes w = (\sum_i a_i v_i) \otimes (\sum_j b_j w_j) = \sum_i \sum_j a_i b_j (v_i \otimes w_j).$$

Thus any pure tensor is in the F-span of $\{v_i \otimes w_j | i \in I, j \in J\}$. We have also seen that an arbitrary element of $V \otimes_F W$ is a finite sum of pure tensors. Hence $\{v_i \otimes w_j | i \in I, j \in J\}$ spans $V \otimes_F W$.

Now suppose that we have an independence relation $\sum_{i,j} a_{ij}(v_i \otimes w_j) = 0$, where $a_{ij} = 0$ for all but finitely many $(i,j) \in I \times J$. Then we have

$$0 = \sum_i \sum_j (a_{ij} v_i \otimes w_j) = \sum_j (\sum_i a_{ij} v_i) \otimes w_j = \sum_j v'_j \otimes w_j$$

for some elements $v'_j \in V$. For each $j \in J$ let $w_j^* \in W^* = \mathrm{Hom}_F(W, F)$ be the linear functional with $w_j^*(w_i) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

For any $k \in J$, the map $\phi : V \times W \to V$ defined by $\phi(v, w) = w_k^*(w)v$ is F-bilinear, as is easy to check. Thus there is an F-linear map $\psi : V \otimes_F W \to V$ with $\psi(v \otimes w) = w_k^*(w)v$. Applying this to our relation above gives $0 = \psi(0) = \sum_j \psi(v'_j \otimes w_j) = \sum_j w_k^*(w_j)v'_j = v'_k$. We conclude that $v'_k = 0$ for all k. Then $v'_k = \sum_i a_{ik} v_i = 0$, and by the linear independence of the v_i, we get $a_{ik} = 0$ for all i. Since k was arbitrary, $a_{ik} = 0$ for all $i \in I, k \in J$, and thus $\{v_i \otimes w_j | i \in I, j \in J\}$ is also F-independent. Hence it is an F-basis as claimed. □

15.5. **Extension of scalars.** Suppose that we have a ring homomorphism $\phi : R \to S$. Recall that if M is a left S-module, there is an easy way to make M into a left R-module: just define $r \cdot m = \phi(r)m$ for $r \in R$, $m \in m$. This is called *restriction of scalars*, as we have already mentioned, since in the case where R is a subring of S and $\phi : R \to S$ is just the inclusion map, then we areliterally just restricting the elements that act on M to a smaller set.

Now that we have developed the tensor product, we can easily define a process that goes the other way.

Definition 15.19. Let $\phi : R \to S$ be a ring homomorphism. Suppose that M is a left R-module. Then $S \otimes_R M$ is naturally a left S-module, where $s \cdot (t \otimes m) = st \otimes m$. This process is called *extension of scalars*.

Again, when R is a subring of S we are extending the ring acting to a larger ring; hence the name.

The fact that $S \otimes_R M$ is a left S-module with action on pure tensors by the given formula is immediate from the fact that S is an (S, R)-bimodule and Proposition 15.14. Here the bimodule structure on S is given by the natural S-action by multiplication on the left, and the right R-action is $s \cdot r = s\phi(r)$, for $s \in S$, $r \in R$. In other words, S is a right R-module by restricting the scalars from the action of S on the right by multiplication.

Let us give several applications of extension by scalars.

Example 15.20. Suppose that $F \subseteq K$ is an inclusion of fields. If V is an F-vector space, then $K \otimes_F V$ is a K-vector space by extension of scalars.

Suppose that $\mathcal{B} = \{v_i | i \in I\}$ is an F-basis of V. Then $\mathcal{B}' = \{1 \otimes v_i | i \in I\}$ is a K-basis of $K \otimes_F V$ (we leave this as an exercise). In other words, $\dim_K(K \otimes_F V) = \dim_F V$. Moreover, suppose that $\phi : V \to V$ is an F-linear transformation. Then $1 \otimes \phi : K \otimes_F V \to K \otimes_F V$ is a K-linear transformation. If $\dim_F V$ is finite and we consider the matrix $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ of ϕ with respect to $\mathcal{B}$, then one may check that $M_{\mathcal{B}'}^{\mathcal{B}'}(1 \otimes \phi)$ is the same matrix $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$. So the linear transformation is given by the same matrix, just working over a larger field.

This is a very useful operation, for example, if we are initially working over a field F and would like to work over an algebraically closed one K. (We will prove later in the notes that F is always a subfield of an algebraically closed field K.) By extending V to $K \otimes_F V$ and the map ϕ to $1 \otimes \phi$ we put ourselves in a setting where the Jordan canonical form is defined.

Example 15.21. Suppose that R is an integral domain. Let K be the field of fractions of R. If M is a left R-module we define the *rank* of M to be $\dim_K(K \otimes_R M)$.

The point is that by extending scalars to a field K, we have access to the notion of dimension of a vector space, which was not available over the original ring R. One may check that if R is a PID and M is a finitely generated R-module, then writing $M \cong R^r \oplus T$ where T is torsion, the rank of M is r. So this notion of rank just recovers the rank of the free part of a finitely generated module in this case.

The rank is defined over any integral domain however, which makes it more generally applicable. One may show that in general the rank of a finitely generated module M is equal to the maximum r such that there is an R-submodule of M isomorphic to R^r. Though we could use this as a definition of rank instead, the properties of rank are easiest to prove by extending scalars to K.

Example 15.22. Suppose R is any commutative ring. We know that R has some maximal ideal $\mathfrak{m}$. We have the homomorphism $\phi : R \to R/\mathfrak{m}$ where $F = R/\mathfrak{m}$ is a field.

This gives another way one can sometimes reduce problems about R-modules to problems about fields. Given an R-module M, when we "extend" scalars using the map ϕ we get the F-module $F \otimes_R M \cong M/\mathfrak{m}M$, using Example 15.15.

If M is a free R-module, say $M \cong \bigoplus_{i \in I} R$, then by picking a basis one may show that $F \otimes_R M \cong \bigoplus_{i \in I} F$ as F-modules; that is, $F \otimes_R M$ is a vector space with dimension $|I|$ over F. This can be used to show that two isomorphic free modules over R must have the same rank, by reducing to the case of vector spaces.

As another example, Let R be a local commutative ring with maximal ideal $\mathfrak{m}$ (that is, $\mathfrak{m}$ is the unique maximal ideal of R). In this case $F = R/\mathfrak{m}$ is called the *residue field*. If M is a finitely generated R-module, then $\dim_F(F \otimes_R M) = \dim_F M/\mathfrak{m}M$ is equal to the minimum number n such that M can be generated by n elements as an R-module. This is a consequence of the result known as Nakayama's Lemma.

15.6. **Tensor products of algebras.**

Definition 15.23. Let R be a commutative ring. A ring A is an *R-algebra* if A is also a left R-module, and for all $r \in A$, $a, b \in A$ we have $r \cdot ab = (r \cdot a)b = a(r \cdot b)$.

A homomorphism of R-algebras is a function $f : A \to B$ which is both a ring homomorphism and a R-module homomorphism. It is an isomorphism of algebras if f is bijective. A *subalgebra* of an algebra A is a subset which is both a submodule over R and a subring; so it is naturally an R-algebra again.

Thus an algebra is both a ring and a module, with those structures being compatible in a certain way. The compatibility can be framed in the following alternative way which is perhaps more natural.

Remark 15.24. If A is an R-algebra, then there is a function $\phi : R \to A$ given by $\phi(r) = r \cdot 1$. This is clearly a homomorphism of abelian groups with $\phi(1) = 1$; moreover $\phi(rs) = (rs) \cdot 1 = r \cdot (s \cdot 1)$ (by module axioms) and $r \cdot (s \cdot 1) = r \cdot ((1)(s \cdot 1)) = (r \cdot 1)(s \cdot 1) = \phi(r)\phi(s)$. Thus ϕ is a homomorphism of rings.

We also check using the axioms of an algebra that $(r \cdot 1)a = r \cdot (1a) = r \cdot a = r \cdot (a1) = a(r \cdot 1)$ for all $r \in R$, $a \in A$. This shows that $\phi(R)$ is contained in the center $Z(A)$ of A.

Conversely, if R and A are arbitrary rings with R commutative and $\phi : R \to A$ is a ring homomorphism such that $\phi(R) \subseteq Z(A)$, then defining a left R-module structure on A by $r \cdot a = \phi(r)a$ it is not hard to check that A is an R-algebra.

In this way, one can see that to make a ring A into an R-algebra is equivalent to giving a homomorphism $\phi : R \to A$ such that $\phi(R) \subseteq Z(A)$.

Example 15.25. Let R be a commutative ring. The polynomial ring $R[x]$ and the power series ring $R[[x]]$ are both R-algebras in an obvious way. Similarly, the polynomial ring $R[x_1, \ldots, x_n]$ in finitely many variables is an R-algebra. A noncommutative example of an R-algebra is the ring $M_n(R)$ of $n \times n$ matrices over R, where R acts by scalar multiplication.

We are especially interested in algebras over fields. If A is an F-algebra for a field F, then by the remark above we can think of this via a homomorphism $\phi : F \to A$ with $\phi(F) \subseteq Z(A)$. Since F is a field, ϕ is injective and we usually identify F with its image $\phi(F)$. Thus an algebra over a field F is just a ring A together with a subring of the center $Z(A)$ which is isomorphic to F, which gives A an F-module structure (i.e. vector space structure) by restriction.

Example 15.26. Let F be a field. The polynomial ring $F[x]$ and the power series ring $F[[x]]$ are both F-algebras in an obvious way. Similarly, the polynomial ring $F[x_1, \ldots, x_n]$ in finitely many variables is an F-algebra. A noncommutative example of an F-algebra is the ring $M_n(F)$ of $n \times n$ matrices over F, where the copy of F in the center is the subring of scalar matrices $\{\lambda I | \lambda \in F\}$.

We do not define the notion of R-algebra when R is not commutative. Remark 15.24 shows why this wouldn't be useful: if we just apply the given definition to an arbitrary ring R, then $\phi(R)$ must lie in the center of A and thus be a commutative ring. If $I = \ker \phi$ then A is an R/I-module where R/I is commutative, and we might as well just think of A as an algebra over the commutative ring R/I.

One of the reasons why algebras are useful is that we can take a tensor product of two algebras and obtain another algebra.

Theorem 15.27. *Let A and B be algebras over a (commutative) ring R. Then $A \otimes_R B$ is again an R-algebra, where it is an R-module via Proposition 15.17 and the product is given by $(a \otimes b)(c \otimes d) = ac \otimes bd$.*

Proof. Since R is commutative, we know that $A \otimes_R B$ is again an R-module, where $r \cdot (a \otimes b) = (ra \otimes b) = (a \otimes rb)$ for $r \in R$, $a \in A$, $b \in B$.

Now for $a \in A, b \in B$, we define a map $\psi_{a,b} : A \otimes_R B \to A \otimes_R B$ by the formula $c \otimes d \mapsto ac \otimes bd$. This exists from the universal property since the function $A \times B \to A \otimes_R B$ given by $(c, d) \mapsto ac \otimes bd$ is R-bilinear. So $\psi_{a,b} \in \mathrm{End}_R(A \otimes_R B)$.

Then define $\widehat{\Psi} : A \times B \to \mathrm{End}_R(A \otimes_R B)$ by $(a, b) \mapsto \psi_{a,b}$. Again this is an R-bilinear map. So we get an R-module homomorphism $\Psi : A \otimes_R B \to \mathrm{End}_R(A \otimes_R B)$ such that $\Psi(a \otimes b) = \psi_{a,b}$.

In other words, we can think of $\Psi(a \otimes b)$ as "left multiplication by $a \otimes b$". This allows us to define a product $*$ on $A \otimes_R B$, where for $x, y \in A \otimes_R B$ we define $x * y = [\Psi(x)](y)$. On pure tensors this product has the formula $(a \otimes b)(c \otimes d) = ac \otimes bd$ as claimed. It is now routine to check that with this product that $A \otimes_R B$ is a ring, and in fact an R-algebra. $\square$

While the product operation in $A \otimes_R B$ is done just my multiplying "coordinate-wise", the tensor product of algebras is a very different operation from the direct product of rings, because the underlying space $A \otimes_R B$ is much different from $A \times B$. The following examples should help illustrate how the tensor product of algebras behaves.

Example 15.28. Let R be a commutative ring. Then the ring $M_n(R)$ consisting of matrices with entries in R is an R-algebra. Let A be another R-algebra; for simplicity assume that the corresponding map $\phi : R \to A$ defined by $\phi(r) = r \cdot 1$ is injective. We claim that $A \otimes_R M_n(R) \cong M_n(A)$ as R-algebras.

We can identify R with its image $\phi(R)$ and thus think of R as a subring of $Z(A)$. There is an R-bilinear map $A \times M_n(R) \to M_n(A)$ given by $(a, B) \mapsto aB$ for $a \in A, B \in M_n(R)$. (a scalar a times a matrix B means multiply every entry of the matrix B by a on the left, or in other words take the matrix product $(aI)B$.) So we get an R-module homomorphism $\psi : A \otimes M_n(R) \to M_n(A)$. Note that for $a \in A$, and any matrix $B \in M_n(R)$, the matrices aI and B commute (this is because $R \subseteq Z(A)$). Thus we have

$$\psi((a \otimes B)(c \otimes D)) = \psi(ac \otimes BD) = (acI)BD = (aI)B(cI)D = \psi(a \otimes B)\psi(c \otimes D).$$

Thus ψ is a homomorphism of R-algebras.

Let e_{ij} be the matrix with 1 in the (i, j)-entry and 0 in all other entries (in $M_n(R)$ or in $M_n(A)$). These n^2 elements are often traditionally called *matrix units* (but they are not units in the ring). An arbitrary matrix $B \in M_n(A)$, where $B_{ij} = a_{ij} \in A$, can be written as $B = \sum_{i,j} a_{ij} e_{ij}$. Now define $\rho : M_n(A) \to A \otimes M_n(R)$ by $\rho(\sum_{i,j} a_{ij} e_{ij}) = \sum_{i,j} a_{ij} \otimes e_{ij}$. It is obvious that $\psi\rho = 1_{M_n(A)}$. Conversely, on pure tensors we have $\rho\psi(a \otimes \sum_{i,j} r_{i,j} e_{i,j}) = \rho(\sum_{i,j} a r_{i,j} e_{i,j}) = \sum_{i,j} a r_{i,j} \otimes e_{i,j} = \sum_{i,j} a \otimes r_{i,j} e_{i,j}$ since $r_{i,j} \in R$. It follows that $\rho\psi = 1_{A \otimes M_n(R)}$. Thus ψ is an isomorphism.

Example 15.29. Let R be a commutative ring. Then $M_n(R) \otimes_R M_m(R) \cong M_{mn}(R)$ as R-algebras.

The previous example actually gives $M_n(R) \otimes_R M_m(R) \cong M_m(M_n(R))$ as R-algebras. We leave the proof that $M_m(M_n(R)) \cong M_{mn}(R)$ as an exercise for the reader (think about multiplication of matrix blocks).

Example 15.30. For any R-algebra A, $A \otimes_R R[x] \cong A[x]$ as R-algebras. This is proved similarly as in Example 15.28 and left as an exercise.

For example, this gives $R[y] \otimes_R R[x] \cong (R[y])[x]$. By definition this is the polynomial ring in two variables $R[y, x]$.

Example 15.31. Tensor product of fields over a common subfield can behave in unexpected ways.

For example, consider $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ as an $\mathbb{R}$-algebra. We know that $\mathbb{C}$ is a vector space over $\mathbb{R}$ with basis $\{1, i\}$ and so by our description of the tensor product over a field, $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is a vector space over $\mathbb{R}$ with basis $\{1 \otimes 1, 1 \otimes i, i \otimes 1, i \otimes i\}$. A pure tensor has the form $(a + bi) \otimes (c + di) = ac(1 \otimes 1) + bc(i \otimes 1) + ad(1 \otimes i) + bd(i \otimes i)$. This is 0 only if $ac = bc = ad = bd = 0$, which happens only if $a = b = 0$ or $c = d = 0$. Thus a pure tensor $(w \otimes z)$ is zero only if $w = 0$ or $z = 0$. In particular, since $(w \otimes z)(x \otimes y) = wx \otimes zy$, a product of nonzero pure tensors is nonzero.

This is a case where thinking about pure tensors only for intuition can lead one astray, however, as $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is not a domain. One may easily check that

$$[(1 \otimes i) - (i \otimes 1)][(1 \otimes i) + (i \otimes 1)] = 0.$$

In fact $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C} \cong \mathbb{C} \times \mathbb{C}$ as $\mathbb{R}$-algebras, as you will be asked to show on the homework.

16. Some homological algebra

In this section we briefly introduce some of the very basic definitions and ideas in homological algebra. This material is most naturally explained in the language of category theory, so we start with an introduction to that subject. We will only be able to scratch the surface here. The reader can find a more detailed treatment in "An Introduction to Homological Algebra" by Rotman, or "An Introduction to Homological Algebra" by Weibel.

16.1. **Categories.** It is useful to introduce here a few definitions from category theory, concentrating on module categories. The homological algebra books mentioned earlier also contain basic introductions to the notions of categories for the reader that would like to learn more.

Definition 16.1. A category is a class $\mathcal{C}$ of objects together with (i) for each pair of objects (X, Y) in $\mathcal{C}$, a set of *morphisms* $\mathrm{Hom}(X, Y)$; (ii) for each object X a *identity morphism* $1_X \in \mathrm{Hom}(X, X)$; and (iii) for each triple of objects (X, Y, Z), a composition rule $\theta_{X,Y,Z} : \mathrm{Hom}(Y, Z) \times \mathrm{Hom}(X, Y) \to \mathrm{Hom}(X, Z)$. For $f \in \mathrm{Hom}(Y, Z)$, $g \in \mathrm{Hom}(X, Y)$, one writes $\theta_{X,Y,Z}(f, g)$ as $f \circ g$.

These are subject to the following axioms: (a) $g \circ 1_X = g$ for any $g \in \mathrm{Hom}(X, Y)$ and $1_X \circ f = f$ for any $f \in \mathrm{Hom}(Z, X)$; and (b) composition is associative, that is $(f \circ g) \circ h = f \circ (g \circ h)$ for $f \in \mathrm{Hom}(Y, Z)$, $g \in \mathrm{Hom}(X, Y)$, $h \in \mathrm{Hom}(W, X)$.

A category is often just referred to by the name $\mathcal{C}$ for its class of objects, with the other data being understood. If we need to emphasize the category a set of morphisms belongs to we write $\mathrm{Hom}_{\mathcal{C}}(X, Y)$ for $\mathrm{Hom}(X, Y)$.

Remark 16.2. We refer in the definition to a class of objects because there are many common and useful categories in which the objects do not form a set. A class is a notion capturing the usual idea of a set (a collection of objects) but which is not assumed to be subject to the rigorous axioms of set theory. We take a naive approach here which ignores set-theoretic subtleties for the most part, as they are not usually relevant for the basic information we want to study.

We are especially interested here in categories of modules.

Example 16.3. Fix a ring R. The category R-Mod has its class of objects all left R-modules. For each $M, N \in R$-Mod, the set of morphisms from M to N is $\mathrm{Hom}_R(M, N)$, the R-module homomorphisms from M to N. Composition of morphisms is usual function composition: If $f \in \mathrm{Hom}_R(M, N)$ and $g \in \mathrm{Hom}_R(N, P)$, then $g \circ f \in \mathrm{Hom}_R(M, P)$. The identity morphism in $\mathrm{Hom}_R(M, M)$ is the identity function 1_M. all axioms are immediate.

Analogously, we have a category Mod-R of all right R-modules, with morphisms being right module homomorphisms.

Example 16.4. If R and S are rings, there is the category (R, S)-Bimod whose objects are (R, S)-bimodules, and where $\mathrm{Hom}(M, N)$ consists of (R, S)-bimodule homomorphisms, that is, maps $f : M \to N$ such that $f(rm) = rf(m)$ and $f(ms) = f(m)s$ for all $m \in M$, $r \in R$, $s \in S$.

In the simplest and most familiar examples of categories, each object of the category is a set with some additional structure, the morphisms consist of certain functions between these sets which "preserve the structure", and composition is given by actual composition of functions. Such a category is called a *concrete category*. Further examples include the category *Set* of sets, with morphisms being functions; *Grp* of Groups, with morphisms being group homomorphisms; *Ab*, the category of abelian groups, with morphisms being group homomorphisms; the category *Rings* of rings, with morphisms being ring homomorphisms; and the category *Top* of topological spaces, with morphisms being continuous maps.

In a general category, however, the objects are not necessarily themselves sets, or even if they are, the morphisms $\mathrm{Hom}(X, Y)$ are not necessarily some subset of the functions from X to Y. So in a general categorical setting one needs to formulate all properties in ways which do not refer to the properties of a function. For example, an isomorphism of modules is usually defined to be a module homomorphism which is injective and surjective, where injective and surjective refer to properties of the underlying function. But instead we could define it to be a module homomorphism for which there is an inverse homomorphism. This gives the natural way to define an isomorphism in a general category:

Definition 16.5. Let $\mathcal{C}$ be a category. A morphism $f \in \mathrm{Hom}(X, Y)$ is called an *isomorphism* if there is a morphism $g \in \mathrm{Hom}(Y, X)$ such that $f \circ g = 1_Y$ and $g \circ f = 1_X$.

We see using this definition that isomorphism reduces to its usual meaning in categories of groups, modules, and rings. (In the category of topological spaces an isomorphism is called a *homeomorphism*.)

Here is an example of a non-concrete category.

Example 16.6. Let M be a fixed monoid. Define a category $\mathcal{C}$ with one object $\{X\}$, and with $\mathrm{Hom}_{\mathcal{C}}(X, X) = M$. For morphisms $g, h \in M$, we define $g \circ h = gh$, using the product operation in M. The identity morphism 1_X is the identity element 1 in M. The category axioms follow from the axioms of M.

Here, X is just an abstract object, not a set. The elements in $\mathrm{Hom}(X, X)$ are elements in M, not functions. "composition" is actually product of monoid elements, which is unrelated to any notion of actual function composition. The example above shows that given a monoid M, we can produce an associated category $\mathcal{C}$ with one object X such that $\mathrm{Hom}(X, X) = M$. Conversely, given any category with one object X, $\mathrm{Hom}(X, X)$ is a monoid. We see in this way that a category with one object is the same as a monoid.

A category $\mathcal{C}$ is called a *groupoid* if every morphism in $\mathcal{C}$ is an isomorphism.

Example 16.7. Let G be a fixed group. Since G is a monoid, we can define a category $\mathcal{C}$ with one object X and $\mathrm{Hom}(X, X) = G$ as in Example 16.6. Because G is a group, we see that $\mathcal{C}$ is a groupoid. Conversely, it is easy to see that given a groupoid with one object X, $\mathrm{Hom}(X, X) = G$ is a group. So a group is the same notion as a groupoid with one object.

In an general category $\mathcal{C}$ with objects X, Y, it does not make sense to ask if $f \in \mathrm{Hom}(X, Y)$ is injective or surjective, since f is not necessarily a function. However, we do have the following natural replacements for these notions in a category.

Definition 16.8. Let $\mathcal{C}$ be a category and $f \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$ a morphism. We say that f is a *monomorphism* if whenever $W \in \mathcal{C}$ and $g, h \in \mathrm{Hom}(W, X)$ with $f \circ g = f \circ h$, then $g = h$. We say that f is an *epimorphism* if whenever $Z \in \mathcal{C}$ and $g, h \in \mathrm{Hom}(Y, Z)$ with $g \circ f = h \circ f$, them $g = h$.

The notions of monomorphism and epimorphism give us a way to talk about properties of morphisms in general categories which are roughly analogous to functions being injective or surjective. In a general category, however, isomorphisms are not necessarily the same as morphisms which

are both monomorphisms and epimorphisms. Even in a concrete category, though, the notions of monomorphism and injection do not necessarily coincide, unlike in the case of the category of sets, and similarly epimorphisms are not necessarily the same as surjections. We give some examples in the following exercises. Fortunately, in the categories of greatest interest to us (categories of modules over a ring) there is no issue.

Exercise 16.9. Let $\mathcal{C}$ be a concrete category. Show that if $f \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$ is an injective function, then it is a monomorphism. Similarly, if f is a surjection, it is an epimorphism.

Exercise 16.10. Let $\mathcal{C}$ be *Set*, *Grp*, R-mod, *Rings*, or *Top*. In each of these concrete categories show that If $X \in \mathcal{C}$ is an object and $x_1, x_2 \in X$, there exists an object $W \in \mathcal{C}$ with an element $w \in W$ and morphisms $g, h \in \mathrm{Hom}_{\mathcal{C}}(W, X)$ such that $g(w) = x_1$ and $h(w) = x_2$. Conclude that in a monomorphism in $\mathcal{C}$ is an injective function, and thus the notions of injection and monomorphism coincide.

Exercise 16.11. Let $\mathcal{C}$ be *Set* or R-mod. In each of these concrete categories show that given $f \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$, if $y \in Y - f(X)$ there is an object $Z \in \mathcal{C}$ and morphisms $g, h \in \mathrm{Hom}_{\mathcal{C}}(Y, Z)$ such that $g(y) \neq h(y)$ while $g(y') = h(y')$ for all $y' \in f(X)$. Conclude that an epimorphism in $\mathcal{C}$ is a surjective function, and thus the notions of surjection and epimorphism coincide.

Exercise 16.12. Let $\mathcal{C}$ be *Top*, the category of topological spaces. Show that f is an epimorphism if and only if $f(X)$ is dense in Y, i.e. every open subset of Y intersects $f(X)$ nontrivially. Give an explicit example of a morphism $f : X \to Y$ which is both a monomorphism and an epimorphism but is not an isomorphism.

Exercise 16.13. Let $\mathcal{C}$ be *Rings*. Suppose that R is a commutative ring and X a multiplicative system, let RX^{-1} be the localization, and let $f : R \to RX^{-1}$ be associated homomorphism of rings. Thinking of $f \in \mathrm{Hom}_{\mathcal{C}}(R, RX^{-1})$, show that f is an epimorphism. Conclude that being an isomorphism in the category $\mathcal{C}$ is not the same as being an isomorphism of rings in the usual sense.

Exercise 16.14. Let M be a monoid, and let $\mathcal{C}$ be a category with one object X and $\mathrm{Hom}(X, X) = M$, where composition in $\mathcal{C}$ is the product in G, as in Example 16.6. What are the monomorphisms and epimorphisms in this category? Is is true that a morphism is an isomorphism if and only if it is both a monomorphism and an epimorphism?

Exercise 16.15. Let S be a set of objects. Show that there is a groupoid whose set of objects is S, and where $\mathrm{Hom}(X, Y)$ consists of exactly one element for all $X, Y \in S$.

16.2. **Functors.** A *functor* is a map from one category to another that preserves the category structure. Essentially, these are the "homomorphisms" between categories.

Definition 16.16. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. A functor $F : \mathcal{C} \to \mathcal{D}$ consists of a choice of object $FX \in \mathcal{D}$ for each $X \in \mathcal{C}$, and for each $X, Y \in \mathcal{C}$ a function $\mathrm{Hom}_{\mathcal{C}}(X, Y) \to \mathrm{Hom}_{\mathcal{D}}(FX, FY)$, written notationally as $f \mapsto F(f)$, such that (i) $F(1_X) = 1_{FX}$ for all $X \in \mathcal{C}$; and (ii) $F(f \circ g) = F(f) \circ F(g)$ for all $X, Y, Z \in \mathcal{C}$ and $f \in \mathrm{Hom}_{\mathcal{C}}(Y, Z)$, $g \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$.

Some easy examples of functors come from trivial relationships between categories. For example, since every Abelian group is a group, there is an *inclusion functor* $F : Ab \to Grp$ which simply takes an abelian group G so $F(G) = G$, thinking of G as an object in the larger category Grp; here if G, H are abelian groups then $\mathrm{Hom}_{Ab}(G, H) = \mathrm{Hom}_{Grp}(G, H)$ and so F also acts as the identity on morphisms. In an opposite direction, we have *forgetful functors* which "forget" some part of the structure. For example, if $\mathcal{C}$ is any concrete category, there is a forgetful functor $F : \mathcal{C} \to Set$ where for any $X \in \mathcal{C}$, $F(X)$ is the underlying set of X, and for $f \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$, $F(f) : X \to Y$ is the underlying map of sets. Similarly, for a ring R, there is a forgetful functor $F : R$ -mod $\to Ab$ which takes an R-module to the underlying abelian group and forgets the R-action. Recall from our earlier study of modules that a $\mathbb{Z}$-module is the essentially the same concept as an Abelian group. In other words, the forgetful functor $\mathbb{Z}$ -mod $\to Ab$ does not actually forget any essential information. We will simply identify the categories $\mathbb{Z}$-mod and Ab below.

More interesting for us will be certain functors from one module category to another.

Example 16.17. Fix a ring R. Let $M \in \text{Mod-}R$ be a right R-module. Then there is a functor $F = M \otimes_R - : R\text{-Mod} \to \mathbb{Z}\text{-Mod}$ given by "tensoring with M on the left". On objects we have $F(N) = M \otimes_R N$; on morphisms, for $f \in \mathrm{Hom}_R(N, P)$ we have $F(f) = 1 \otimes f : M \otimes_R N \to M \otimes_R P$ as defined in Lemma 15.12, which is a homomorphism of abelian groups (i.e. $\mathbb{Z}$-modules). Recall that $[1 \otimes f](m \otimes n) = m \otimes f(n)$. The axioms of a functor are easy to verify.

Similarly, for a fixed $P \in R\text{-Mod}$ there is a functor $G = - \otimes_R P : \text{Mod-}R \to \mathbb{Z}\text{-Mod}$ given by "tensoring with P on the right", where $G(L) = L \otimes_R P$ and for $f \in \mathrm{Hom}_R(L, X)$, $G(f) = f \otimes 1 : L \otimes_R P \to X \otimes_R P$ as in Lemma 15.12.

When the tensor products have additional module structure, so do the functors in Example 16.17. For example, if M is an (S, R)-bimodule, then $M \otimes_R -$ is a functor from R-Mod to S-Mod. Similarly, if P is an (R, T)-bimodule then $- \otimes_R P$ is a functor from Mod-R to Mod-T. And if R is commutative, then as usual we define everything in terms of left modules only, and for $M \in R\text{-Mod}$, both $M \otimes_R -$ and $- \otimes_R M$ are functors from R-Mod to R-Mod.

Example 16.18. Fix a ring R and let $M \in R$-Mod. Then there is a functor $F = \mathrm{Hom}_R(M, -) : R\text{-Mod} \to \mathbb{Z}\text{-Mod}$, where $F(N) = \mathrm{Hom}_R(M, N)$ and for $f \in \mathrm{Hom}_R(N, P)$, we have $F(f) : \mathrm{Hom}_R(M, N) \to \mathrm{Hom}(M, P)$ given by $g \mapsto f \circ g$. The axioms of a functor are easy to check.

There is also a kind of functor given by $\mathrm{Hom}_R(-, N)$ for a fixed N, but we need another definition first. A *contravariant functor* from a category $\mathcal{C}$ to a category $\mathcal{D}$ is defined similarly as a usual functor, except that it reverses the order of composition. So such a functor is given by a choice of object $FX \in \mathcal{D}$ for each $X \in \mathcal{C}$, and a function on sets of morphisms $F(-) : \mathrm{Hom}_{\mathcal{C}}(X, Y) \to \mathrm{Hom}_{\mathcal{D}}(FY, FX)$, such that $F(f \circ g) = F(g) \circ F(f)$ for $f \in \mathrm{Hom}(Y, Z)$, $g \in \mathrm{Hom}(X, Y)$.

16.3. **Exact sequences and flatness.**

Definition 16.19. Fix a ring R and let M, N, P be left R-modules. Consider a sequence of R-module homomorphisms

$$M \xrightarrow{f} N \xrightarrow{g} P.$$

The sequence is called *exact* at N if $\ker g = f(M)$. More generally, it is called a *complex* if $f \circ g = 0$, i.e. $f(M) \subseteq \ker g$. In this case, the *homology group* at N is the R-module $H = (\ker g)/f(M)$. So a complex is exact at N if and only if $H = 0$.

Note that $0 \longrightarrow N \xrightarrow{g} P$ is automatically a complex, and it is exact at N if and only if $0 = \ker(g)$, that is, g is an injective homomorphism. Dually, $M \xrightarrow{f} N \longrightarrow 0$ is a complex which is exact at N if and only if $f(M) = N$, that is, f is surjective.

A longer sequence of modules and homomorphisms is called an exact sequence if it is exact at every spot that has an arrow both entering and leaving. In particular we have

Definition 16.20. A sequence of R-modules and maps

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

which is exact (that is, exact at M, N, and P) is called a *short exact sequence*.

From the comments above, we see that explicitly the definition of short exact sequence is equivalent to f being injective, g being surjective, and $f(M) = \ker(g)$. Given such a short exact sequence, we also say that N is an *extension* of P by M. This is because by the first isomorphism theorem applied to g, $N/f(M) = N/(\ker g) \cong P$. Also since f is injective, $f(M) \cong M$. So N is built out of the submodule $f(M) \cong M$ and the factor module $N/f(M) \cong P$; the short exact sequence tells us precisely how M and P are put together to form N. In this point of view, a short exact sequence is just a convenient way to represent the information of an extension of two modules.

Below we will sometimes write $\text{im}(f)$ instead of $f(M)$ for the image of a map $f : M \to N$, so that image and kernel are given similar notations.

If you have taken a course in algebraic topology, you will recognize the definitions above. The development of Algebraic Topology had a lot of influence on algebra. A whole field of homological algebra developed from this which abstracts definitions and techniques derived from topology. These homological techniques have had and continue to have great importance in the study of algebra. In this short section we will just be able to scratch the surface. We recommend the book of Weibel, ("An Introduction to Homological Algebra"), or of Rotman (also called "An Introduction to Homological Algebra") if you would like to learn more about this interesting subject.

We would like to consider the following question. To what extent does the operation of tensoring with a module preserve exact sequences? More precisely, suppose that

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

is a short exact sequence where M, N, and P are left R-modules and f and g are R-module homomorphisms. If Q is a right R-module, then we can apply the operation $Q \otimes_R -$ to the entire sequence using the functoriality result of Lemma 15.12, to obtain a sequence

$$0 \longrightarrow Q \otimes_R M \xrightarrow{1\otimes f} Q \otimes_R N \xrightarrow{1\otimes g} Q \otimes_R P \longrightarrow 0$$

in which the maps are homomorphisms of abelian groups. If Q is an (S, R)-bimodule the maps are even left S-module homomorphisms. This sequence is a complex but it turns out to not always be a short exact sequence, as we will see next; the problem is at the $Q \otimes_R M$ spot.

Theorem 16.21. *Let $0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$ be a short exact sequence of left R-modules. Let Q be a right R-module. Then*

$$Q \otimes_R M \xrightarrow{1\otimes f} Q \otimes_R N \xrightarrow{1\otimes g} Q \otimes_R P \longrightarrow 0$$

is exact, i.e. exact at the $Q \otimes_R N$ and $Q \otimes_R P$ spots.

This result can be described by saying that the operation of tensoring with a module is *right exact*. It preserves the exactness at the right two terms of the sequence only.

Proof. Let $q \otimes p$ be a pure tensor in $Q \otimes_R P$. Since g is surjective, there is $n \in N$ such that $g(n) = p$. Then $[1 \otimes g](q \otimes n) = q \otimes p$. Thus all pure tensors are in the image of $1 \otimes g$. Since all elements in $Q \otimes_R P$ are sums of pure tensors and $1 \otimes g$ is a homomorphism of abelian groups, $1 \otimes g$ is surjective. Thus we have exactness at the $Q \otimes_R P$ spot.

Now since $g \circ f = 0$, note that $(1 \otimes g) \circ (1 \otimes f) = 0$. Thus our supposed right exact sequence is at least a complex, in other words we have $\operatorname{im}(1 \otimes f) \subseteq \ker(1 \otimes g)$. Let $L = (Q \otimes_R N)/\operatorname{im}(1 \otimes f)$, which is again a left R-module. Note that because we have $\operatorname{im}(1 \otimes f) \subseteq \ker(1 \otimes g)$, we can define a map $\overline{1 \otimes g} : L \to Q \otimes_R P$ by $\overline{1 \otimes g}(x + \operatorname{im}(1 \otimes f)) = [1 \otimes g](x)$ for $x \in Q \otimes_R N$. Since $1 \otimes g$ is surjective, so is $\overline{1 \otimes g}$.

Now we claim that there is a homomorphism of abelian groups

$$\psi : Q \otimes P \to L = (Q \otimes_R N)/\operatorname{im}(1 \otimes f)$$

given by the formula $\psi(q \otimes p) = q \otimes n + \operatorname{im}(1 \otimes f)$, where $n \in N$ is any element such that $g(n) = p$. To see that this formula does not depend on the choice of n, note that if $g(n) = g(n') = p$ then $n - n' \in \ker(g) = \operatorname{im}(f)$ and so $n' - n = f(m)$ for some $m \in M$. Then $q \otimes n - q \otimes n' = q \otimes (n - n') = q \otimes f(m) \in \operatorname{im}(1 \otimes f)$. thus $q \otimes n + \operatorname{im}(1 \otimes f) = q \otimes n' + \operatorname{im}(1 \otimes f)$. The existence of the map ψ is then proved in the usual way by first defining an R-balanced map and applying the universal property.

Now for $(q \otimes n) + \operatorname{im}(1 \otimes f) \in L$ applying $\overline{1 \otimes g}$ gives $q \otimes g(n)$ and then applying ψ gives $q \otimes n' + \operatorname{im}(1 \otimes f)$ for some n' such that $g(n) = g(n')$. As we just saw, $q \otimes n + \operatorname{im}(1 \otimes f) = q \otimes n' + \operatorname{im}(1 \otimes f)$. Thus $\psi \circ \overline{1 \otimes g} = 1_L$.

In particular $\overline{1 \otimes g}$ must be injective. So $0 = \ker(\overline{1 \otimes g}) = \ker(1 \otimes g)/\operatorname{im}(1 \otimes f)$ and thus $\ker(1 \otimes g) = \operatorname{im}(1 \otimes f)$. Thus we have exactness at the $Q \otimes_R N$ spot and we are done. □

It is easy to give an example showing that the result above is the best we can do; the operation of tensoring with a module does not preserve exactness at the left in general.

Example 16.22. Let R be an integral domain which is not a field. Let x be a nonunit in R. Consider the module homomorphism $f : R \to R$ given by $f(a) = ax$. Since R is a domain, f is injective. Thus f is the left map in a short exact sequence $0 \to R \xrightarrow{f} R \xrightarrow{\pi} R/xR \to 0$ where π is the natural surjection, and $R/xR \neq 0$ since x is not a unit.

Now let us tensor this short exact sequence with R/xR, obtaining

$$0 \to (R/xR) \otimes_R R \xrightarrow{1 \otimes f} (R/xR) \otimes_R R \xrightarrow{1 \otimes \pi} (R/xR) \otimes_R (R/xR) \to 0.$$

We know the resulting sequence will be exact at the right by Theorem 16.21. Let us see that it is not exact at the left; in other words $1 \otimes f$ is not injective.

We have seen that there is an isomorphism $(R/xR) \otimes_R R \to R/xR$ given by $(a + xR) \otimes b \mapsto ab + xR$; see Example 15.15. In particular, $(R/xR) \otimes_R R \neq 0$. On the other hand,

$$[1 \otimes f]((a + xR) \otimes b) = (a + xR) \otimes xb = x(a + xR) \otimes b = (xa + xR) \otimes b = (0 + xR) \otimes b = 0$$

which implies that $1 \otimes f = 0$. In particular, $1 \otimes f$ is not injective.

The following definition is made to focus on those modules that do not have the problem of failure of left exactness as in the previous example.

Definition 16.23. Let R be any ring. A right R-module Q is called *flat* (over R) if for all short exact sequences of left R-modules $0 \to M \xrightarrow{f} N \xrightarrow{g} P \to 0$, the sequence obtained by applying $Q \otimes_R -$ to this short exact sequence,

$$0 \to Q \otimes_R M \xrightarrow{1 \otimes f} Q \otimes_R N \xrightarrow{1 \otimes g} Q \otimes_R N \to 0,$$

is again short exact.

Because tensoring with Q always preserves exactness on the right, by Theorem 16.21, it is easy to see that Q is a flat right R-module if and only if for all injective homomorphisms of left R-modules $f : M \to N$, then $1 \otimes f : Q \otimes_R M \to Q \otimes_R N$ is still injective.

In the terminology of category theory, $Q \otimes_R -$ is what is known as a *functor*: it is an operation that sends every left R-module M to an abelian group $Q \otimes_R M$, and comes along with an action on homomorphisms which sends a module homomorphism $f : M \to N$ to an abelian group homomorphism $1 \otimes f : Q \otimes_R M \to Q \otimes_R N$. A functor which when applied to a short exact sequence returns another short exact sequence is called an *exact* functor. So Q is flat if $Q \otimes_R -$ is an exact functor, by definition. We don't have time here to go more into the details of category theory, but may occasionally adopt this terminology.

Of course there is nothing special about the side on which we have chosen to make this definition: a left R-module L is called flat if $- \otimes_R L$ is an exact functor when applied to short exact sequences of right R-modules.

Example 16.24. Let R be an integral domain which is not a field. If x is not a unit in R, then R/xR is an R-module which is not flat, by Example 16.22. In fact, it is possible to show that a flat module over an integral domain must be torsionfree.

Now let us give examples of modules that are flat by showing that free modules are always flat. We leave the following result to the reader; it is a good exercise in applying the universal property of the tensor product.

Lemma 16.25. *Let $\{M_\alpha | \alpha \in I\}$ be an indexed family of right R-modules. For any left R-module N, we have an isomorphism of abelian groups*

$$\Phi : \Big(\bigoplus_{\alpha \in I} M_\alpha\Big) \otimes_R N \to \bigoplus_{\alpha \in I} \big(M_\alpha \otimes_R N\big),$$

where $\Phi((m_\alpha)_{\alpha\in I} \otimes n) = (m_\alpha \otimes n)_{\alpha\in I}$.

Of course, there is also a symmetric result showing that a direct sum in the second coordinate pulls out of a tensor product.

Remark 16.26. The corresponding statement for products is not true in general: there is a natural homomorphism $\Psi : (\prod_{\alpha\in I} M_\alpha) \otimes_R N \to \prod_{\alpha\in I}(M_\alpha \otimes_R N)$ given by the same formula, but it need not be an isomorphism when the index set I is infinite.

Proposition 16.27. *Let F be a free right R-module. Then F is flat.*

The same result holds for left R-modules, as one would expect, by a symmetric proof.

Proof. We know that $F \cong \bigoplus_{\alpha\in I} R$ for some index set I. It is easy to see that flatness is an invariant of a module up to isomorphism, so we just need to prove that $\bigoplus_{\alpha\in I} R$ is flat. Now for any right module M we have

$$(\bigoplus_{\alpha\in I} R) \otimes_R M \cong \bigoplus_{\alpha\in I}(R \otimes_R M) \cong \bigoplus_{\alpha\in I} M$$

using Lemma 16.25 and the fact that $R \otimes_R M \cong M$, as in Example 15.15.

If $f : M \to N$ is an injective homomorphism of right R-modules, using the isomorphisms above it is straightforward to check that $1 \otimes f : (\bigoplus_{\alpha\in I} R) \otimes_R M \to (\bigoplus_{\alpha\in I} R) \otimes_R N$ can be identified with the homomorphism $\bigoplus f : \bigoplus_{\alpha\in I} M \to \bigoplus_{\alpha\in I} N$ which simply applies f in every coordinate. But since f is injective, this homomorphism is clearly also injective. □

Another important class of flat modules is given by localization.

Example 16.28. Let R be any commutative ring and let X be a multiplicative system in R. Then one can define the localization RX^{-1} as in Section 9.3. The localization comes along with a ring homomorphism $\phi : R \to RX^{-1}$ defined by $\phi(r) = r/1$, and this makes RX^{-1} into an R-module by restriction of scalars. Then RX^{-1} is a flat R-module, as you will check in an exercise on the homework. For example, if R is an integral domain with field of fractions K, then K is a flat R-module.

The case where a module Q is flat over R and so $Q \otimes_R -$ preserves short exact sequences is very nice, but it is not the usual situation. In general one only has right exactness of the tensor product. In a further study of homological algebra one develops the theory of Tor functors which can be used to understand more precisely any failure of exactness of the tensor product on the left. We refer the reader to either of the books on homological algebra already mentioned for more details.

16.4. **Projective modules.** In this optional section we give another class of modules which are flat, namely projective modules.

Lemma 16.29. *Suppose that F is a free right R-module and that $F \cong P \oplus Q$ for right R-submodules P and Q. Then P is a flat module.*

Proof. Let $f : M \to N$ be an injective homomorphism of left R-modules. We have seen that free modules are flat and hence

$$(P \oplus Q) \otimes_R M \xrightarrow{1 \otimes f} (P \oplus Q) \otimes_R N$$

is also injective.

Now we have an injection map $i : P \to P \otimes Q = F$ given by $i(p) = (p, 0)$, and a projection map $\pi : F = P \oplus Q \to P$ given by $\pi(p, q) = p$, where $\pi \circ i = 1_P$. Moreover, there is a commutative diagram

$$\begin{array}{ccc} P \otimes_R M & \xrightarrow{1 \otimes f} & P \otimes_R N \\ \downarrow{\scriptstyle i \otimes 1} & & \downarrow{\scriptstyle i \otimes 1} \\ (P \oplus Q) \otimes_R M & \xrightarrow{1 \otimes f} & (P \oplus Q) \otimes_R N \end{array}$$

Now if $x \in P \otimes_R M$ satisfies $[1 \otimes f](x) = 0$, then $[1 \otimes f] \circ [i \otimes 1](x) = 0$; by the flatness of $F = P \oplus Q$, we have $[i \otimes 1](x) = 0$. But since $\pi \circ i = 1_P$, $[\pi \otimes 1] \circ [i \otimes 1] = 1_{P \otimes_R M}$. It follows that $x = 0$. Hence $1 \otimes f : P \otimes_R M \to P \otimes_R N$ is injective and P is flat. □

We have proved that direct summands of free modules are flat. These modules have a name and another interesting description.

Definition 16.30. Let P be a right R-module. P is a *projective* module if given any surjective homomorphism of right modules $g : M \to N$ and a homomorphism of right modules $f : P \to N$, there is a homorphism $h : P \to M$ such that $g \circ h = f$.

This property can be respresented by the following commutative diagram:

$$\begin{array}{ccccc} & & P & & \\ & {\scriptstyle \exists h} \swarrow & \downarrow{\scriptstyle f} & & \\ M & \xrightarrow{g} & N & \longrightarrow & 0 \end{array}$$

The additional arrow to the right of N pointing to 0 is to remind one that in this property g is assumed to be surjective, that is, that the bottom row is exact at N. The property satisfied by P is not a universal property the way those are usually understood, because the map h is only assumed to exist, and need not be (in fact almost never is) unique.

One of the important properties of projective modules is that any surjection onto a projective module is split.

Lemma 16.31. *Let P be a projective right R-module. Suppose $g : N \to P$ is a surjective homomorphism of right R-modules. Then g is a split surjection.*

Recall that for g to be a split surjection means that there is a homomorphism $h : P \to N$ such that $g \circ h = 1_P$, and that by Lemma 12.49 this has as a consequence that N is an internal direct sum $N = \ker(g) \oplus \mathrm{im}(h) \cong \ker(g) \oplus P$.

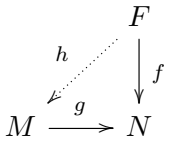
Proof. Taking $f = 1_P : P \to P$ and $g : N \to P$ as given, the existence of $h : N \to P$ such that $g \circ h = 1_P$ is immediate from the definition of projective module. □

Let us now relate this definition of projective module to the modules that appeared in Corollary 16.29.

Theorem 16.32. *A right R-module P is projective if and only if there exists a right R-module Q such that $P \oplus Q$ is a free right R-module. In particular, free modules are projective.*

As usual, in the definition of projective and this theorem, there exist left-sided versions which are stated and proved in the analogous way. We have focused on right modules here only because our primary definition of flatness was for right modules, so we have stayed on that side for consistency.

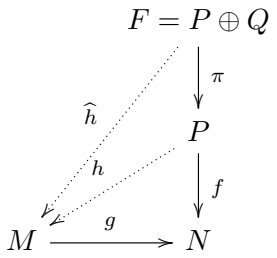
Proof. First we claim that a free right R-module F is projective. Suppose that $f : F \to N$ and $g : M \to N$ are given right module homomorphisms with g surjective. Let $\{x_\alpha | \alpha \in I\}$ be a basis for F as a free right R-module. For each $\alpha \in I$ we can choose $m_\alpha \in M$ such that $g(m_\alpha) = f(x_\alpha)$, since g is surjective. We need to find h completing the diagram



By the universal property of a free module, there exists a unique homomorphism $h : F \to M$ such that $h(x_\alpha) = m_\alpha$ for all α. Then $f(x_\alpha) = [h \circ g](x_\alpha)$ for all α. This shows that the diagram commutes for the basis elements x_α. Since the basis generates F as an R-module and all maps are R-module homomorphisms, $h \circ g = f$ and the diagram commutes. Thus F is projective as claimed.

Now we need to extend this result to show that a direct summand of a free module F is projective. Suppose that we have an internal direct sum $F = P \oplus Q$. Let $i : P \to P \oplus Q = F$ and $\pi : F =$

$P \oplus Q \to P$ be the injection and projection maps associated to the first coordinate of the direct sum, as in the proof of Corollary 16.29. Given f and g as above we have a diagram



Now there exists a homomorphism $\widehat{h}$ making the outer triangle commute, i.e. $g \circ \widehat{h} = f \circ \pi$, since F is free. Define $h = \widehat{h} \circ i$. Then

$$g \circ h = g \circ \widehat{h} \circ i = f \circ \pi \circ i = f$$

because $\pi \circ i = 1_P$. This proves that P is projective.

For the converse, we need to prove that a projective module is necessarily a direct summand of a free module. Let P be projective and pick any generating set $\{p_\alpha | \alpha \in I\}$ whatsoever for P as a right R-module. By the universal property of a free module, there is a unique homomorphism $g : F = \bigoplus_{\alpha \in I} R \to P$ such that $g(e_\alpha) = p_\alpha$ for all α, where $\{e_\alpha\}$ is the standard basis of F. Since the $\{p_\alpha\}$ are a generating set for P, g is surjective. Now because P is projective, Lemma 16.31 shows that g is split. Thus $F \cong (\ker g) \oplus P$ by lemma 12.49. Taking $Q = \ker g$ we see that $P \oplus Q$ is free as required. □

Corollary 16.33. *Projective modules are flat.*

Proof. This follows from Theorem 16.32 and Corollary 16.29. □

Note that Corollary 12.50, which showed that a surjective homomorphism onto a free module splits, can now just be seen as a special case of Lemma 16.31 since free modules are projective. We needed the fact that surjections onto free modules are split for the theory of modules over a PID, but did not wish to introduce projective modules at that point.

Let us note that not all projective modules are free.

Example 16.34. Let $R = M_2(F)$ where F is a field. Then $R = I \oplus J$ as right R-modules, for right ideals $I = \left\{ \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} \middle| a, b \in F \right\}$ and $J = \left\{ \begin{pmatrix} 0 & 0 \\ c & d \end{pmatrix} \middle| c, d \in F \right\}$. Thus each of I and J is a summand of a free module of rank 1 and so I and J are projective right R-modules. Note however that every right R-module is also an F-vector space, since R is an F-algebra. Since $\dim_F R = 4$,

any free R-module has F-dimension which a multiple of 4 (or infinite). Since $\dim_F I = 2$, I cannot be free. Of course J is also not free for the same reason, but in fact it is easy to see that $I \cong J$ as right R-modules.

Example 16.35. We claim that $\mathbb{Q}$ is a flat $\mathbb{Z}$-module which is not projective. $\mathbb{Q}$ is flat since it is a localization of $\mathbb{Z}$, as in Example 16.28.

A $\mathbb{Z}$-module M is *divisible* if for all $x \in M$ and positive integer $n > 0$, there exists $y \in M$ such that $ny = x$. It is obvious that $\mathbb{Q}$ is a divisible $\mathbb{Z}$-module.

On the other hand, for a free $\mathbb{Z}$-module $F \cong \bigoplus \mathbb{Z}$, it is easy to see that no nonzero element x of F can satisfy the divisibility property above for all $n > 0$; thus F has no nonzero divisible submodules. In particular, $\mathbb{Q}$ cannot be isomorphic to a summand of a free module and hence it is not projective.

Example 16.36. If R is a PID, then it is true that all projective R-modules are free. This is easy to prove for finitely generated projective modules using the classification theorem. In general it follows from the fact that submodules of free modules over R are free (which is true in general but we only proved for finitely generated modules).

16.5. **Exactness of Hom.**

17. Field basics

17.1. **Field extensions.** Recall that a *field* F is a commutative ring such that every nonzero element is a unit. In our study of fields we will sometimes want to refer to auxiliary commutative rings which are not themselves fields; but we will not have any use for noncommutative rings in this section. Every ring should be assumed to be commutative unless told otherwise.

There is a short list of fields that arise naturally as fields of numbers appearing throughout mathematics, and which we have already encountered: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}/p\mathbb{Z}$ for a prime p. In this section we will write the field of integers mod p as $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, to emphasize that we are considering it as a field and not just a group. Actually, how "natural" the field of real numbers $\mathbb{R}$ is may be debatable, since to define it precisely involves a nontrivial limiting process of some kind. But we will take the existence of $\mathbb{R}$ and its basic properties as a given. It is certainly natural in the sense of its many applications to the physical sciences.

Besides the basic examples above we saw in the ring theory section two ring theoretic constructions which lead to many new examples of fields. Both will be of fundamental importance in our study of fields.

First, if R is any commutative ring whatsoever, and I is a maximal ideal of R, then the factor ring R/I is a field. This gives a way of producing potentially a number of different fields from a given commutative ring. For example, taking $R = \mathbb{Z}$ then the maximal ideals are those of the form $p\mathbb{Z}$ for primes p, and we recover all of the fields $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ through this construction. Even if we have a general commutative ring R about which we know nothing, we have seen as an application of Zorn's lemma that R must have some maximal ideal I; thus R has at least one factor ring R/I that is a field.

Second, if R is an integral domain, then we defined the field of fractions K of R to be the set of formal fractions $\{r/s | r \in R, 0 \neq s \in R\}$ under a natural equivalence relation that $r/s = t/u$ if $ru = st$, with addition and multiplication defined as usual for fractions. We always identify R with the subring $\{r/1 | r \in R\}$ of K, so that $R \subseteq K$. Thus for any integral domain we can always produce at least one field by taking the field of fractions. Of course $\mathbb{Q}$ can be produced in this way by taking the field of fractions of $\mathbb{Z}$. Recall also that if F is a field, the field of fractions of the polynomial ring $F[x]$ is called the *field of rational functions* $F(x)$. Its elements consists of formal ratios of polynomials.

Usually field theory does not study a field in isolation but rather its relation to other fields. The basic object of our study of fields will be the following.

Definition 17.1. A *field extension* or *extension of fields* is an inclusion of fields $F \subseteq K$; that is, K is a field and F is a subring of K which is also a field, which we also call a *subfield*.

K is always a left K-module by multiplication, so it is a left F-module by restriction. In other words, K is a vector space over F. As such it has a dimension and we define the *degree* of the field extension $F \subseteq K$ to be the number $[K : F] = \dim_F K$.

It is also common to use the notation K/F for a field extension $F \subseteq K$; K/F is read as "K over F" and is meant to emphasize that we are considering K in relation to the subfield F it lies over. The notation K/F is one whole unit and does not indicate any kind of quotient construction. The field F is also called the *base field*.

Example 17.2. $\mathbb{R} \subseteq \mathbb{C}$ is a field extension and $[\mathbb{C} : \mathbb{R}] = 2$ because by construction an $\mathbb{R}$-basis for $\mathbb{C}$ is given by $\{1, i\}$.

Example 17.3. In our study of ring theory we introduced for any square-free integer D the ring $\mathbb{Q}(\sqrt{D}) = \{a + b\sqrt{D} | a, b \in \mathbb{Q}\}$, as a subring of $\mathbb{C}$, and proved this ring is a field. Since D is squarefree $\sqrt{D} \notin \mathbb{Q}$, so $\{1, \sqrt{D}\}$ form a basis for $\mathbb{Q}(\sqrt{D})$ over $\mathbb{Q}$ and $[\mathbb{Q}(\sqrt{D}) : \mathbb{Q}] = 2$.

Actually both field extensions $\mathbb{R} \subseteq \mathbb{C}$ and $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})$ arise from a general ring-theoretic construction which will be of crucial importance from now on. There is nothing here we haven't already seen in our study of rings, but we remind the reader of the details since they are so important.

Lemma 17.4. *Let F be a field. Let $f \in F[x]$ be an irreducible polynomial.*

(1) *$F[x]/(f) = K$ is again a field.*

(2) *The function $\theta : F \to K = F[x]/(f)$ defined by $\theta(a) = a + (f)$ is an injective ring homomorphism.*

(3) *Identifying F with $\theta(F)$ we have a field extension $F \subseteq K$. As such, $[K : F] = \deg f$.*

Proof. (1) Since $F[x]$ is a PID, we know that an irreducible element generates a maximal ideal (f) by Lemma 10.27.

(2) This function is obviously a ring homomorphism by the definition of multiplication in $F[x]/(f)$. Since f is irreducible, by definition it is not a unit and so $\deg f \geq 1$. Since f has minimal degree among nonzero elements of (f), the ideal (f) contains no nonzero constant polynomials. Thus $\ker \theta = 0$ and θ is injective.

(3) Since θ is injective it gives an isomorphism from F to $\theta(F)$, so we can take this to be an identification, under which F now consists of the elements in $F[x]/(f)$ whose coset representatives are constant polynomials. Let $\deg f = n$. If $h \in F[x]$, then $h = qf + r$ under polyomial long division, where $\deg r < \deg f$. Hence $h + (f) = r + (f)$ with $\deg r \leq n-1$. Moreover, if $r' + (f) = r + (f)$, where $\deg r \leq n-1$ as well, then $r - r' \in (f)$ with $\deg r - r' \leq n-1$; since the smallest degree of nonzero elements of (f) is n, $r - r' = 0$ and $r = r'$. We see that every element of K is of the form $r + (f)$ for a *unique* polynomial r of degree at most $n-1$. It follows that $\{1 + (f), x + (f), \ldots, x^{n-1} + (f)\}$ is a basis for $F[x]/(f)$ as an F-vector space, and so $[K : F] = \deg f = n$. □

Whenever we are in the situation of the previous lemma, it is convenient to always identify F with its image under θ and consider K as an extension of F.

Example 17.5. Let us revisit Example 17.2. Usually the construction of $\mathbb{C}$ from $\mathbb{R}$ is done by defining $\mathbb{C}$ to a be an $\mathbb{R}$ vector space with basis $\{1, i\}$ made into a ring by defining the multiplication explicitly using $i^2 = 1$; technically, one needs to check this multiplication is associative and that the resulting ring is a field.

Lemma 17.4 gives a more abstract but cleaner way to go: It is immediate that $x^2 + 1$ is an irreducible polynomial in $\mathbb{R}[x]$, since it has no real roots. Thus $F = \mathbb{R}[x]/(x^2 + 1)$ is a field extension of $\mathbb{R}$ such that $[F : \mathbb{R}] = 2$. It is immediate that F contains an element $x + (x^2 + 1)$

whose square is equal to $-1 + (x^2 + 1)$, so abstractly we have found a field extension of $\mathbb{R}$ in which -1 has a square root. Of course, $F \cong \mathbb{C}$ as we already saw in Example 10.11; we have just defined the usual complex numbers in a different way.

Just as for groups and modules, the notion of a subfield generated by a subset will be important.

Definition 17.6. Let $F \subseteq K$ be a field extension. For any subset X of K, the *subfield of K generated by X over F* is the intersection of all subfields of K which contain both F and X. It is written as $F(X)$. When $X = \{\alpha_1, \ldots, \alpha_n\}$ is finite we write this as $F(\alpha_1, \ldots, \alpha_n)$. A field generated by one element over F, $F(\alpha)$, is called a *simple extension* of E.

It is easy to see that an arbitrary intersection of subfields is again a subfield, so that $F(X)$ is the unique smallest subfield of K which contains both F and X. Note that this notation only makes sense when working inside some larger field K that contains the elements in X. The field K is understood and not part of the notation.

Example 17.7. Let D be a squarefree integer. We have already defined a subfield of $\mathbb{C}$ called $\mathbb{Q}(\sqrt{D})$ in Example 17.3, where taking $\sqrt{D}$ to be either of the square roots of D in $\mathbb{C}$, we have $\mathbb{Q}(\sqrt{D}) = \mathbb{Q} + \mathbb{Q}\sqrt{D} \subseteq \mathbb{C}$. We have seen this is a field, and it obviously contains $\mathbb{Q}$ and $\sqrt{D}$. Conversely, any subfield of $\mathbb{C}$ which contains $\mathbb{Q}$ and $\sqrt{D}$ would contain $\mathbb{Q} + \mathbb{Q}\sqrt{D}$ just because it is closed under addition and multiplication. Thus $\mathbb{Q}(\sqrt{D})$ is indeed the subfield of $\mathbb{C}$ generated by $\sqrt{D}$ over $\mathbb{Q}$, so this notation we have used for it agrees with our new notation for subfield generation.

In our study of module theory over a commutative ring R, we saw that R-modules generated by a single element (cyclic modules) are the modules of the form R/I for an ideal I. Now we will see that field extensions generated by a single element, or simple extensions as we have named them, also have a very rigid structure.

Theorem 17.8. *Let $F \subseteq K$ be a field extension. Let $\alpha \in K$. There is a canonical homomorphism of rings*

$$\phi : F[x] \to F(\alpha)$$
$$f(x) \mapsto f(\alpha)$$

and moreover exactly one of the following two cases occurs:

(i) $\ker \phi = (f)$ *for a unique, monic irreducible polynomial $f \in F[x]$. Moreover, ϕ is surjective, $F(\alpha) \cong F[x]/(f)$ as fields, and $[F(\alpha) : F] = \deg f < \infty$.*

(ii) $\ker\phi = 0$. *In this case ϕ extends to an isomorphism $F(x) \to F(\alpha)$, where $F(x)$ is the field of rational functions in one variable over F. Moreover $[F(\alpha) : F] = \infty$.*

Proof. The map ϕ is simply the evaluation at $\alpha \in K$, where $f(x) = \sum_{i=0}^{n} a_i x^i \in F[x]$ maps to $f(\alpha) = \sum_{i=0}^{n} a_i \alpha^i$ (with $\alpha^0 = 1$). We saw in our study of ring theory that such a map is a homomorphism of rings $F[x] \to K$. But every element of the image is of the form $\sum_{i=0}^{n} a_i \alpha^i$ with $a_i \in F$, which is clearly contained in any subfield of K which contains F and α. So $\mathrm{im}(\phi) \subseteq F(\alpha)$ and we can think of ϕ as a map $F[x] \to F(\alpha)$.

Assume that $\ker\phi \neq 0$, so we are in case (i). Since $F[x]$ is a PID, $\ker\phi = (f)$ for some unique monic polynomial $f \in F[x]$. Now by the 1st isomorphism theorem for rings, $F[x]/(f) \cong \mathrm{im}(\phi)$. Since $\mathrm{im}(\phi)$ is a subring of K, it is a domain, so $F[x]/(f)$ is a domain. Thus (f) is a prime ideal, but in a PID, nonzero prime ideals are maximal. So $F[x]/(f)$ is a field and hence so is $\mathrm{im}(\phi)$. Now $\mathrm{im}(\phi)$ is a subfield of K which clearly contains F (the image of the constant polynomials) and α (the image of x). Thus $F(\alpha) \subseteq \mathrm{im}(\phi)$ by the definition of subfield generation. On the other hand, we already saw that $\mathrm{im}(\phi) \subseteq F(\alpha)$. Thus $\mathrm{im}(\phi) = F(\alpha)$ and we have an isomorphism $F[x]/(f) \to F(\alpha)$. In a PID maximal ideals are generated by irreducible polynomials, so f is irreducible. By Lemma 17.4, $[F(\alpha) : F] = \dim_F(F(\alpha)) = \dim_F(F[x]/(f)) = \deg f$ since ϕ is also an F-vector space isomorphism.

Otherwise, $\ker\phi = 0$ and we are in case (ii). Now the homomorphism of rings $\phi : F[x] \to F(\alpha)$ has the property that for every nonzero element $f \in F[x]$, $0 \neq \phi(f)$ is a unit, since $F(\alpha)$ is a field. By the universal property of the localization in Theorem 9.22, ϕ extends uniquely to a homomorphism $\widetilde{\phi} : F(x) \to F(\alpha)$, where $F(x)$ is the field of fractions of $F[x]$, in other words the localization of $F[x]$ at the set $X = F[x] - \{0\}$. The formula for $\widetilde{\phi}$ is $\widetilde{\phi}(f/g) = \phi(f)(\phi(g))^{-1} = f(\alpha)g(\alpha)^{-1}$ for all $0 \neq g, f \in F[x]$. Since $F(x)$ is a field, $\widetilde{\phi}$ must be injective and so $F(x) \cong \mathrm{im}(\widetilde{\phi})$. Thus $\mathrm{im}(\widetilde{\phi})$ is a subfield of K containing F and α and since $F(\alpha)$ is the unique smallest such, $F(\alpha) \subseteq \mathrm{im}(\widetilde{\phi})$. On the other hand, the formula for $\widetilde{\phi}$ above shows that $\mathrm{im}(\widetilde{\phi}) \subseteq F(\alpha)$. Thus $\mathrm{im}(\widetilde{\phi}) = F(\alpha)$ and in this case we have $F(x) \cong F(\alpha)$ as fields. Since $\dim_F F[x] = \infty$ already, certainly $\dim_F F(x) = \infty$. Thus $[F(\alpha) : F] = \dim_F F(\alpha) = \dim_F F(x) = \infty$. □

Definition 17.9. Let $F \subseteq K$ be a field extension, and let $\alpha \in K$. If case (i) occurs in Theorem 17.8 we say that α is *algebraic over F*; the monic irreducible polynomial f with $(f) = \ker(\phi)$ is called the *minimal polynomial* of α over F and is written as $\mathrm{minpoly}_F(\alpha)$. Otherwise, case (ii) occurs and we say that α is *transcendental over F*.

Note that α is algebraic over F precisely when there is some nonzero polynomial $g \in F[x]$ such that $g(\alpha) = 0$; the set of all such polynomials is then $\ker \phi = (f)$ in the notation of Theorem 17.8. Thus the minimal polynomial of α over F is the monic polynomial f of minimal possible degree such that $f(\alpha) = 0$. We also know that f is irreducible, and so clearly f is the unique monic irreducible polynomial such that $f(\alpha) = 0$ as well.

Example 17.10. Consider the extension $\mathbb{Q} \subseteq \mathbb{R}$. $f(x) = x^3 - 2$ is irreducible in $\mathbb{Q}[x]$ since it has no root in $\mathbb{Q}$, or by the Eisenstein criterion. Since the real cube root $\alpha = \sqrt[3]{2} \in \mathbb{R}$ satisfies $f(\alpha) = 0$, $f = \text{minpoly}_{\mathbb{Q}}(\alpha)$. Thus α is algebraic over $\mathbb{Q}$ and $\mathbb{Q}(\alpha) \cong \mathbb{Q}[x]/(x^3 - 2)$. In particular $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$.

On the other hand, it is known that π is a transcendental number over $\mathbb{Q}$, though this is not particularly easy to prove. Thus $\mathbb{Q}(\pi) \cong \mathbb{Q}(x)$. In general *transcendence theory* refers to the study of which particular real or complex numbers are transcendental over $\mathbb{Q}$. This subject tends to involve sensitive results in analysis. As an example of the difficulty of this theory, the number e has also been proved to be transcendental over $\mathbb{Q}$, but it is unknown whether $e + \pi$ and $e\pi$ are transcendental.

We now know how to understand simple extensions $F(\alpha)$ of a field F, in terms of the properties of α. But in fact, an extension generated by a finite set of elements can be expressed as a series of simple extensions.

Lemma 17.11. *Let $F \subseteq K$ be a field extension. For any elements $\alpha_1, \ldots, \alpha_n \in K$, we have $F(\alpha_1, \ldots, \alpha_n) = F(\alpha_1)(\alpha_2)\ldots(\alpha_n)$.*

Proof. Let us argue the case $n = 2$; the general case follows easily by induction. Note that when we write $F(\alpha_1)(\alpha_2)$, we mean $[F(\alpha_1)](\alpha_2)$; that is we are now applying the definition of generation to the field extension $F(\alpha_1) \subseteq K$ and the element α_2. In other words, $F(\alpha_1)(\alpha_2)$ is the unique smallest subfield of K which contains the field $F(\alpha_1)$ and the element α_2. On the other hand, the field $F(\alpha_1, \alpha_2)$ is the unique smallest subfield of K which contains F, α_1, and α_2.

The field $F(\alpha_1)(\alpha_2)$ contains $F(\alpha_1)$ and α_2, so it contains F, α_1, and α_2. Thus $F(\alpha_1, \alpha_2) \subseteq F(\alpha_1)(\alpha_2)$. Conversely, the field $F(\alpha_1, \alpha_2)$ contains F and α_1, so it contains $F(\alpha_1)$, the unique smallest such subfield. Thus $F(\alpha_1, \alpha_2)$ contains $F(\alpha_1)$ and α_2 and so $F(\alpha_1)(\alpha_2) \subseteq F(\alpha_1, \alpha_2)$. □

Note that when we write $F(\alpha_1, \ldots, \alpha_n)$, the order in which we write the elements α_i is immaterial; we are taking the subfield generated by them as a set. On the other hand, when we treat this

as $F(\alpha_1)(\alpha_2)\dots(\alpha_n)$ we have chosen a specific order, though the end result must be the same regardless. If we are doing calculations, one order might be easier to handle than another.

Example 17.12. Let $\mathbb{Q} \subseteq \mathbb{C}$ and consider $\mathbb{Q}(\sqrt{2}, i)$. Think of this as $\mathbb{Q}(\sqrt{2})(i)$. We know that $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$, since $\text{minpoly}_{\mathbb{Q}}(\sqrt{2}) = x^2 - 2$ as $x^2 - 2$ does not have a root in $\mathbb{Q}$. Now i is a root of $x^2 + 1 \in \mathbb{Q}[x] \subseteq \mathbb{Q}(\sqrt{2})[x]$. So $\text{minpoly}_{\mathbb{Q}(\sqrt{2})}(i)$ has degree at most 2. If it has degree 1, that means that $i \in \mathbb{Q}(\sqrt{2})$, but this is false since $\mathbb{Q}(\sqrt{2})$ consists of real numbers. Hence $\text{minpoly}_{\mathbb{Q}(\sqrt{2})}(i) = x^2 + 1$. We conclude that $\mathbb{Q}(\sqrt{2}, i) \cong \mathbb{Q}(\sqrt{2})[x]/(x^2 + 1)$ and $[\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}(\sqrt{2})] = 2$. By results we will see in the next section this means that $[\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}(\sqrt{2})][\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 4$.

It is also possible to analyze this extension by considering it as $\mathbb{Q}(i)(\sqrt{2})$ instead, but the other order is a bit easier because we can use the trick of considering real versus complex numbers. In this order we would have to determine whether or not $\sqrt{2} \in \mathbb{Q}(i)$; this is not difficult but it is a bit more work.

17.2. **Algebraic extensions.**

Definition 17.13. Let $F \subseteq K$ be a field extension. The extension is called *algebraic* if every $\alpha \in K$ is algebraic over F.

We will explore the properties of algebraic extensions in this section. Note that in an algebraic extension, for every $\alpha \in K$ we have $[F(\alpha) : F]$ is finite, equal to the degree of $\text{minpoly}_F(\alpha)$. But these degrees can vary widely as we range over elements α, and $[K : F]$ could still be infinite overall, as we will see.

The key lemma for calculating the degree of an algebraic extension is the following.

Lemma 17.14. *Let $E \subseteq F \subseteq K$ be fields.*

(1) *If either $[K : F]$ or $[F : E]$ is infinite, then so is $[K : E]$.*

(2) *If $[K : F]$ and $[F : E]$ are finite, then so is $[K : E]$ and in fact $[K : E] = [K : F][F : E]$.*

Proof. (1) If $[F : E] = \infty$, in other words $\dim_E F = \infty$, then since the E-vector space K contains the E-subspace F, $\dim_E K = \infty$ also. If instead $[K : F] = \infty$, then a basis of K as an F-space is certainly also still a linearly independent set over E. Since any linearly independent set can be extended to a basis, $\dim_E K = \infty$.

(2) The proof will show more than the statement; we will see how to construct, given a basis of K as an F-space and a basis of F as an E-space, an explicit basis of K as an E-space.

Thus let $\{\alpha_1, \dots, \alpha_m\}$ be an E-basis for F and $\{\beta_1, \dots, \beta_n\}$ an F-basis for K. We claim that $S = \{\alpha_i \beta_j | 1 \leq i \leq m, 1 \leq j \leq n\}$ is an E-basis for K.

First, if $\gamma \in K$, then $\gamma = \sum_{j=1}^{n} a_j \beta_j$ for $a_j \in F$ since $\{\beta_j\}$ spans K as an F-space. But then each $a_j = \sum_{i=1}^{m} b_{ij}\alpha_i$ for some $b_{ij} \in E$ since the α_i space F as an E-space. We conclude that $\gamma = \sum_{i,j} b_{ij}\alpha_i\beta_j$ and so the set S spans K over E.

Next, suppose that $\sum_{i,j} b_{ij}\alpha_i\beta_j = 0$ for some $b_{ij} \in E$. Then $\sum_j (\sum_i b_{ij}\alpha_i)\beta_j = 0$ with $\sum_i b_{ij}\alpha_i \in F$, since $E \subseteq F$. Since the β_j are independent over F we get $\sum_i b_{ij}\alpha_i = 0$ for all i. But then since the α_i are independent over E we get $b_{ij} = 0$ for all j, for all i. This shows that S is independent over E. We have shown that S is a basis for K over E as claimed.

Now $[K : E] = |S| = nm = [K : F][F : E]$. □

Corollary 17.15. *If $E \subseteq F \subseteq K$ are fields with $[K : E] < \infty$, then $[F : E]$ and $[K : F]$ divide $[K : E]$.*

Corollary 17.16. *If $E \subseteq K$ is a field extension with prime degree $[K : E] = p$, then for any field F with $E \subseteq F \subseteq K$, either $E = F$ or $E = K$.*

Lemma 17.14 and its corollaries are reminiscent of Lagrange's theorem in finite group theory. This is not an accident; the Fundamental Theorem of Galois Theory which we prove later gives strong connections between fields and groups.

Many important examples involve considering extensions generated over $\mathbb{Q}$ inside the field extension $\mathbb{Q} \subseteq \mathbb{C}$. From now on when we write $\mathbb{Q}(\alpha_1, \ldots, \alpha_n)$ for certain complex numbers α_i, it should be assumed that we are taking this extension inside $\mathbb{C}$ even if that is not explicitly mentioned.

Lemma 17.14 is very useful for doing calculations of degrees of extensions. Here is one example.

Example 17.17. Consider $K = \mathbb{Q}(\sqrt[3]{2}, \sqrt{3})$ where $\sqrt[3]{2}$ is the real cube root of 2. We claim that $[K : \mathbb{Q}] = 6$. We know that $x^3 - 2$ and $x^2 - 3$ are irreducible over $\mathbb{Q}$, so $\text{minpoly}_{\mathbb{Q}}(\sqrt[3]{2}) = x^3 - 2$ and $\text{minpoly}_{\mathbb{Q}}(\sqrt{3}) = x^2 - 3$. Thus $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ and $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$. Since of course $x^3 - 2 \in \mathbb{Q}(\sqrt{3})[x]$ as well, $\text{minpoly}_{\mathbb{Q}(\sqrt{3})}(\sqrt[3]{2})$ has degree at most 3 and

$$[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}(\sqrt{3})] = [\mathbb{Q}(\sqrt{3})(\sqrt[3]{2}) : \mathbb{Q}(\sqrt{3})] \leq 3.$$

By Lemma 17.14, $[K : \mathbb{Q}] \leq 6$. On the other hand, since $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{3}) \subseteq K$ we have $[K : \mathbb{Q}]$ is divisible by $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$ (using Lemma 17.14 again), and since $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2}) \subseteq K$ it also follows that $[K : \mathbb{Q}]$ is divisible by $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$. The only possibility is $[K : \mathbb{Q}] = 6$.

As a consequence, we conclude that $[K : \mathbb{Q}(\sqrt{3})] = 3$. Since $K = \mathbb{Q}(\sqrt{3})(\sqrt[3]{2})$, this means that $\sqrt[3]{2}$ must have minimal polynomial $x^3 - 2$ over $\mathbb{Q}(\sqrt{3})$. It is not obvious that $x^3 - 2$ is irreducible over $\mathbb{Q}(\sqrt{3})$, or equivalently that it has no roots in this field, and this would be more awkward to prove directly. Similarly, we conclude that the minimal polynomial of $\sqrt{3}$ over $\mathbb{Q}(\sqrt[3]{2})$ is still $x^2 - 3$.

As well as being useful for calculations, Lemma 17.14 leads to a number of interesting general results about algebraic elements and extensions.

First we see that a finite degree extension is always algebraic.

Lemma 17.18. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$. Then K/F is an algebraic extension.*

Proof. Let $\alpha \in K$. We have $F \subseteq F(\alpha) \subseteq K$. Because $[K : F] < \infty$, we have $[F(\alpha) : F] < \infty$ by Lemma 17.14. But then we know that α is algebraic over F by Theorem 17.8. Since $\alpha \in K$ was arbitrary, K/F is algebraic. □

Proposition 17.19. *Let $F \subseteq K$ be a field extension and suppose that $\alpha, \beta \in K$ are both algebraic over F. Then $\alpha - \beta$, $\alpha\beta$, and α^{-1} (if $\alpha \neq 0$) are also algebraic over F.*

Proof. By Theorem 17.8, since α is algebraic over F we have $[F(\alpha) : F] < \infty$. Since β is algebraic over F, it is certainly algebraic over $F(\alpha)$; if $f(\alpha) = 0$ with $0 \neq f \in F[x]$ then just consider $f \in F(\alpha)[x]$. Thus $[F(\alpha)(\beta) : F(\alpha)] < \infty$. By Lemma 17.14, $[F(\alpha, \beta) : F] < \infty$. Thus $F(\alpha, \beta)/F$ is an algebraic extension, by Lemma 17.18.

Now notice that since $F(\alpha, \beta)$ is a subfield of K containing α and β, it certainly also contains $\alpha - \beta$, $\alpha\beta$, and α^{-1} (if $\alpha \neq 0$). Hence these elements are all algebraic over F. □

Corollary 17.20. *Let $F \subseteq K$ be a field extension. Define $L = \{\alpha \in K | \alpha$ is algebraic over $F\}$. Then L is a subfield of K containing F and L/F is algebraic.*

Proof. Proposition 17.19 shows that L is closed under difference, product, and inverses, which implies that L is a subfield of K. It is obvious that L/F is algebraic by definition. □

Definition 17.21. The subset

$$\overline{\mathbb{Q}} = \{\alpha \in \mathbb{C} | \alpha \text{ is algebraic over } \mathbb{Q}\}$$

is called the *field of algebraic numbers*. It is also called the *algebraic closure* of $\mathbb{Q}$.

Note that $\overline{\mathbb{Q}}$ really is a subfield of $\mathbb{C}$ by Corollary 17.20. For a fixed prime number p, and each $n \geq 2$, note that $x^n - p$ is irreducible over $\mathbb{Q}$ by the Eisenstein criterion. Thus $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[n]{p}) \subseteq \overline{\mathbb{Q}}$ with $[\mathbb{Q}(\sqrt[n]{p}) : \mathbb{Q}] = n$. It follows that $[\overline{\mathbb{Q}} : \mathbb{Q}] = \infty$. Thus $\overline{\mathbb{Q}}/\mathbb{Q}$ is an example of an infinite degree algebraic extension.

Proposition 17.19 showed that the algebraic numbers over $\mathbb{Q}$ are closed under the field operations. The proof is abstract, however, and does not show how one can find a polynomial that has a given difference, product, or inverse of algebraic numbers as a root.

Example 17.22. Since $\sqrt{2}$ and $\sqrt{3}$ are both algebraic over $\mathbb{Q}$, with minimal polynomials x^2-2 and x^2-3 respectively, we know that $\alpha = \sqrt{2}+\sqrt{3}$ is algebraic over $\mathbb{Q}$. Here is a way to find a polynomial in $\mathbb{Q}$ with α as a root (this method is rather special to the case of a sum of two square roots, though).

Note that $\alpha^2 = (\sqrt{2}+\sqrt{3})^2 = 2+2\sqrt{6}+3 = 5+2\sqrt{6}$. Thus $2\sqrt{6} = \alpha^2-5$. Squaring both sides, $24 = (\alpha^2-5)^2$. It follows that $f(x) = (x^2-5)^2 - 24 = x^4 - 10x^2 + 1$ has α as a root. In fact one may show that $f = \operatorname{minpoly}_{\mathbb{Q}}(\alpha)$.

The following definition is similar to what we have defined for generation in other contexts like groups and modules.

Definition 17.23. Let $F \subseteq K$ be a field extension. We say that the extension K/F is *finitely generated* if $K = F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ for some $\alpha_1, \alpha_2, \ldots, \alpha_n \in K$.

Lemma 17.24. *Let $F \subseteq K$ be a field extension. Then $[K:F] < \infty$ if and only if K/F is finitely generated and algebraic.*

Proof. If $[K:F] < \infty$ then we have seen that K/F is algebraic in Lemma 17.18. It is also easy to see that K/F is finitely generated; for example, if $\alpha_1, \ldots, \alpha_n$ is an F-basis of K then certainly $K = F(\alpha_1, \ldots, \alpha_n)$ (though likely fewer than n elements suffice).

Conversely, suppose that $K = F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ for some elements $\alpha_i \in K$, and that K/F is algebraic. Thus each α_i is algebraic over F. Define $d_i = [F(\alpha_1, \ldots, \alpha_i) : F(\alpha_1, \ldots, \alpha_{i-1})]$ for each $1 \leq i \leq n$. Since $F(\alpha_1 \ldots, \alpha_i) = F(\alpha_1, \ldots, \alpha_{i-1})(\alpha_i)$, we see that $d_i = \deg \operatorname{minpoly}_{F(\alpha_1,\ldots,\alpha_{i-1})}(\alpha_i)$. If $e_i = \deg \operatorname{minpoly}_F(\alpha_i)$, then $d_i \leq e_i$ for all i since any polynomial in $F[x]$ with α_i as a root is also a polynomial in $F(\alpha_1, \ldots, \alpha_{i-1})[x]$. Now

$$\begin{aligned}[K:F] &= [F(\alpha_1, \ldots, \alpha_n) : F(\alpha_1, \ldots, \alpha_{n-1})][F(\alpha_1, \ldots, \alpha_{n-1}) : F(\alpha_1, \ldots, \alpha_{n-2}] \ldots [F(\alpha_1) : F] \\ &= d_n d_{n-1} \ldots d_1 \leq e_n e_{n-1} \ldots e_1 < \infty,\end{aligned}$$

by repeated use of Lemma 17.14. In particular $[K:F] < \infty$. □

Examining the proof of the lemma, we immediately have the following consquence.

Corollary 17.25. *If $K = F(\alpha_1, \ldots, \alpha_n)$ where each α_i is algebraic over F with $e_i = \deg \operatorname{minpoly}_F(\alpha_i) = [F(\alpha_i) : F]$, then $[K:F] \leq e_1 e_2 \ldots e_n$.*

Example 17.17, where $K = \mathbb{Q}(\sqrt[3]{2}, \sqrt{3})$, is an example where the upper bound given by the corollary is actually acheived. In general it is just an upper bound, as can be seen from silly

examples like $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{6})$, where $\sqrt{6} \in \mathbb{Q}(\sqrt{2}, \sqrt{3})$ already, so in fact $[K : \mathbb{Q}] \leq 4$, while applying the corollary with the 3 generators blindly gives $[K : \mathbb{Q}] \leq 8$.

We may now show that an algebraic extension of an algebraic extension is algebraic.

Theorem 17.26. *Let $E \subseteq F \subseteq K$ where both F/E and K/F are algebraic extensions. Then K/E is algebraic.*

Proof. Let $\alpha \in K$. We know that α is algebraic over F. Let $f = \text{minpoly}_F(\alpha) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in F[x]$. Now each coefficient $a_i \in F$ is algebraic over E. By Lemma 17.24, $[E(a_0, \ldots, a_{n-1}) : E] < \infty$. Now note that $f \in E(a_0, \ldots, a_{n-1})[x]$. This means that α is algebraic over the field $E(a_0, \ldots, a_{n-1})$. Hence $[E(a_0, \ldots, a_{n-1}, \alpha) : E(a_0, \ldots, a_{n-1})] < \infty$. Now applying Lemma 17.14 gives $[E(a_0, \ldots, a_{n-1}, \alpha) : E] < \infty$. In particular, since α belongs to the field $E(a_0, \ldots, a_{n-1}, \alpha)$, α is algebraic over E by Lemma 17.18. □

Example 17.27. Suppose that $F \subseteq K$ is a field extension. We say that F is *algebraically closed in K* if given any chain of subfields $F \subseteq L \subseteq K$ with L/F algebraic, then $L = F$. In other words, no subfield of K properly containing F is algebraic over F.

Now given an arbitrary extension $E \subseteq K$, we can define $F = \{\alpha \in K | \alpha \text{ is algebraic over } E\}$. As we have seen, $E \subseteq F \subseteq K$ with F a subfield of K algebraic over E. Now we can see that F is algebraic closed in K. For if $F \subseteq L \subseteq K$ with L/F algebraic, then L/E is algebraic by Theorem 17.26. Thus every element of L is algebraic over E, which means $L \subseteq F$ and hence $L = F$.

In particular, considering $\mathbb{Q} \subseteq \overline{\mathbb{Q}} \subseteq \mathbb{C}$, we see that $\overline{\mathbb{Q}}$ is algebraically closed in $\mathbb{C}$.

17.3. **Splitting fields.** Let F be a field and let $f(x) \in F[x]$ be an irreducible polynomials. Recall that if $\deg f \geq 2$, then f has no roots in F (Corollary 11.13). Of course, if we have a field extension $F \subseteq K$, then f might well have a root in K.

It is natural to wonder if we just start with a field F and have no prior knowledge about any extension fields, does there always exist a field extension $F \subseteq K$ such that the irreducible polynomial $f \in F[x]$ has a root in K? The next lemma answers this question. While the proof of the lemma is easy, the idea behind it is rather subtle.

Lemma 17.28. *Let $f \in F[x]$ be irreducible. Then $K = F[x]/(f)$ is a field, and identifying F with a subfield of K as usual, then f has a root α in K.*

Proof. We already know that K is a field and we saw that the natural map $\theta : F \to F[x]/(f)$ is an injective homomorphism, allowing us to identify F with the cosets of constant polynomials in $K = F[x]/(f)$. This was Lemma 17.4.

But now observe that for $\alpha = x + (f) \in K$, evaluating $f = \sum_{i=0}^{n} a_i x^i$ at α gives

$$\sum_{i=0}^{n}(a_i + (f))(x + (f))^i = \sum_{i=0}^{n}(a_i x^i + (f)) = (\sum_{i=0}^{n} a_i x^i) + (f) = f + (f) = 0 + (f) = 0.$$

Thus f has a root in K as we wished. □

Recall that a polynomial $f \in F[x]$ *splits* (over F) if f factors as a product of degree 1 irreducibles. Since every degree 1 irreducible is of the form $(x - a)$ up to associates, this is the same as saying that we can write $f = c(x - \alpha) \dots (x - \alpha_n)$ for some $c, \alpha_1, \dots, \alpha_n \in F$. A field F is *algebraically closed* if every polynomial in $F[x]$ splits over F, or equivalently if every irreducible polynomial in $F[x]$ is of degree 1. We will freely use in examples that $\mathbb{C}$ is algebraically closed, though we don't prove this until much later.

Definition 17.29. Let $F \subseteq K$ be a field extension and let $f \in F[x]$. We say that K is a *splitting field for f over F* if

(i) $f = c(x - \alpha_1) \dots (x - \alpha_n)$ with $c, \alpha_1, \dots, \alpha_n \in K$, i.e. f splits over K.

(ii) $K = F(\alpha_1, \dots, \alpha_n)$.

Roughly, this definition is saying that a splitting field K is a larger field in which f splits, but where K is no larger than necessary for this to happen. Certainly any field over which f splits must contain the field generated over F by the roots of f.

We have seen that we can always find a larger field in which any given irreducible polynomial has a root. We can now extend this to see that any polynomial has a splitting field.

Lemma 17.30. *Let F be a field and let $f \in F[x]$. Then there exists a field extension $F \subseteq K$ such that K is a splitting field for f over F.*

Proof. First we prove that there exists a field extension $F \subseteq L$ such that f splits in $L[x]$. We induct on $\deg f$. If f already splits in $F[x]$, take $L = F$. Otherwise, a factorization of f into a product of irreducibles in $F[x]$ contains at least one irreducible factor, say g, with $\deg g \geq 2$. By Lemma 17.28, there is an extension $F \subseteq L'$ such that g has a root α_1 in L'. Then by the factor theorem, in $L'[x]$ the polynomial f factors as $f = (x - \alpha_1)f'$ for $f' \in L'[x]$. Since $\deg f' < \deg f$, by the induction hypothesis there is an extension $L' \subseteq L$ such that f' splits in $L[x]$, say $f' = c(x - \alpha_2) \dots (x - \alpha_n)$. Then $f = c(x - \alpha_1)(x - \alpha_2) \dots (x - \alpha_n)$ in $L[x]$, so f splits in $L[x]$.

Finally, define $K = F(\alpha_1, \ldots, \alpha_n) \subseteq L$. It is clear that K is a splitting field for f over F. $\square$

As we saw in the proof above, to find a splitting field K for a polynomial $f \in F[x]$, it suffices to find a field extension $F \subseteq L$ such that f splits in $L[x]$ with $f = c(x - \alpha_1) \ldots (x - \alpha_n)$ and then let $K = F(\alpha_1, \ldots, \alpha_n) \subseteq L$. In particular, since we know any polynomial in $\mathbb{C}[x]$ splits, if $F \subseteq \mathbb{C}$ we can find a splitting field for $f \in F[x]$ by finding the roots of f in $\mathbb{C}$ and adjoining them to F inside $\mathbb{C}$.

Example 17.31. Let $f = x^n - 1 \in \mathbb{Q}[x]$ for some $n \geq 1$. We know that $\mathbb{C}$ has n distinct nth roots of 1, namely $\alpha_j = e^{2j\pi i/n}$ for $1 \leq j \leq n$. Thus $x^n - 1 = (x - \alpha_1) \ldots (x - \alpha_n)$ since both are monic and have the same distinct n roots in $\mathbb{C}$. So $K = \mathbb{Q}(\alpha_1, \ldots, \alpha_n)$ is a splitting field for $x^n - 1$ over $\mathbb{Q}$. Now note that setting $\zeta = \alpha_1 = e^{2\pi i/n}$, we have $\alpha_i = \zeta^i$ for all i. It follows that $K = \mathbb{Q}(\zeta, \zeta^2, \ldots, \zeta^n) = \mathbb{Q}(\zeta)$ and so K is a simple extension of $\mathbb{Q}$. It is called the *nth cyclotomic field*.

Consider the special case where $n = p$ is prime. In this case $x^p - 1 = (x - 1)g(x)$ where $g(x) = x^{p-1} + \cdots + x + 1$ was shown to be irreducible over $\mathbb{Q}$ in Example 11.20. Clearly $\zeta = e^{2\pi i/p}$ is a root of g and therefore $g = \text{minpoly}_{\mathbb{Q}}(\zeta)$. Hence $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$ in this case.

We will calculate the degree $[\mathbb{Q}(\zeta) : \mathbb{Q}]$ for arbitrary n later.

Example 17.32. Generalizing the previous example, let us consider the splitting field of $f = x^n - a \in \mathbb{Q}[x]$ over $\mathbb{Q}$ for some $n \geq 1$ and $a \neq 0$. Let α be any nth root of a in $\mathbb{C}$. As in the previous example, let $\zeta = e^{2\pi i/n}$ so that $\{1, \zeta, \ldots, \zeta^{n-1}\}$ is the set of distinct complex nth roots of 1. Then the set $\{\alpha, \alpha\zeta, \alpha\zeta^2, \ldots, \alpha\zeta^{n-1}\}$ consists of n distinct complex numbers and they are all roots of $x^n - a$; thus $x^n - a = (x - \alpha)(x - \alpha\zeta) \ldots (x - \alpha\zeta^{n-1})$ in $\mathbb{C}[x]$. Now a splitting field for f over $\mathbb{Q}$ can be constructed inside $\mathbb{C}$ as $\mathbb{Q}(\alpha, \alpha\zeta, \ldots \alpha\zeta^{n-1}) = \mathbb{Q}(\alpha, \zeta)$.

Let us consider the special case where $f = x^p - q$ for prime numbers p, q (not necessarily distinct). The number q has a unique positive pth root $\alpha = \sqrt[p]{q} \in \mathbb{R}$. By the Eisenstein criterion applied to the prime q, f is irreducible over $\mathbb{Q}$. Thus $f = \text{minpoly}_{\mathbb{Q}}(\alpha)$ and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = p$. We have seen that $\text{minpoly}_{\mathbb{Q}}(\zeta) = x^{p-1} + \cdots + x + 1$ in this case, so $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$. By Corollary 17.25, $d = [\mathbb{Q}(\alpha, \zeta) : \mathbb{Q}] \leq p(p - 1)$. On the other hand, $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$ and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = p$ are both divisors of d by Lemma 17.14. Since $\gcd(p, p - 1) = 1$, we see that this forces $d = p(p - 1)$. It also means that $[\mathbb{Q}(\alpha, \zeta) : \mathbb{Q}(\alpha)] = p - 1$ and thus $\deg \text{minpoly}_{\mathbb{Q}(\alpha)}(\zeta) = p - 1$; this forces $\text{minpoly}_{\mathbb{Q}(\alpha)}(\zeta) = x^{p-1} + \cdots + x + 1$. It would be rather awkward to prove that this polynomial is irreducible over $\mathbb{Q}(\alpha)$ directly. Similarly, $[\mathbb{Q}(\alpha, \zeta) : \mathbb{Q}(\zeta)] = p$ and $\text{minpoly}_{\mathbb{Q}(\zeta)}(\alpha) = x^p - q$.

It turns out that any two splitting fields of a polynomial $f \in F[x]$ are isomorphic as fields. This is perhaps not too surprising, since by definition, in some sense the splitting field is a "smallest" field extension over which f splits, and that ought to be determined by f. However, along the way we will actually show something stronger: if f is irreducible in $F[x]$, then given two splitting fields K and K', we can choose an isomorphism between them which sends a specified root of f in K to any root of f in K' we please. Applying this to a single splitting field will give us a way to produce automorphisms of a field which move the roots of f around. This will lay the foundation for our study of Galois theory later.

Before attacking the general case of splitting fields, we first example what happens when we adjoin to a field F two possibly different roots of the same irreducible polynomial $f \in F[x]$. For technical reasons, it is necessary to state this result in a generality that at first seems rather awkward: instead of fixing a base field, we work over two different base fields with an isomorphism between them. It doesn't really make the proof harder, and its utility will be seen in the proof of the proposition to follow.

Lemma 17.33. *Let $\phi : F \to F'$ be an isomorphism of fields. This induces an isomorphism of polynomial rings $\phi : F[x] \to F'[x]$ we give the same name, by applying ϕ to the coefficients. Let $f \in F[x]$ be an irreducible polynomial (over F) and let $f' = \phi(f) \in F'[x]$. Suppose that $F \subseteq K$ and $F' \subseteq K'$ are field extensions such that $\alpha \in K$ is a root of f and $\alpha' \in K'$ is a root of f'. Then considering $F \subseteq F(\alpha) \subseteq K$ and $F' \subseteq F'(\alpha') \subseteq K'$, there is an isomorphism $\theta : F(\alpha) \to F'(\alpha')$ such that $\theta(\alpha) = \alpha'$ and $\theta|_F = \phi$.*

Proof. Since f is irreducible over F, $f = \text{minpoly}_F(\alpha)$. Since $\phi : F[x] \to F'[x]$ is an isomorphism of rings, $f' = \phi(f)$ is irreducible over F', and so $f' = \text{minpoly}_{F'}(\alpha')$. Now by our theorem on the structure of a simple extension, there are isomorphisms $\sigma_1, \sigma_2, \sigma_3$ forming a chain

$$F(\alpha) \xrightarrow{\sigma_1} F[x]/(f) \xrightarrow{\sigma_2} F'[x]/(f') \xrightarrow{\sigma_3} F'(\alpha')$$

Where $\sigma_1^{-1} : F[x]/(f) \to F(\alpha)$, $\sigma_3 : F'[x]/(f') \to F'(\alpha')$ are the isomorphisms coming from Theorem 17.8(i), and σ_2 is induced by the isomorphism $\phi : F[x] \to F'[x]$ and the fact that $\phi(f) = f'$. Now take $\theta = \sigma_3\sigma_2\sigma_1$. Then

$$\theta(\alpha) = [\sigma_3\sigma_2\sigma_1](\alpha) = [\sigma_3\sigma_2](x + (f)) = \sigma_3(x + (f')) = \alpha',$$

and $\theta|_F = \phi$ since $\sigma_2|_F = \phi$ while $\sigma_1|_F = 1_F$ and $\sigma_3|_{F'} = 1_{F'}$. □

Corollary 17.34. *Let $F \subseteq K$ be a field extension, and let $f \in F[x]$ be irreducible over F. Suppose that $\alpha_1, \alpha_2 \in K$ are roots of f. Then there is an isomorphism $\sigma : F(\alpha_1) \to F(\alpha_2)$ such that $\sigma(\alpha_1) = \alpha_2$ and $\sigma|_F = 1_F$.*

Proof. This is immediate by applying the lemma to $F = F'$, $\phi = 1_F$, $K' = K$. □

We see thus that *adding two roots of the same irreducible polynomial give isomorphic simple extensions*. All roots of an irreducible polynomial in a larger field are "equal" in this sense.

Example 17.35. Let $\zeta = e^{2\pi i/3}$ be a primitive 3rd root of 1. Consider the splitting field K of $f(x) = x^3 - 2$ over $\mathbb{Q}$. As we saw in Example 17.32, f is irreducible over $\mathbb{Q}$. If $\alpha = \sqrt[3]{2}$ is the positive real cube root of 2, then the roots of f in $\mathbb{C}$ are $\{\sqrt[3]{2}, \sqrt[3]{2}\zeta, \sqrt[3]{2}\zeta^2\}$. By Corollary 17.34, there is an isomorphism $\phi : \mathbb{Q}(\sqrt[3]{2}) \to \mathbb{Q}(\sqrt[3]{2}\zeta)$ fixing $\mathbb{Q}$ and sending $\sqrt[3]{2}$ to $\sqrt[3]{2}\zeta$. The two fields $\mathbb{Q}(\sqrt[3]{2})$ and $\mathbb{Q}(\sqrt[3]{2}\zeta)$ are certainly different, as one of these fields is contained in $\mathbb{R}$ and the other isn't. But as abstract fields they are isomorphic and so must have the same structure.

Now we are ready to prove the uniqueness up to isomorphism of splitting fields.

Proposition 17.36. *Let $\phi : F \to F'$ be an isomorphism of fields, inducing the isomorphism of rings $\phi : F[x] \to F'[x]$. Suppose K is a splitting field of $f \in F[x]$ over F, and K' is a splitting field of $\phi(f) \in F'[x]$ over F'.*

(1) *There is an isomorphism $\sigma : K \to K'$ such that $\sigma|_F = \phi$.*
(2) *If $g \in F[x]$ is any irreducible factor of f in $F[x]$, then for any $\alpha \in K$ which is a root of g, and any $\alpha' \in K'$ which is a root of $\phi(g)$, we can choose a σ in (1) with $\sigma(\alpha) = \alpha'$.*

Proof. Suppose that we have proved (1) and that g is an irreducible factor of degree 1 in part (2). It is no loss of generality to assume that g is monic, so $g = x - \alpha \in F[x]$. Then $\phi(g) = x - \phi(\alpha) \in F'[x]$. So the only root of g is α and the only root of $\phi(g)$ is $\phi(\alpha)$, and since $\sigma|_F = \phi$ we certainly have $\sigma(\alpha) = \phi(\alpha)$. So part (2) is automatic for irreducibles of degree 1.

The proof in general is by induction on degree f. If f splits over F already, then $K = F$. This also means that f' splits over F', so $K' = F'$; thus we take $\sigma = \phi$ in part (1). In this case all irreducible factors of f have degree 1 so (2) is also clear by the remark above. In particular, if $\deg f \leq 1$ then f splits, so the base case holds.

Now we assume that f has some irreducible factor $g \in F[x]$ with $\deg g \geq 2$, and that the result is true for all polynomials of degree smaller than f. Let $g' = \phi(g)$. Then g' is irreducible in $F'[x]$. Since f and hence g splits in K, we can pick a root $\alpha \in K$ of g. Similarly we pick a root $\alpha' \in K'$ of g'.

By Lemma 17.33, there is an isomorphism $\theta : F(\alpha) \to F'(\alpha')$ such that $\theta(\alpha) = \alpha'$ and $\theta|_F = \phi$. Now the key is to treat θ as the new isomorphism of base fields. Since α is a root of g and hence of f, we have $f = (x - \alpha)h$ for some $h \in F(\alpha)[x]$. Since f splits over K, all of the roots of f in K other than α, say $\beta_1, \ldots, \beta_m$, are roots of h. It follows that $K = F(\alpha)(\beta_1, \ldots, \beta_m)$ is a splitting field of $h \in F(\alpha)[x]$ over $F(\alpha)$. As usual we extend θ to an isomorphism of polynomial rings $\theta : F(\alpha)[x] \to F'(\alpha')[x]$, and as such $\theta(f) = (x - \theta(\alpha))\theta(h) = (x - \alpha')h'$ where $h' = \theta(h) \in F'(\alpha')[x]$. Thus we similarly conclude that K' is a splitting field of h' over $F'(\alpha')$.

We thus have an isomorphism $\theta : F(\alpha) \to F'(\alpha')$ and splitting fields K of $h \in F(\alpha)[x]$ and K' of $\theta(h) = h' \in F'(\alpha')[x]$. Since $\deg h < \deg f$, by the induction hypothesis there is an isomorphism $\sigma : K \to K'$ such that $\sigma|_{F(\alpha)} = \theta$ (in particular, $\sigma(\alpha) = \theta(\alpha) = \alpha'$). We also have $\sigma|_F = \theta|_F = \phi$.

This proves part (1). We have proved part (2) along the way, since we saw that (2) is trivial if f splits over F, and otherwise to do the proof we chose an arbitrary irreducible factor g of f with $\deg g \geq 2$ and constructed a σ with $\sigma(\alpha) = \alpha'$, where α and α' were arbitrary roots of g and $\phi(g) = g'$, respectively. □

It should be clear now why Lemma 17.33 and Proposition 17.36 were stated using an isomorphism of base fields $\phi : F \to F'$ rather than a fixed base field. In the proof of the induction step in Proposition 17.36, even if we had started with $F = F'$ and $\phi = 1_F$ at the beginning, we would be forced to consider an isomorphism of base fields $\theta : F(\alpha) \to F(\alpha')$ at the induction step, where these fields are isomorphic but different in general. To make the induction work it is necessary for the statement to be in terms of an isomorphism of base fields from the start.

Nonetheless, we usually apply Proposition 17.36 in the case where $F = F'$ and $\phi = 1_F$. In this case, it tells us that if $F \subseteq K$ and $F \subseteq K'$ are both splitting fields of $f \in F[x]$, then there is an isomorphism $\sigma : K \to K'$ with $\sigma|_F = 1_F$. This tells us that splitting fields are unique up to isomorphism, as we claimed earlier. The proposition also tells us how to construct automorphisms of a splitting field which move roots around; this special case is worth singling out:

Corollary 17.37. *Let f be a polynomial in $F[x]$ and let K be a splitting field for f over F. If g is an irreducible factor of f in $F[x]$ and $\alpha, \alpha' \in K$ are both roots of g, then there exists an automorphism $\sigma : K \to K$ such that $\sigma(\alpha) = \alpha'$.*

Proof. Just take $F = F'$, $\phi = 1_F$, and $K = K'$ in Proposition 17.36. □

Example 17.38. Let us revisit Example 17.35. A splitting field of $x^3 - 2$ over $\mathbb{Q}$ is $K = \mathbb{Q}(\alpha, \zeta)$ where $\alpha = \sqrt[3]{2}$, $\zeta = e^{2\pi i/3}$, and $[K : \mathbb{Q}] = 6$. Let us construct 6 different automorphisms of K.

Since $x^3 - 2$ is irreducible, by Corollary 17.37 we can find automorphisms σ, τ of K such that $\sigma(\alpha) = \alpha\zeta$ and $\tau(\alpha) = \alpha\zeta^2$. We also saw in Example 17.32 that $g = x^2 + x + 1 = \text{minpoly}_{\mathbb{Q}}(\zeta)$ remains irreducible over $\mathbb{Q}(\alpha)$. Clearly K is the splitting field of g over $\mathbb{Q}(\alpha)$. Thus by the Corollary again, there is an automorphism ρ of K such that $\rho|_{\mathbb{Q}(\alpha)} = 1_{\mathbb{Q}(\alpha)}$ but $\rho(\zeta) = \zeta^2$.

Now it is easy to check that the 6 automorphisms $\{1_K, \sigma, \tau, \rho, \sigma\rho, \tau\rho\}$ of K are all different, as no two act the same way on both α and ζ.

17.4. **Separability.** Suppose that $f \in F[x]$ is a monic polynomial over a field F and that $F \subseteq K$ is a field extension such that f splits in $K[x]$, say $f = (x - \alpha_1) \dots (x - \alpha_n) \in K[x]$. Could it be that some of the α_i are equal? Of course there are easy ways to make this happen; for example we could have $f = (x - \alpha)^2$ with $\alpha \in F$ already; or slightly less trivially, $f = g^2$ for some irreducible polynomial $g \in F[x]$ which splits over K, so that each root appears twice when we factor f into linear factors over K. An example of the latter phenomenon would be $f = (x^2 + 1)^2 \in \mathbb{Q}[x]$.

What is less obvious is whether there could be an *irreducible* polynomial $f \in F[x]$ where some of the α_i are equal in the factorization of f over K. In fact this does happen, but only for special kinds of fields and field extensions which are quite different from the examples given so far. The goal of this section is to study this phenomenon, and also show that it is something that doesn't happen in many of the most common situations.

Definition 17.39. A polynomial $f \in F[x]$ is called *separable* if given a splitting field K for f over F, f factors as $f = c(x - \alpha_1) \dots (x - \alpha_n) \in K[x]$ with $\alpha_1, \alpha_2, \dots, \alpha_n$ distinct elements of K. Otherwise we say that the polynomial f is *inseparable*.

We have seen that if $F \subseteq K$ and $F \subseteq K'$ are both splitting fields for $f \in F[x]$, then there is an isomorphism $\sigma : K \to K'$ such that $\sigma|_F = 1_F$. Using this it is easy to see that the definition above is independent of the choice of splitting field; if f splits with distinct roots in one splitting field, the same will be true in any other. Note that f is separable if and only if f has $\deg f$ distinct roots in a splitting field K.

Example 17.40. $(x^2 + 1)^2 \in \mathbb{Q}[x]$ is an inseparable polynomial, as already mentioned; in $\mathbb{C}[x]$ it factors as $(x + i)(x + i)(x - i)(x - i)$. If $a \neq 0$, the polynomial $x^n - a \in \mathbb{Q}[x]$ is separable over $\mathbb{Q}$ for all $n \geq 1$, as we have seen that it has n distinct roots in $\mathbb{C}$ in Example 17.32.

Example 17.41. Here is an example of a polynomial which is inseparable and also irreducible. Let $F = \mathbb{F}_2(y)$ be the field of rational functions in one variable over the field $\mathbb{F}_2$ with two elements. Note that $\text{char}\, F = 2$. We claim that $f = x^2 - y \in F[x]$ is an irreducible polynomial over F. This

follows from the Eisenstein criterion, thinking of F as the field of fractions of $\mathbb{F}_2[y]$, since y is prime in $\mathbb{F}_2[y]$. Now let $K = F[x]/(x^2 - y)$ and think of $F \subseteq K$ as a field extension as usual. In K there is a root $\alpha = x + (x^2 - y)$ of the polynomial $x^2 - y$. In other words, $\alpha^2 = y$ in K. Now note that $(x - \alpha)^2 = x^2 - 2\alpha + \alpha^2 = x^2 + y = x^2 - y$ since we are in characteristic 2. Thus the irreducible polynomial $x^2 - y$ factors in $K[x]$ as the square $(x - \alpha)^2$ and thus has only the single root α in K. Hence f is inseparable.

The example above may seem complicated at first, but it is in some sense the simplest example of an inseparable irreducible polynomial. That will become clear from the next results.

A useful technical tool in studying separability is given by the formal derivative of a polynomial.

Definition 17.42. Let F be any field. If $f = a_n x^n + \cdots + a_1 x + a_0 \in F[x]$ we define its *derivative* as

$$f' = n a_n x^n + (n-1) a_{n-1} x^{n-1} + \cdots + 2a_2 x + a_1 = \sum_{i=1}^{n} i a_i x^{i-1} \in F[x].$$

Remark 17.43. This definition requires some interpretation. In the formula $\sum_{i=0}^{n} i a_i x^{i-1}$ for f', the coefficient ia_i is the "ith multiple" of a_i, which is defined in any field F. In other words, i really means the ith multiple of 1, or the image of i under the canonical ring homomorphism $\mathbb{Z} \to F$. In particular, if $\operatorname{char} F = p > 0$ then some coefficients of f' may become 0; for example $(x^p)' = p x^{p-1} = 0$.

It is also worth pointing out that there is no limiting process involved here as is used in the definition of the derivative in calculus. We are only defining the derivative of a polynomial, which is done with an explicit formula. Nonetheless, it is easy to check that this definition satisfies all of the usual differentation formulas. In particular, if $f, g \in F[x]$ then $(f+g)' = f' + g'$; $(fg)' = fg' + f'g$; and $(f^d)' = d f^{d-1} f'$ for any positive integer d.

The next result gives an explicit connection between the derivative of a polynomial and separability.

Lemma 17.44. *Let $f \in F[x]$. Then f is separable if and only if* $\gcd(f, f') = 1$.

Proof. Let $F \subseteq K$ be a splitting field for f over F. In $K[x]$ we have

$$f = c(x - \alpha_1)^{e_1} (x - \alpha_2)^{e_2} \dots (x - \alpha_m)^{e_m},$$

where $\alpha_1, \alpha_2, \dots, \alpha_m$ are distinct in K; f is separable if and only if $e_i = 1$ for all i.

Note that the derivative f' of f is independent of whether we are thinking of f as a polynomial over $F[x]$ or over $K[x]$. Also, the product rule for derivatives extends to more than 2 factors as $(f_1 f_2 \dots f_m)' = \sum_{i=1}^m f_1 f_2 \dots f_{i-1} (f_i)' f_{i+1} \dots f_m$. Thus we have

$$f' = \sum_{i=1}^{m} c e_i (x-\alpha)^{e_1} \dots (x-\alpha_{i-1})^{e_{i-1}} (x-\alpha_i)^{e_i - 1} (x-\alpha_{i+1})^{e_{i+1}} \dots (x-\alpha_m)^{e_m}.$$

From this we see that if $e_i \geq 2$, then $(x-\alpha_i)$ divides every term of the sum and so $(x-\alpha_i) | f'$ (in $K[x]$). Of course $(x-\alpha_i)|f$ also, so $(x-\alpha_i)|\gcd_{K[x]}(f,f')$. Conversely, if $e_i = 1$, then $(x-\alpha_i)$ divides every term of the sum except the ith; so $(x-\alpha_i)$ does not divide f'. Since the $(x-\alpha_i)$ are the only irreducible factors of f in $K[x]$, if $e_i = 1$ for all i then we get $\gcd_{K[x]}(f,f') = 1$.

We have proved that f is separable if and only if $\gcd_{K[x]}(f,f') = 1$. However, the gcd of two polynomials in $F[x]$ is the same whether calculated over $F[x]$ or over $K[x]$; this can be shown by noting that the steps in the Euclidean algorithm are the same in either case. Thus f is separable if and only if $\gcd_{F[x]}(f,f') = 1$. □

The lemma has immediate interesting consequences for what an irreducible inseparable polynomial could possibly look like.

Proposition 17.45. *Suppose that $f \in F[x]$ is irreducible over F. Then f is inseparable if and only if*

(i) $\operatorname{char} F = p$ *for some* $p > 0$*; and*

(ii) $f = \sum_{i=0}^n b_i x^{ip}$ *for some* $b_i \in F$.

Proof. Suppose that f is inseparable. By Lemma 17.44, $\gcd(f,f') \neq 1$. However, since f is irreducible, its only divisors (up to associates) are 1 and f. Thus $\gcd(f,f') = f$. But how can this happen? Note that $\deg f' < \deg f$ always. If $f|f'$ this forces $f' = 0$.

Now since f is irreducible, $\deg f \geq 1$. If $f = \sum_{i=0}^n a_i x^i$ then $f' = \sum_{i=1}^n i a_i x^{i-1}$ so $i a_i = 0$ for all $i \geq 1$. We can think of this as $(i \cdot 1) a_i = 0$ where $i \cdot 1$ is the ith multiple of 1. Since F is a domain, for each i either $a_i = 0$ or else $i \cdot 1 = 0$; the latter happens if and only if $\operatorname{char} F = p > 0$ and i is a multiple of p. It follows that if $\operatorname{char} F = 0$ then f can have only a constant term, so $\deg f = 0$ and f is not irreducible, a contradiction. Thus we must have as in (i) that $\operatorname{char} F = p > 0$, and we see that f is a polynomial whose only nonzero coefficents are the a_i where i is a multiple of p. By reindexing such a polynomial we get one in the form of (ii).

Conversely, if (i) and (ii) hold, a similar argument shows that $f' = 0$, and thus $\gcd(f,f') = f \neq 1$. Thus f is inseparable by Lemma 17.44. □

The proposition implies the useful fact that for a field of characteristic 0, all irreducible polynomials are separable. In particular, when working over $\mathbb{Q}$, as is the main setting for many investigations in field theory, separability becomes a non-issue. The characteristic p setting is still very important to applications, however, and so it is interesting to push our results in this case further. We will see shortly that certain special fields of characterstic p also have no irreducible inseparable polynomials.

Definition 17.46. Let R be a commutative ring with $\operatorname{char} R = p > 0$. Then the *Frobenius homomorphism* is the map $\phi : R \to R$ given by $\phi(a) = a^p$.

Note that while the pth power map preserves multiplication in any commutative ring, the preservation of addition works here only because we are in characteristic p. This follows from the binomial formula:

$$\phi(a+b) = (a+b)^p = \sum_{i=0}^{p} \binom{p}{i} a_i b^{p-i} = a^p + b^p$$

because $\binom{p}{i} = p!/(i)!(p-i)!$ is a multiple of p for all $0 < i < p$. Thus the Frobenius homomorphism really is a homomorphism of rings.

Proposition 17.47. *Let F be a field with $\operatorname{char} F = p > 0$. If the Frobenius homomorphism $\phi : F \to F$ is surjective, then every irreducible polynomial $f \in F[x]$ is separable.*

Proof. Let $f \in F[x]$ be irreducible and suppose that f is inseparable over F. By Proposition 17.45, $f = \sum_{i=0}^{n} b_i x^{ip}$ for $b_i \in F$. Now since the Frobenius is surjective, every element of F is a pth power. In particular, $b_i = (a_i)^p$ for some $a_i \in F$. Then

$$f = \sum_{i=0}^{n} b_i x^{ip} = \sum_{i=0}^{n} (a_i)^p (x^i)^p = \sum_{i=0}^{n} (a_i x^i)^p = (\sum_{i=0}^{n} a_i x^i)^p.$$

But now f factors as a product of p copies of the polynomial $g = \sum_{i=0}^{n} a_i x^i \in F[x]$, so f is not irreducible, a contradiction. □

When F is a field, we write the image of the Frobenius map $\phi : F \to F$ as $F^p = \{a^p | a \in F\}$. Since the Frobenius is a homomorphism of fields, it is injective and so F^p is a subfield of F which is isomorphic to F. The Frobenius need not be surjective, however.

Definition 17.48. A field F is *perfect* if either $\operatorname{char} F = 0$ or else $\operatorname{char} F = p > 0$ and $F^p = F$.

The following result justifies including these two very different cases in one definition.

Proposition 17.49. *A field F is perfect if and only if every irreducible polynomial in $F[x]$ is separable.*

Proof. We have seen that if either $\operatorname{char} F = 0$ or $\operatorname{char} F = p$ and $F^p = F$, then every irreducible $f \in F[x]$ is separable; see Proposition 17.45 and Proposition 17.47.

Conversely, suppose that F is not perfect. Thus there is a prime p such that $\operatorname{char} F = p > 0$ and $F^p \neq F$. thus we can pick $a \in F$ such that a is not a pth power of an element in F. Now let $f = x^p - a \in F[x]$ and find a field extension $F \subseteq K$ such that f has a root in K, say $\alpha \in K$. This means that $\alpha^p - a = 0$, so α is a pth root of a in K. Now $(x-\alpha)^p = x^p - \alpha^p = x^p - a = f$, so $f = (x-\alpha)^p$ already splits in $K[x]$ as a pth power. Thus f is inseparable.

On the other hand, we claim that f is irreducible over F. If not, then $f = gh$ with $g, h \in F[x]$ where $\deg g \geq 1$, $\deg h \geq 1$. Now because of the factorziation of f in $K[x]$ is a pth power of a degree 1 irreducible, we must have $g = (x-\alpha)^i$ and $h = (x-\alpha)^j$ in $K[x]$, where $i + j = p$ and $i, j \geq 1$. But now the constant term of g is $\pm\alpha^i$, so $\alpha^i \in F$. Since $\gcd(i, p) = 1$ and $\alpha^p = a \in F$, writing $1 = mi + np$ we get $\alpha^1 = (\alpha^i)^m(\alpha^p)^n \in F$, a contradiction since a has no pth root in F. Thus f is irreducible over F as claimed. □

Example 17.50. Let F be a field with $\operatorname{char} F = p > 0$, and suppose that F is finite. Then F must be perfect. Indeed, the Frobenius map $\phi : F \to F$ is always injective (since F is a field). Then since F is a finite set this forces ϕ to be surjective as well.

Example 17.51. Let $F = \mathbb{F}_p(y)$ be a field of fractions functions in one variable over $\mathbb{F}_p$. Then F is not perfect. In fact, this must be true, we already saw Example 17.41 that $F[x]$ has an irreducible inseparable polynomial when $p = 2$, and a similar example works for arbitrary p.

But we can also check it directly from the definition of perfect, by showing that $y \in F$ has no pth root. Indeed, suppose that $f/g \in F$, where $f, g \in \mathbb{F}_p[y]$, and that $(f/g)^p = y$. Then $f^p = yg^p$ in $\mathbb{F}_p[y]$. But then considering degrees we have $p \deg f = p \deg g + 1$, which is absurd.

There are infinite fields of characteristic p which are perfect, as well. An easy example is any algebraically closed field of characteristic p (we will see later that these exist). Over an algebraically closed field there are no irreducible polynomials except those of degree 1, which are trivially separable.

17.5. **Finite fields and the Theorem of the Primitive Element.** In the first part of this section, we give the basic structure theory of fields with finitely many elements.

We start with an easy group theory result and its application to the structure of the multiplicative group of a field.

Lemma 17.52. *Let G be a finite abelian group of order n. Suppose that for each divisor d of n that G has at most d elements of order dividing d. Then G is cyclic.*

Proof. We use the classification of finite abelian groups in the invariant factor form. This tells us that G is isomorphic to an additive group $\mathbb{Z}/(a_1) \oplus \cdots \oplus \mathbb{Z}/(a_m)$, where $a_1|a_2|\ldots|a_m$ are integers greater than 1. Suppose that $m \geq 2$. Then since $a_{m-1}|a_m$, every element of the form $g = (0, 0, \ldots, 0, b, c)$ satisfies $a_m g = 0$. There are $(a_m)(a_{m-1}) > a_m$ such elements. In other words, G has more than a_m elements of order dividing a_m, contradicting the hypothesis. Thus $m = 1$ and $G \cong \mathbb{Z}/(a_1)$ is cyclic. □

Corollary 17.53. *Let F a be a field and $F^\times = F - \{0\}$ its multiplicative group. If G is a finite subgroup of $F^\times$ then G is cyclic. In particular, if F is a finite field then $F^\times$ is cyclic.*

Proof. Using multiplicative notation, the elements in G of order dividing d are $\{g \in G | g^d = 1\}$. In other words, these are roots in G of the polynomial $x^d - 1 \in F[x]$. This polynomial can have at most d roots in F by Corollary 11.11. Thus the hypotheses of the lemma are satisfied and we conclude that G is cyclic. The last statement is clear. □

Now let us prove the basic structural results of finite fields. Note that if F is a finite field, then certainly 1 has finite additive order and so F has positive characteristic, say p for a prime p. Then the additive subgroup generated by 1 is a subfield isomorphic to $\mathbb{F}_p$, so we can think of F as an extension of $\mathbb{F}_p$.

Theorem 17.54. *Let p be prime. For each $n \geq 1$, setting $q = p^n$ the splitting field of $x^q - x$ over $\mathbb{F}_p$ is a field $\mathbb{F}_q$ with $|\mathbb{F}_q| = q$. Conversely, if F is any finite field of characteristic p then $F \cong \mathbb{F}_q$ for $q = p^n$, some $n \geq 1$.*

Proof. Let F be the splitting field of $f = x^{p^n} - x$ over $\mathbb{F}_p$. Note that $f' = p^n x^{p^n - 1} - 1 = -1$ since we are in characteristic p, so we must have $\gcd(f, f') = 1$, and by Lemma 17.44 f is separable. Thus f has p^n distinct roots in F. Let E be the set of these roots in F, so $E = \{\alpha \in F | \alpha^{p^n} = \alpha\}$. Note that $\phi : F \to F$ given by $\phi(x) = x^{p^n}$ is the nth power of the Frobenius homomorphism; in particular, ϕ is a homomorphism of fields. Then $E = \{x \in F | \phi(x) = x\}$ is automatically a subfield of F. Now since E contains all of the roots of f, the polynomial f already splits in $E[x]$, and obviously the subfield of F generated over $\mathbb{F}_p$ by these roots must be E. Hence $F = E$ and the spliting field of f is actually equal to the set of roots of f. In particular $|F| = p^n$. Letting $q = p^n$ and writing $\mathbb{F}_q = F$, we have $|\mathbb{F}_q| = q$.

Conversely, let F be a finite field of characteristic p. As we remarked above, F contains a copy of $\mathbb{F}_p$. Thus we have a field extension $\mathbb{F}_p \subseteq F$. The degree $[F : \mathbb{F}_p] = n$ is certainly finite, since F is. Moreover, a vector space of dimension n over $\mathbb{F}_p$ clearly has precisely p^n elements, because this is the number of distinct $\mathbb{F}_p$-linear combinations of n basis vectors. Now $F^\times = F - \{0\}$ is a multiplicative group of order $p^n - 1$; so for any $\alpha \in F^\times$ we have $\alpha^{p^n-1} = 1$. Then $\alpha^{p^n} = \alpha$. This latter equation is also true for $\alpha = 0$. Hence every element of F is a root of $f = x^{p^n} - x \in \mathbb{F}_p[x]$. Since F consists of $p^n = \deg f$ distinct roots of f, the polynomial f must already split over F as $x^{p^n} - x = \prod_{\alpha \in F}(x - \alpha) \in F[x]$. So F is a splitting field of f over $\mathbb{F}_p$. Now by the uniqueness of splitting fields up to isomorphism, setting $q = p^n$ we have that $F \cong \mathbb{F}_q$ for the field $\mathbb{F}_q$ constructed above. □

Thinking of the field $\mathbb{F}_q$ for $q = p^n$ as the splitting field of $x^{p^n} - x$ was useful for theoretical reasons, but to actually construct a field $\mathbb{F}_q$, in practice it is helpful to find it as a splitting field of a polynomial of smaller degree. This can be done with the help of our result that the multiplicative group of a finite field is cyclic.

Lemma 17.55. *Let p be prime. For each $n \geq 1$ there is at least one irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n; and for any such f, $\mathbb{F}_p[x]/(f)$ is a field isomorphic to $\mathbb{F}_{p^n}$.*

Proof. Consider the field $F = \mathbb{F}_{p^n}$ as an extension of its prime subfield $\mathbb{F}_p$. We know that $F^\times$ is a cyclic group by Corollary 17.53, say with generator γ. Since the powers of γ fill up $F^\times$, clearly $\mathbb{F}_p(\gamma) = F$, so F is a simple extension of $\mathbb{F}_p$. Consequently $F \cong \mathbb{F}_p[x]/(f)$ where $f = \mathrm{minpoly}_{\mathbb{F}_p}(\gamma)$. But we also know that $[F : \mathbb{F}_p] = n = \deg f$, so f is an irreducible polynomial over $\mathbb{F}_p$ of degree n.

Conversely, it is clear that for any irreducible polynomial $g \in \mathbb{F}_p[x]$ of degree n, then $K = \mathbb{F}_p[x]/(g)$ is a field with $[K : \mathbb{F}_p] = n$ and hence $|K| = p^n$. □

Example 17.56. Irreducible polynomials of low degree over $\mathbb{F}_p$ can be found explicitly as described in Example 11.16. For example, as shown there, $f = x^4 + x + 1$ is irreducible over $\mathbb{F}_2$ and thus $\mathbb{F}_{16} \cong \mathbb{F}_2[x]/(x^4 + x + 1)$. This gives an explicit description of $\mathbb{F}_{16}$ that allows one to do calculations in this field, by writing its elements as $\{a_0 + a_1 x + a_2 x^2 + a_3 x^3 + (f) | a_i \in \mathbb{F}_2\}$ and doing arithmetic of such elements modulo (f).

For example, consider $x \in \mathbb{F}_{16}$ (ommitting the $+(f)$ and just remembering to do calculations modulo f). The relation tells us that $x^4 = -x - 1 = x + 1$. Then $x^5 = x^2 + x$ and $x^3 = x^3$ are not equal to 1 (since the representative of a coset with degree ≤ 3 is uniquely determined). This shows that the order of x in $\mathbb{F}_{16}^\times$ is 15 and so x is a generator of the cyclic group $\mathbb{F}_{16}^\times$.

Corollary 17.53 is also relevant to the proof of a basic theorem often called the "Theorem of the Primitive Element". It gives a useful criterion for when a finite degree extension is a simple extension, i.e. generated by one element. The name arises from the fact that in a simple extension $F \subseteq F(\gamma)$ the element γ is sometimes called a primitive element for the extension.

Theorem 17.57. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$. The following are equivalent:*

(i) *$K = F(\gamma)$ for some $\gamma \in K$.*

(ii) *There are finitely many subfields E with $F \subseteq E \subseteq K$.*

Proof. Different arguments are required here depending on whether F is finite or infinite. The finite case follows quickly from results we have already proved. If $|F| < \infty$ then $|K| = [K : F]|F| < \infty$ also. Then $K^\times$ is a cyclic group by Corollary 17.53, generated by some γ, say. Thus $K = F(\gamma)$, so every finite degree extension of a finite field is simple and (i) is automatic. Condition (ii) also trivially holds when F and hence K is finite, since K has finitely many distinct subsets.

Assume for the rest of the proof that $|F| = \infty$. Suppose as in (ii) that there are finitely many subfields E with $F \subseteq E \subseteq K$ (these are called *intermediate* fields for the extension $F \subseteq K$). Suppose that $\alpha, \beta \in K$ and consider the fields $E_a = F(\alpha + a\beta)$ as a ranges over elements of F. Since each E_a is an intermediate field, of which there are only finitely many, yet $|F| = \infty$, we must have $E_a = E_b$, for some $a \neq b$. Then E_a contains $\alpha + a\beta - (\alpha + b\beta) = (a-b)\beta$ and since $0 \neq a - b \in F$, $\beta \in E_a$. But then $a\beta \in E_a$ and so $\alpha \in E_a$ also. It follows that $E_a = F(\alpha + a\beta) = F(\alpha, \beta)$. This shows that the subfield of K generated over F by any two elements can actually be generated by one element. Since $[K : F] < \infty$, K/F is certainly finitely generated. By iteratively replacing pairs of generators by a single generator we obtain that $K = F(\gamma)$ for some γ.

Conversely, suppose that $K = F(\gamma)$ for some $\gamma \in K$. Let $F \subseteq E \subseteq K$ where E is an intermediate field. Since $[K : F] < \infty$, certainly $[K : E] < \infty$, so γ is algebraic over E. Let $f = \operatorname{minpoly}_E(\gamma) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in E[x]$. Since all $a_i \in E$, we have $E' = F(a_0, \ldots, a_{n-1}) \subseteq E$. Since f is irreducible in $E[x]$, and also $f \in E'[x]$, certainly f is irreducible in $E'[x]$. Thus $f = \operatorname{minpoly}_{E'}(\gamma)$ also. But this means that $[K : E'] = [E'(\gamma) : E'] = \deg f = [E(\gamma) : E] = [K : E]$. This forces $[E : E'] = 1$ and hence $E' = E$. In particular, E is generated over F by the coefficients of $f = \operatorname{minpoly}_E(\gamma)$. Now let $g = \operatorname{minpoly}_F(\gamma)$. Then since $g(\gamma) = 0$, we have $f | g$ in $E[x]$, hence in $K[x]$. Since there are only finitely many monic polynomials that divide g in $K[x]$, there are finitely many possible f and so finitely many intermediate fields E. □

We will see later that if $F \subseteq K$ is a finite degree separable extension, then condition (ii) above holds and hence $K = F(\gamma)$ is a simple extension. For example, an arbitrary finite degree extension of $\mathbb{Q}$ can be generated by one element, which is quite surprising.

18. Galois Theory

18.1. **Separable and normal extensions.** In this section we study two special properties of field extensions, separability and normality, and their relations to automorphisms of fields. This will lay the main groundwork for the fundamental theorem of Galois Theory in the next section. This theory is named for the French mathematician Évariste Galois (pronounced "gal-wah"), who developed its main ideas at a young age before his life was tragically cut short in a duel. At the time, there was no notion of a "group"—the groups that occured in Galois's work were explicit subsets of the group of permutations of the roots of a polynomial. These ideas nonetheless helped to point the way to the idea of an abstract group which was formulated later in the 1800's.

Definition 18.1. Let $F \subseteq K$ be a field extension. We define

$$\mathrm{Aut}(K) = \{\sigma : K \to K \mid \sigma \text{ is an automorphism of fields}\},$$

which is a group under composition. The *Galois group* of the field extension K/F is

$$\mathrm{Gal}(K/F) = \{\sigma \in \mathrm{Aut}(K) \mid \sigma(a) = a \text{ for all } a \in F\}.$$

It is a subgroup of $\mathrm{Aut}(K)$.

We read $\mathrm{Gal}(K/F)$ as "Galois K over F". We say that the elements of $\mathrm{Gal}(K/F)$ *fix* F. Note that these elements are required to fix F pointwise, not just as an overall set.

One reason it is interesting to consider automorphisms of K that fix F is the following. If $\sigma \in \mathrm{Gal}(K/F)$ and $f \in F[x]$, then if $\alpha \in K$ is a root of f, then $\sigma(\alpha)$ is also a root of f. Explicitly, if $f = \sum a_i x^i$ with $a_i \in F$, then $\sum a_i \alpha^i = 0$, so applying σ we get

$$\sum \sigma(a_i)\sigma(\alpha)^i = \sum a_i \sigma(\alpha)^i = f(\sigma(\alpha)) = 0,$$

since σ fixes F. Thus elements in $\mathrm{Gal}(K/F)$ must permute any roots of $f \in F[x]$ that lie in K. We will use this fact frequently.

It turns out that finite degree extensions which have "enough" automorphisms have particularly good properties and will be the focus of the fundamental theorem later. The relevant definition is the following.

Definition 18.2. A finite degree extension K/F is called *Galois* if $|\mathrm{Gal}(K/F)| = [K : F]$.

Here are two examples that show the sort of things that might go wrong and prevent an extension from being Galois.

Example 18.3. Consider $\mathbb{Q} \subseteq K = \mathbb{Q}(\sqrt[3]{2})$ inside $\mathbb{C}$. Since $K \subseteq \mathbb{R}$ and the other two roots of $x^3 - 2$ in $\mathbb{C}$ are not real, $\sqrt[3]{2}$ is the only root of $x^3 - 2$ in K. Now if $\sigma \in G = \mathrm{Gal}(K/\mathbb{Q})$, then by the remark above $\sigma(\sqrt[3]{2}) = \sqrt[3]{2}$ is forced. Since $\sigma \in G$ fixes $\mathbb{Q}$ and the element $\sqrt[3]{2}$ which generates K, it follows that $\sigma = 1_K$. Thus G is trivial and $|\mathrm{Gal}(K/\mathbb{Q})| = 1 < [K : \mathbb{Q}] = 3$.

Example 18.4. Consider $F = \mathbb{F}_2(y)$. As we have already seen in Example 17.41, the polynomial $f = x^2 - y \in F[x]$ is inseparable and irreducible. If K is a splitting field of f over F then there is $\alpha \in K$ such that $f(\alpha) = 0$, that is $\alpha^2 = y$, and where $f = (x - \alpha)^2 \in K[x]$. So $K = F(\alpha)$. Again if $\sigma \in \mathrm{Gal}(K/F)$ then $\sigma(\alpha) = \alpha$ because σ permutes the roots of $f \in F[x]$ (and α is the only root). This implies that $\sigma = 1_K$ and so $|\mathrm{Gal}(K/F)| = 1 < [K : F] = 2$.

In the next results we will see that the things that happened in the two examples above are the *only* things that can go wrong in an extension K/F that is not Galois—either there is an irreducible polynomial in $F[x]$ that does not entirely split in $K[x]$, so it doesn't have enough roots in K, or a polynomial in $F[x]$ that splits in $K[x]$ but with indistinct roots, so again there are not enough different places for an automorphism to send the roots. This leads to the following two definitions.

Definition 18.5. Let $F \subseteq K$ be an algebraic field extension. We say the extension K/F is *separable* if for all irreducible polynomials $f \in F[x]$, if f has a root in K then f is separable.

Definition 18.6. Let $F \subseteq K$ be an algebraic field extension. We say the extension K/F is *normal* if for all irreducible polynomials $f \in F[x]$, if f has a root in K then f splits over K.

An alternative way to define these notions is using minimal polynomials: given an extension K/F, it is normal if for all $\alpha \in K$, $\mathrm{minpoly}_F(\alpha)$ splits in $K[x]$, and it is separable if for all $\alpha \in K$, $\mathrm{minpoly}_F(\alpha)$ is a separable polynomial. This follows immediately from the fact that any irreducible polynomial in $F[x]$ that has α as a root must be the minimal polynomial of α.

Note that in Example 18.3, the extension $\mathbb{Q} \subseteq K$ is not normal, because the irreducible polynomial $f = x^3 - 2 \in \mathbb{Q}[x]$ has the root $\sqrt[3]{2} \in K$ but does not split over K. In Example 18.4, the extension $F \subseteq K$ fails to be separable, because the minimal polynomial of α is the irreducible inseparable polynomial $x^2 - y \in F[x]$.

It is useful to see how the separable and normal properties pass to smaller extensions.

Lemma 18.7. *Let $F \subseteq E \subseteq K$ be field extensions where K/F is algebraic.*

(1) *If K/F is separable, then E/F and K/E are separable.*

(2) *if K/F is normal, then K/E is normal.*

Proof. (1) It is obvious that E/F is separable from the definition–we are just checking $\text{minpoly}_F(\alpha)$ is separable for those $\alpha \in E$, rather than for all $\alpha \in K$. Now if $\alpha \in K$, consider $g = \text{minpoly}_E(\alpha)$. If $f = \text{minpoly}_F(\alpha)$, then since $f(\alpha) = 0$ and $f \in F[x] \subseteq E[x]$, we have $g|f$ in $E[x]$. But now since f has distinct roots in a splitting field, so does its factor g. So K/E is also separable.

(2) Similarly as in part (1), for $\alpha \in K$, $g = \text{minpoly}_E(\alpha)$ divides $f = \text{minpoly}_F(\alpha)$. Now since f splits over K, so does its factor g. □

It turns out that finite degree normal extensions are the same as splitting fields of polynomials. Both points of view are useful, since the notion of normality doesn't depend on a choice of polynomial, so it is easier to work with abstractly; while thinking in terms of the splitting field of a particular polynomial is important in calculations.

Lemma 18.8. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$. Then K/F is normal if and only if there is a polynomial $f \in F[x]$ such that K is the splitting field of f over F.*

Proof. Suppose first that K is the splitting field over F of $f \in F[x]$. Thus $f = a(x-\alpha_1)\dots(x-\alpha_m)$ in $K[x]$, with $K = F(\alpha_1, \dots, \alpha_m)$. Let $g \in F[x]$ be irreducible over F and assume that $g(\beta_1) = 0$ with $\beta_1 \in K$. Let $K \subseteq L$ where L is the splitting field of g over K. So the polynomial g splits as $g = a(x - \beta_1)(x - \beta_2)\dots(x - \beta_n)$ in $L[x]$, and $L = K(\beta_1, \dots, \beta_n)$.

Now note that f and g both split over L, and that $L = F(\alpha_1, \dots, \alpha_m, \beta_1, \dots \beta_n)$. This implies that L is a splitting field of $h = fg \in F[x]$ over F. By Corollary 17.37, since g is an irreducible factor of h, for any root β_i of g there is an automorphism σ of L such that σ satisfies $\sigma|_F = 1_F$, in other words σ fixes F, and $\sigma(\beta_1) = \beta_i$. On the other hand, since f has coefficients in F, σ must permute the roots $\{\alpha_j\}$ of f as well. Since K is generated by the $\{\alpha_j\}$ over F, $\sigma(K) = K$. In particular, since $\beta_1 \in K$, we get $\sigma(\beta_1) = \beta_i \in K$ for all i. This shows that $L \subseteq K$ and hence $K = L$. Thus g already splits over K. We see that any irreducible polynomial in $F[x]$ with a root in K splits over K, so K/F is normal.

The converse is easier. If K/F is normal, then since $[K : F] < \infty$ the extension K/F is certainly finitely generated; say $K = F(\gamma_1, \dots, \gamma_r)$. Let $g_i = \text{minpoly}_F(\gamma_i)$. By the normality condition, each g_i must split over K. But then $f = g_1 g_2 \dots g_r$ is a polynomial that splits over K, and the set of all of its roots generates K over F since these roots are contained in K and include all of the γ_i. So K is the splitting field of f over F. □

The lemma allows for an alternate proof of Lemma 18.7: If $F \subseteq E \subseteq K$ with K/F normal, then we know that K is the splitting field over F over some $f \in F[x]$. It is easy to see then that K is also the splitting field over E of the same f, so K/E must also be normal.

Example 18.9. We can also now give an example showing that if $F \subseteq E \subseteq K$ and K/F is normal, then E/F need not be normal. Consider $F = \mathbb{Q} \subseteq E = \mathbb{Q}(\sqrt[3]{2}) \subseteq K = \mathbb{Q}(\sqrt[3]{2}, \zeta)$ where $\zeta = e^{2\pi i/3}$. We have seen in Example 17.32 that K is the splitting field over F of $x^3 - 2$. On the other hand, we saw that E is not normal over F in Example 18.3. Thus normality of K/F does not pass in general to the subextension E/F.

We now give the main result of this section. It shows that the failure of an extension to be Galois is always because there are two few automorphisms, not too many; and this always happens essentially because of one of the two problems exhibited in Examples 18.3 and 18.4, namely lack of normality or lack of separability.

Theorem 18.10. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$.*

(1) $|\operatorname{Gal}(K/F)| \leq [K : F]$.

(2) *The following are equivalent:*

(i) *K/F is Galois, i.e. $|\operatorname{Gal}(K/F)| = [K : F]$.*

(ii) *K/F is separable and normal.*

(iii) *K is the splitting field over F of a separable polynomial $f \in F[x]$.*

Proof. (1) If $K = F$ the result is vacuous, so assume that $[K : F] \geq 2$. Pick any $\alpha_1 \in K - F$ and let $g = \operatorname{minpoly}_F(\alpha_1)$, so g is irreducible over F and $n = \deg g \geq 2$, with $[F(\alpha_1) : F] = n$. Let $\{\alpha_1, \ldots, \alpha_m\}$ be the set of all elements in K that are roots of g, so $m \leq n$.

If $\sigma \in \operatorname{Gal}(K/F)$, then $\sigma(\alpha_1) = \alpha_i$ for some i, because $g \in F[x]$. Let

$$S = \{i \in \{1, 2, \ldots, m\} | \text{there exists an automorphism } \sigma \in \operatorname{Gal}(K/F) \text{ such that } \sigma(\alpha_1) = \alpha_i\}.$$

For each $i \in S$ fix an automorphism $\sigma_i \in \operatorname{Gal}(K/F)$ such that $\sigma_i(\alpha_1) = \alpha_i$.

Now let us prove that $|\operatorname{Gal}(K/F)| \leq [K : F]$ by induction on the degree $[K : F]$. We have $[K : F(\alpha_1)] < [K : F]$. By the induction hypothesis, $|\operatorname{Gal}(K/F(\alpha_1))| \leq [K : F(\alpha_1)]$. Now if $\tau \in \operatorname{Gal}(K/F)$ is arbitrary, then $\tau(\alpha_1) = \alpha_i$ for some $i \in S$; then $\rho = (\sigma_i)^{-1} \circ \tau \in \operatorname{Gal}(K/F)$ and $\rho(\alpha_1) = \alpha_1$. Thus ρ fixes $F(\alpha_1)$ pointwise, and so $\rho \in \operatorname{Gal}(K/F(\alpha_1))$. Let $H = \operatorname{Gal}(K/F(\alpha_1))$. We conclude that $\tau = \sigma_i \circ \rho \in \sigma_i H$. Thus

$$|\operatorname{Gal}(K/F)| \leq |S||H| \leq m|H| \leq n|\operatorname{Gal}(K/F(\alpha_1))| \leq |F(\alpha_1) : F||K : F(\alpha_1)| = |K : F|. \tag{18.11}$$

(2) Consider the argument in part (1). In order to have equality in (18.11), so that $|\operatorname{Gal}(K/F)| = |K : F|$, it is necessary and sufficient that $|S| = m = n$ and $|\operatorname{Gal}(K/F(\alpha_1))| = [K : F(\alpha_1)]$.

$(i) \implies (ii)$. The case $K = F$ is trivial. Assume that K/F is Galois with $[K : F] > 1$. Choose any $\alpha_1 \in K - F$ and let $g = \operatorname{minpoly}_F(\alpha_1)$. As just remarked, we must have the number m of elements in K that are roots of g equal to the degree n of g in the argument of part (1). In particular, this forces g to split over K, and with distinct roots, so that g is a separable polynomial. By definition, K/F is normal and separable.

$(ii) \implies (iii)$. Since K/F is normal, K is the splitting field over F of some polynomial $f \in F[x]$ by Lemma 18.8. We may assume that f is monic. Write $f = (g_1)^{e_1} \dots (g_m)^{e_m}$ where the g_i are monic irreducibles in $F[x]$ such that $g_1, \dots, g_m$ are all distinct, and $e_i \geq 1$. It is clear that K is also the splitting field of $h = g_1 \dots g_m$ over F. Now if g_i and g_j have a common root $\alpha \in K$, then $g_i = \operatorname{minpoly}_F(\alpha) = g_j$, a contradiction. Also, since K/F is separable, each irreducible polynomial g_i is separable and thus splits with distinct roots in K. We conclude that h has distinct roots in K and so is a separable polynomial, and K is the splitting field of h over F.

$(iii) \implies (i)$. The proof is by induction on $[K : F]$, with the case $F = K$ trivial as usual. Assume that K is the splitting field of a separable polynomial f over F, where f does not split over F (otherwise we are back to the trivial case $F = K$). Let g be an irreducible factor of f in $F[x]$ with $\deg g \geq 2$, and apply the argument in (1) to a root $\alpha_1 \in K$ of g. Since K is a splitting field of f, the polynomial g splits over K, and since f is separable, so is g, so g has $m = n = \deg g$ distinct roots in K. In addition, for any root α_i of g, we can find an automorphism $\sigma_i \in \operatorname{Gal}(K/F)$ such that $\sigma_i(\alpha_1) = \alpha_i$, by Corollary 17.37. So $|S| = m = n$ in the argument of part (1). Finally, K is also the splitting field of the separable polynomial f over $F(\alpha_1)$. Since $[K : F(\alpha_1)] < [K : F]$, by induction we may assume that $|\operatorname{Gal}(K/F(\alpha_1))| = [K : F(\alpha_1)]$. Now (18.11) implies that $|\operatorname{Gal}(K/F)| = [K : F]$. □

Corollary 18.12. *Let K/F be a finite degree Galois extension. Then for every intermediate field E with $F \subseteq E \subseteq K$, the extension K/E is also Galois.*

Proof. By Theorem 18.10, we know that for a finite degree extension being Galois is equivalent to being normal and separable. But both properties pass from K/F to K/E by Lemma 18.7. □

Example 18.9 is an example of a Galois extension K/F such that E/F is not Galois.

Suppose $F \subseteq K$ is a finite-degree extension that is not necessarily normal. Then we can embed it canonically in a normal extension, as follows.

Proposition 18.13. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$.*

(1) *There is an extension $K \subseteq L$ with $[L : K] < \infty$ such that L/F is normal, and where L is minimal in the sense that if $K \subseteq E \subseteq L$ with E/F normal, then $E = L$.*

(2) *If K/F is separable in (1), then L/F is Galois.*

Proof. (1) K/F is certainly finitely generated, say $K = F(\alpha_1, \ldots, \alpha_m)$. Let $g_i = \text{minpoly}_F(\alpha_i) \in F[x]$. Some of the g_i may be equal; let f be the product of the distinct g_i, each once.

Now take L to be a splitting field of f over K. We define the splitting field over K, not over F, because we need to ensure the resulting field L contains K. However, if $\beta_1, \ldots, \beta_n$ are the roots of f in L then

$$L = K(\beta_1, \ldots, \beta_n) = F(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n) = F(\beta_1, \ldots, \beta_n)$$

because the elements α_i are roots of f and hence lie among the β_j. Thus L is in fact the splitting field of f over F as well. By Lemma 18.8, L/F is now a normal extension.

To see that L is minimal, note that if $K \subseteq E \subseteq L$ is an extension where E/F is normal, then each irreducible polynomial g_i has a root in K and so must split in E. But then all of the roots β_j of f in L are already in E, so $E = L$ since L is generated over K by the β_j.

(2) Now suppose that K/F is separable. Then each g_i is a separable polynomial, by definition. As we saw previously, distinct monic irreducible polynomials in $F[x]$ cannot have a common root. Since each g_i has distinct roots in the splitting field L, the product f of the distinct g_i is a separable polynomial. This L is a splitting field over F of a separable polynomial f. By Theorem 18.10, L is Galois over F. □

The extension L/F constructed in (1) above is called the *normal closure* of K/F. It is unique up to isomorphism, as the reader may check. When K/F is separable, so that the normal closure L/F is Galois as in part (2), then it is called the *Galois closure*.

18.2. **The Fundamental Theorem of Galois Theory.** The fundamental theorem we will prove in this section gives a surprisingly tight connection between group theory and field theory. Namely, the set of intermediate fields E such that $F \subseteq E \subseteq K$ where K/F is a finite degree Galois extension will be in one-to-one correspondence with the set of subgroups of the group $G = \text{Gal}(K/F)$. Moreover, the correspondence will have many other special properties.

The way the correspondence is set up is not particularly complicated to describe. Let $F \subseteq K$ be an extension of fields. Let $G = \text{Gal}(K/F)$, which we know is a group under composition. If E is an intermediate field with $F \subseteq E \subseteq K$, then we can define a subgroup H of G by $H = \text{Gal}(K/E)$. In other words, these are the automorphisms of K that fix (pointwise) the larger field E, not just

F. This is obviously a subset of G, and since the elements fixing E are closed under products and inverses, H is a subgroup of G.

Conversely, if we start with a subgroup H of $G = \mathrm{Gal}(K/F)$, then we define an intermediate field E by $E = \mathrm{Fix}(H) = \{\alpha \in K | \sigma(\alpha) = \alpha \text{ for all } h \in H\}$. This is called the *fixed field of* H. Because the elements in H are automorphisms of K it is clear that $\mathrm{Fix}(H)$ is a subfield of K; and $F \subseteq \mathrm{Fix}(H)$ since every element of G fixes F. So $F \subseteq \mathrm{Fix}(H) \subseteq K$ and $\mathrm{Fix}(H)$ is an intermediate field.

Thus for a fixed field extension $F \subseteq K$ we have the following setup and notation:

(18.14)

$$\left\{\text{intermediate fields } E \text{ with } F \subseteq E \subseteq K\right\} \underset{\Phi=\mathrm{Fix}(-)}{\overset{\Gamma=\mathrm{Gal}(K/-)}{\rightleftarrows}} \left\{\text{subgroups } H \text{ of } G = \mathrm{Gal}(K/F)\right\}$$

There are some basic properties of the maps Γ and Φ that we follow directly from the definitions. First, if E is an intermediate field, then $E \subseteq \Phi\Gamma(E) = \mathrm{Fix}\,\mathrm{Gal}(K/E)$; this is clear since every element of $\mathrm{Gal}(K/E)$ fixes E at least (but there could be a larger field of elements fixed by $\mathrm{Gal}(K/E)$). Similarly, if H is a subgroup of $G = \mathrm{Gal}(K/F)$ then $H \subseteq \Gamma\Phi(H) = \mathrm{Gal}(K/\mathrm{Fix}(H))$; because certainly H is contained in the set of those elements of G that fix everything in $\mathrm{Fix}(H)$ (though there could be more elements that do). Second, both Φ and Γ are *inclusion reversing*. Namely, if $F \subseteq E \subseteq L \subseteq K$, then $\Gamma(L) = \mathrm{Gal}(K/L) \subseteq \Gamma(E) = \mathrm{Gal}(K/E)$, since an automorphism that fixes L pointwise certainly also fixes E. Similarly, if $H \subseteq J \subseteq G$ are subgroups of G, then $\Phi(J) = \mathrm{Fix}(J) \subseteq \Phi(H) = \mathrm{Fix}(H)$, since the elements fixed pointwise by everything in J are certainly also fixed by everything in H.

While the set up above in (18.14) makes sense for any field extension $F \subseteq K$, we will prove it has especially good properties when $[K : F] < \infty$ and K/F is Galois. In that case we will see shortly that Γ and Φ are inverse bijections of sets, so they will define a one-to-one (inclusion reversing) correspondence between intermediate fields and the subgroups of the Galois group.

Understanding the action of $\Phi\Gamma$ on intermediate subfields of a Galois extension is easy from what we have already done.

Lemma 18.15. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$ and K/F Galois. For every intermediate field E with $F \subseteq E \subseteq K$, we have $\Phi\Gamma(E) = \mathrm{Fix}\,\mathrm{Gal}(K/E) = E$.*

Proof. By Corollary 18.12, K/E is a Galois extension with $[K : E] < \infty$. Let $E' = \mathrm{Fix}\,\mathrm{Gal}(K/E)$, so $E \subseteq E' \subseteq K$. Now by the inclusion reversing property of $\mathrm{Gal}(K/-)$ we have $\mathrm{Gal}(K/E') \subseteq \mathrm{Gal}(K/E)$. On the other hand, if $\sigma \in \mathrm{Gal}(K/E)$ then by definition σ fixes pointwise everything in

$E' = \operatorname{Fix}\operatorname{Gal}(K/E)$, and so $\sigma \in \operatorname{Gal}(K/E')$. Hence $\operatorname{Gal}(K/E) = \operatorname{Gal}(K/E')$. Also, K/E' is a Galois extension by Corollary 18.12. It follows that $[K : E] = |\operatorname{Gal}(K/E)| = |\operatorname{Gal}(K/E')| = [K : E']$. But this forces $[E' : E] = 1$ and hence $E = E'$. □

An immediate corollary is the following version of the theorem of the primitive element.

Corollary 18.16. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$. If K/F is separable, then it has a primitive element; in other words $K = F(\gamma)$ for some $\gamma \in K$.*

Proof. Using Proposition 18.13, we can take a Galois closure L/F of K/F; so $F \subseteq K \subseteq L$ and L/F is Galois. Now by Lemma 18.15, for any $F \subseteq E \subseteq L$ we have $E = \operatorname{Fix}\operatorname{Gal}(K/E)$. If $G = \operatorname{Gal}(L/F)$ then G is finite and so has finitely many subgroups. Since every E is the fixed field of the subgroup $\operatorname{Gal}(K/E)$ of G, there are finitely many intermediate fields E. Since the extension L/F has finitely many intermediate fields, of course the smaller extension K/F also has this property. Now apply Theorem 17.57. □

We note that a finite degree inseparable extension might well have infinitely many intermediate fields.

To understand the action of $\Gamma\Phi$ on subgroups of $\operatorname{Gal}(K/F)$ we need one more new idea, which is a way of determining the minimal polynomial of an element using the action of the Galois group.

Lemma 18.17. *Let K/F be a finite degree Galois extension. Suppose that H is a subgroup of $\operatorname{Gal}(K/F)$ such that $\operatorname{Fix}(H) = F$.*

1. *For any $\alpha \in K$, let $\mathcal{O}_\alpha = \{\sigma(\alpha) | \sigma \in H\}$. Then $\operatorname{minpoly}_F(\alpha)$ is equal to $\prod_{\beta \in \mathcal{O}_\alpha}(x - \beta)$.*
2. *$H = \operatorname{Gal}(K/F)$.*

Proof. (1) Given any automorphism $\sigma \in \operatorname{Gal}(K/F)$, $\sigma : K \to K$ extends to an automorphism σ of $K[x]$ in the usual way, by acting on the coefficients of a polynomial. Apriori the polynomial $f = \prod_{\beta \in \mathcal{O}_\alpha}(x - \beta)$ just lies in $K[x]$, but we claim it is actually in $F[x]$.

Since $\mathcal{O}_\alpha$ is an orbit of the action of H on K, it is clear that for any $\sigma \in H$, if $\mathcal{O}_\alpha = \{\beta_1, \ldots, \beta_m\}$ then $\{\sigma(\beta_1), \ldots, \sigma(\beta_m)\} = \mathcal{O}_\alpha$ as well. Now we have $\sigma(f) = \prod_{\beta \in \mathcal{O}_\alpha}(x - \sigma(\beta)) = \prod_{\beta \in \mathcal{O}_\alpha}(x - \beta) = f$. This is true for all $\sigma \in H$. But that means that every coefficient of f is fixed by all $\sigma \in H$. In other words the coefficients of f lie in $\operatorname{Fix}(H) = F$ and so $f \in F[x]$ as claimed.

Since $\alpha \in \mathcal{O}_\alpha$ it is clear that $f(\alpha) = 0$. Let $g = \operatorname{minpoly}_F(\alpha)$. Then for each $\sigma \in H \subseteq \operatorname{Gal}(K/F)$, we know that σ permutes the roots of g in K. In particular $\sigma(\alpha)$ must be a root of g for all $\sigma \in H$. But now every element of $\mathcal{O}_\alpha$ is a root of g, and so $(x - \beta)$ is a factor of g for all $\beta \in \mathcal{O}_\alpha$. This forces $f | g$ and hence $f = g$ since g is irreducible.

(2) By Corollary 18.16, we know there is $\gamma \in K$ such that $K = F(\gamma)$. Now $|\operatorname{Gal}(K/F)| = [K : F] = \deg \operatorname{minpoly}_F(\gamma)$. By part (1), $\deg \operatorname{minpoly}_F(\gamma) = |\mathcal{O}_\gamma| \leq |H|$. This shows that $|\operatorname{Gal}(K/F)| \leq |H|$ and since H is a subgroup of $\operatorname{Gal}(K/F)$ we have $H = \operatorname{Gal}(K/F)$. □

Corollary 18.18. *Let $G = \operatorname{Gal}(K/F)$ for a finite degree Galois extension K/F. If H is a subgroup of G then $\Gamma\Phi(H) = \operatorname{Gal}(K/\operatorname{Fix}(H)) = H$.*

Proof. Since K/F is Galois, we know that $K/\operatorname{Fix}(H)$ is Galois by Corollary 18.12. Now applying Lemma 18.17 to the subgroup H and the extension $K/\operatorname{Fix}(H)$ yields that $H = \operatorname{Gal}(K/\operatorname{Fix}(H))$. □

We also now get an additional characterization of Galois extensions to add to those we found in Theorem 18.10. In fact, the property in the next theorem is sometimes taken to be the definition of a Galois extension.

Theorem 18.19. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$. Then K/F is Galois if and only if $F = \operatorname{Fix}(\operatorname{Gal}(K/F))$.*

Proof. If K/F is Galois, then we saw that $F = \operatorname{Fix}(\operatorname{Gal}(K/F))$ in Lemma 18.15. On the other hand, if $F = \operatorname{Fix}(\operatorname{Gal}(K/F))$ then Lemma 18.17(1) applies with $H = \operatorname{Gal}(K/F)$. For any $\alpha \in K$ the minimal polynomial $\operatorname{minpoly}_F(\alpha)$ calculated there clearly has distinct roots and splits over K. This shows that K/F is normal and separable, and hence it is Galois by Theorem 18.10. □

We now have all of the ingredients we need to prove the fundamental theorem.

Theorem 18.20 (Fundamental Theorem of Galois Theory)**.** *Let $F \subseteq K$ be a finite degree Galois extension of fields. Let $G = \operatorname{Gal}(K/F)$.*

(1) *The functions Γ, Φ defined as in (18.14) by*

$$\left\{\text{intermediate fields } E \text{ with } F \subseteq E \subseteq K\right\} \underset{\Phi=\operatorname{Fix}(-)}{\overset{\Gamma=\operatorname{Gal}(K/-)}{\rightleftarrows}} \left\{\text{subgroups } H \text{ of } G = \operatorname{Gal}(K/F)\right\}$$

are inverse inclusion-reversing bijections between the indicated sets.

(2) *For $F \subseteq E \subseteq K$, we have $[K : E] = |\operatorname{Gal}(K/E)|$ and $[E : F] = |G : \operatorname{Gal}(K/E)|$.*

(3) *For $F \subseteq E \subseteq K$, E/F is normal (and hence Galois) if and only if $H = \operatorname{Gal}(K/E)$ is a normal subgroup of G, and in this case $\operatorname{Gal}(E/F) \cong G/H$ as groups.*

Proof. (1) We have done almost all of the work needed for the proof in the preceding lemmas. We saw that Γ and Φ make sense for a general field extension and that they always reverse inclusions. Now using that K/F is finite degree Galois, by Lemma 18.15 we have $\Phi\Gamma$ is the identity function on

intermediate fields. By Lemma 18.18, $\Gamma\Phi$ is the identity function on subgroups of $G = \mathrm{Gal}(K/F)$. Thus Φ and Γ are inverse bijections.

(2) We also saw in Corollary 18.12 that for any intermediate field E, K/E is also Galois. Thus by definition $[K : E] = |\mathrm{Gal}(K/E)|$. Of course we have in particular that $[K : F] = |G| = |\mathrm{Gal}(K/F)|$. Now we have $[E : F] = [K : F]/[K : E] = |G|/|\mathrm{Gal}(K/E)| = |G : \mathrm{Gal}(K/E)|$.

(3) Since K/F is Galois, it is separable and so E/F is separable. This is why if E/F is normal it is automatically Galois as commented.

Let $H = \mathrm{Gal}(K/E)$. It is easy to see that the conjugate $\sigma H \sigma^{-1}$ is equal to $\mathrm{Gal}(K/\sigma(E))$. We see that H is normal in G if and only if $\mathrm{Gal}(K/\sigma(E)) = \mathrm{Gal}(K/E)$ for all $\sigma \in G$. But since Γ is a bijection this is if and only if $\sigma(E) = E$ for all $\sigma \in G$.

Now we claim that $\sigma(E) = E$ for all $\sigma \in G$ if and only if E/F is a normal extension. If E/F is normal and $\sigma \in G$, then for any $\alpha \in E$, $f = \mathrm{minpoly}_F(\alpha)$ splits in $E[x]$, so every root of f in K is already in E. On the other hand, σ must permute the roots of f, so $\sigma(\alpha) \in E$ and thus $\sigma(E) \subseteq E$; applying this argument with σ^{-1} yields $\sigma^{-1}(E) \subseteq E$ and so $E \subseteq \sigma(E)$; thus $\sigma(E) = E$ for all $\sigma \in G$. Conversely, suppose that $\sigma(E) = E$ for all $\sigma \in G$. For $\alpha_1 \in E$, let $f = \mathrm{minpoly}_F(\alpha_1)$. Now f splits in $K[x]$, say with roots $\alpha_1, \ldots, \alpha_m \in K$. By Corollary 17.37, for any i there is $\sigma \in \mathrm{Gal}(K/F)$ such that $\sigma(\alpha_1) = \alpha_i$. By hypothesis since $\alpha_1 \in E$, $\alpha_i \in \sigma(E) = E$. So f splits over E and E/F is normal. This proves the claim.

We have seen that E/F is a normal extension if and only if $H = \mathrm{Gal}(K/E)$ is a normal subgroup of G. Now assume this is the case. The map

$$\phi : \mathrm{Gal}(K/F) \longrightarrow \mathrm{Gal}(E/F)$$
$$\sigma \longmapsto \sigma|_E$$

is well defined because $\sigma(E) = E$ for all $\sigma \in G$. It is easy to see that ϕ is a homomorphism of groups since it is simply defined by restriction. Also, $\ker \phi = H = \mathrm{Gal}(K/E)$ by definition. Then $G/H \cong \phi(G)$ as groups by the 1st isomorphism theorem. The orders satisfy

$$|\phi(G)| = |G|/|H| = [G : \mathrm{Gal}(K/E)] = [E : F] = |\mathrm{Gal}(E/F)|$$

since E/F is Galois, and this forces ϕ to be surjective, so $G/H \cong \mathrm{Gal}(E/F)$. □

18.3. **Examples of the fundamental theorem.**

Example 18.21. Let K be the splitting field over $\mathbb{Q}$ of $f = x^3 - 2 \in \mathbb{Q}[x]$. We have seen that $[K : \mathbb{Q}] = 6$. since K is the splitting field of a separable polynomial, $K/\mathbb{Q}$ is Galois and so

$|\operatorname{Gal}(K/\mathbb{Q})| = 6$. In fact, in Example 17.38 we already constructed 6 different automorphisms, but let us revisit this yet again from another perspective.

We have $K = \mathbb{Q}(\alpha, \zeta)$, where $\alpha = \sqrt[3]{2}$ and $\zeta = e^{2\pi i/3}$. Any $\sigma \in G = \operatorname{Gal}(K/\mathbb{Q})$ must permute the roots $\{\zeta, \zeta^2\}$ of $x^2 + x + 1$ and the the roots $\{\alpha, \alpha\zeta, \alpha\zeta^2\}$ of $x^3 - 2$. Moreover, since α and ζ generate K over $\mathbb{Q}$, any automorphism in G is determined by its action on α and ζ. Since there are 3 possible things to send α to and 2 possible things to send ζ to, there are at most 6 different automorphisms in G. Since we know $|G| = 6$, all of these possibilities occur. Thus by applying our general results which tell us that $K/\mathbb{Q}$ is Galois in advance, we do not have to directly construct these automorphisms using the theory of splitting fields, as in Example 17.38.

In particular, there is an automorphism σ with $\sigma(\alpha) = \alpha\zeta$ and $\sigma(\zeta) = \zeta$, and an automorphism ρ with $\rho(\alpha) = \alpha$ and $\rho(\zeta) = \zeta^2$. Then $\sigma\rho(\alpha) = \sigma(\alpha) = \alpha\zeta$ while $\rho\sigma(\alpha) = \rho(\alpha\zeta) = \alpha\zeta^2$. It follows that $\sigma\rho \neq \rho\sigma$ and hence G is non-abelian. The only non-abelian group of order 6 is S_3, so $G \cong S_3$.

Now σ clearly has order 3, so $\langle\sigma\rangle$ is the unique subgroup of order 3 in G. Since this group has index 2 in G, the corresponding intermediate field $\operatorname{Fix}\langle\sigma\rangle$ has degree 2 over $\mathbb{Q}$. Clearly this must be $\operatorname{Fix}\langle\sigma\rangle = \mathbb{Q}(\zeta)$.

The elements of order 2 in the group are then $\rho, \sigma\rho$, and $\sigma^2\rho$. The corresponding fixed fields must have degree 3 over $\mathbb{Q}$. It is now easy to check that $\operatorname{Fix}\langle\rho\rangle = \mathbb{Q}(\alpha)$, $\operatorname{Fix}\langle\sigma\rho\rangle = \mathbb{Q}(\alpha\zeta^2)$, and $\operatorname{Fix}\langle\sigma\rho\rangle = \mathbb{Q}(\alpha\zeta)$. The fields we have found must be all of the intermediate fields strictly between $\mathbb{Q}$ and K, by the fundamental theorem.

Now let us work through a more elaborate example of the fundamental theorem.

Example 18.22. Consider the splitting field of $f = x^4 - 2$ over $\mathbb{Q}$. The 4th roots of 1 in $\mathbb{C}$ are $\{\pm 1, \pm i\}$. Let $\alpha = \sqrt[4]{2}$ be the positive real 4th root of 2. Then the roots of f in $\mathbb{C}$ are $\{\alpha, \alpha i, -\alpha, -\alpha i\}$. It is clear that the splitting field K of f is equal to $\mathbb{Q}(\alpha, i)$.

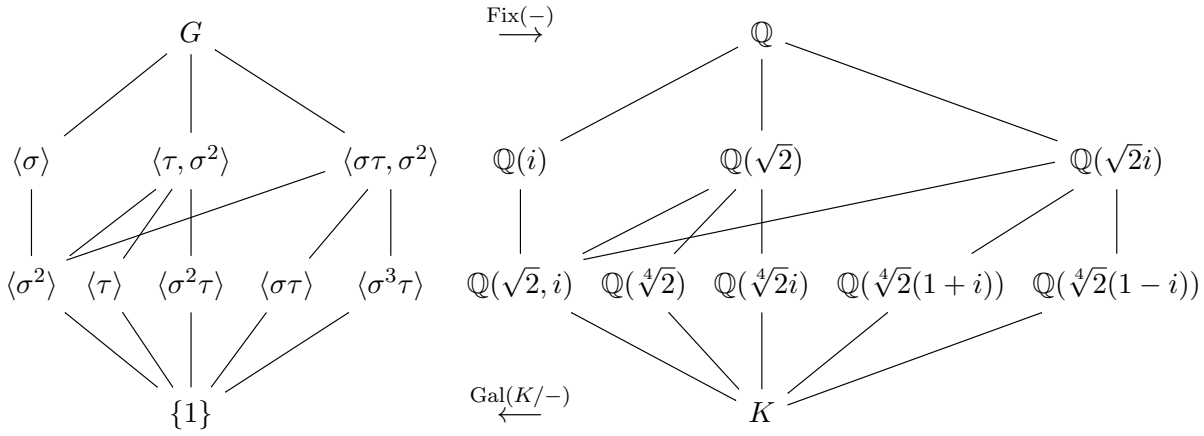
Now $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$, since f is irreducible by Eisenstein and thus $f = \operatorname{minpoly}_{\mathbb{Q}}(\alpha)$. Since $\mathbb{Q}(\alpha) \subseteq \mathbb{R}$, $\mathbb{Q}(\alpha) \neq K$; since i is a root of the degree 2 polynomial $x^2 + 1 \in \mathbb{Q}[x]$, the only possibility is $[\mathbb{Q}(\alpha, i) : \mathbb{Q}(\alpha)] = 2$ and so $[K : \mathbb{Q}] = 8$.

$K/\mathbb{Q}$ is certainly a Galois extension, being the splitting field of a separable polynomial, so $|\operatorname{Gal}(K/\mathbb{Q})| = 8$. Let $G = \operatorname{Gal}(K/\mathbb{Q})$ and let us determine the isomorphism type of G. If $\sigma \in G$, then σ sends α to one of the 4 roots of f and sends i to one of the 2 roots of $\operatorname{minpoly}_{\mathbb{Q}}(i) = x^2 + 1$. Since α and i generate K over $\mathbb{Q}$, any $\sigma \in G$ is determined by where it sends α and i. Since there are only $(4)(2) = 8$ possibilities they must all occur.

Let us call σ the automorphism in G with $\sigma(\alpha) = \alpha i$ and $\sigma(i) = i$. We let τ be the automorphism in G with $\tau(\alpha) = \alpha$ and $\tau(i) = -i$. Now it is easy to see that $|\sigma| = 4$ and $|\tau| = 2$. It is clear that $\langle\sigma\rangle \cap \langle\tau\rangle = \{1\}$ and so we must have $\langle\sigma\rangle\langle\tau\rangle = G$ as this product has order 8. Thus $G = \{1, \sigma, \sigma^2, \sigma^3, \tau, \sigma\tau, \sigma^2\tau, \sigma^3\tau\}$.

Now one easily calculates that $\tau\sigma = \sigma^3\tau = \sigma^{-1}\tau$. We recognize from this relation that G is isomorphic to the dihedral group D_8. From our knowledge of D_8 we can write down all of the subgroups of G. Of course we have the trivial subgroup and all of G; there is the rotation subgroup $\langle\sigma\rangle$ of order 4, which has a subgroup $\langle\sigma^2\rangle$ of order 2; each reflection generates a subgroup of order 2, and these are the subgroups $\langle\tau\rangle$, $\langle\sigma\tau\rangle$, $\langle\sigma^2\tau\rangle$, and $\langle\sigma^3\tau\rangle$. Any missing subgroups have order 4 and so must intersect $\langle\sigma\rangle$ in a subgroup of order 2, thus containing σ^2. This leads to two further subgroups $\langle\sigma^2, \tau\rangle$ and $\langle\sigma^2, \sigma\tau\rangle$ which are isomorphic to the Klein 4-group.

We display the lattice diagram of these subgroups on the left below, with a line drawn when one is included in the other. For each subgroup H we write the subfield $\mathrm{Fix}(H)$ on the right in the same position; the resulting diagram of all intermediate fields E looks the same, but due to the inclusion-reversing nature of the correspondence, the larger fields are below the fields they contain. For each intermediate field we have written elements that generate that field over $\mathbb{Q}$.



The verification that the fixed fields of each subgroup are what we have displayed is largely routine. For example, the subgroup $H = \langle\sigma\tau\rangle$ has order 2 and hence index 4 in G. Thus $E = \mathrm{Fix}(H)$ has degree 4 over $\mathbb{Q}$. To find elements in $\mathrm{Fix}(H)$, one method is to take any $\gamma \in K$ and note that $\sum_{\rho\in H} \rho(\gamma)$ is fixed by H. (This idea already appeared in the proof of Lemma 18.17.) Applying this to $\alpha = \sqrt[4]{2}$ gives that $\sqrt[4]{2} + \sqrt[4]{2}i = \sqrt[4]{2}(1 + i) \in \mathrm{Fix}(\langle\sigma\tau\rangle)$. On the other hand, $(\sqrt[4]{2}(1 + i))^4 = 2(-4) = -8$ and so $\sqrt[4]{2}(1 + i)$ is a root of $x^4 + 8$. One may check that this polynomial is irreducible

over $\mathbb{Q}$, and so $[\mathbb{Q}(\sqrt[4]{2}(1+i)) : \mathbb{Q}] = 4$. Thus $\mathbb{Q}(\sqrt[4]{2}(1+i)) = \mathrm{Fix}(H)$. We leave the other verifications to the reader.

We know by the theorem of the primitive element that every intermediate field can be generated by one element over $\mathbb{Q}$; for most of the intermediate fields above this has already been done. Suppose we want to write $K = \mathbb{Q}(\sqrt[4]{2}, i)$ in the form $\mathbb{Q}(\gamma)$ for some γ. It suffices to choose a γ which is not contained in any of the 5 displayed fields that have index 4 over $\mathbb{Q}$. This could be done by showing it is not fixed by any of the 5 elements generating order 2 subgroups of G. For example, $\sigma\tau(\sqrt[4]{2} + i) = \sqrt[4]{2}i - i = (\sqrt[4]{2} - 1)i$, which has 0 real part and so certainly is not equal to $\sqrt[4]{2} + i$. Similarly, σ^2, τ, $\sigma^2\tau$, and $\sigma^3\tau$ do not fix $\sqrt[4]{2} + i$, so $K = \mathbb{Q}(\sqrt[4]{2} + i)$.

We may also easily determine which subfields are normal and hence Galois over $\mathbb{Q}$. The normal subgroups of G include the three subgroups of index 2 and the subgroup $\langle\sigma^2\rangle$ (which is actually the center of D_8). The other 4 subgroups of order 2 generated by the reflections are not normal. Thus the intermediate fields which are normal over $\mathbb{Q}$ are $\mathbb{Q}(i)$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{2}i)$ and $\mathbb{Q}(\sqrt{2}, i)$. It is also easy to see directly that these are all splitting fields over $\mathbb{Q}$. On the the other hand, for example, $\mathbb{Q}(\sqrt[4]{2})$ is not normal over $\mathbb{Q}$ since it contains one root of $x^4 - 2$ but this polynomial does not split over that field. Similarly, $\mathbb{Q}(\sqrt[4]{2}(1 + i))$ contains one root of $x^4 + 8$ but that polynomial does not split over it. In fact, it is easy to see that K is also the splitting field over $\mathbb{Q}$ of $x^4 + 8$.

Finally, for any field E which is normal over $\mathbb{Q}$, the fundamental theorem tells us that $\mathrm{Gal}(E/\mathbb{Q}) \cong \mathrm{Gal}(K/\mathbb{Q})/\mathrm{Gal}(K/E)$. For instance, taking $E = \mathbb{Q}(\sqrt{2}, i)$ we must have $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, i)/\mathbb{Q}) \cong D_8/\langle\sigma^2\rangle$. It is also easy to see directly that both sides are Klein 4-groups.

We observe that it was easy to compute the subgroups of the Galois group above, and relatively easy then to find the fixed fields. Without Galois theory it is extremely unclear how one would go about finding all of the intermediate fields in the extension, or even why there should be finitely many.

Let us give an example of a Galois group over $\mathbb{Q}$ where determining the group is a bit less straightforward.

Example 18.23. Let $\alpha = \sqrt{2 + \sqrt{2}}$. First let us calculate its minimal polynomial over $\mathbb{Q}$. We have $\alpha^2 = 2 + \sqrt{2}$ so $(\alpha^2 - 2)^2 = 2$. Then α is a root of $f = (x^2 - 2)^2 - 2 = x^4 - 4x^2 + 2$. This polynomial is irreducible by the Eisenstein criterion, so $f = \mathrm{minpoly}_{\mathbb{Q}}(\alpha)$. Then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$.

We claim that $K = \mathbb{Q}(\alpha)$ is already the splitting field of f over $\mathbb{Q}$. First, applying the quadratic formula to $x^2 - 4x + 2 = 0$ yields roots $2 \pm \sqrt{2}$. Thus $x^4 - 4x^2 + 2 = 0$ has roots $\pm\sqrt{2 \pm \sqrt{2}}$, where all of these roots are real. Write $\beta = \sqrt{2 - \sqrt{2}}$, so the roots are $\pm\alpha, \pm\beta$.

Now notice that $\sqrt{2-\sqrt{2}}\sqrt{2+\sqrt{2}} = \sqrt{(2-\sqrt{2})(2+\sqrt{2})} = \sqrt{4-2} = \sqrt{2}$. Moreover, $\sqrt{2} = \alpha^2 - 2$ is already in K. Hence $\beta = \sqrt{2-\sqrt{2}} \in K$ already and so all of the roots of f are in K. Thus K is the splitting field of f as claimed.

Let $G = \mathrm{Gal}(K/\mathbb{Q})$. Since K is a splitting field of a polynomial over $\mathbb{Q}$, $K/\mathbb{Q}$ is Galois and so $|G| = 4$. Either $G \cong \mathbb{Z}_4$ or $G \cong \mathbb{Z}_2 \times \mathbb{Z}_2$ and we would like to determine which occurs.

We know from our results on splitting fields that we can find $\sigma \in G$ that sends α to any other root of f. Let us choose $\sigma \in G$ with $\sigma(\alpha) = \beta$. How does σ act on the other roots of f? Well, $\sigma(\alpha^2) = \beta^2$. This says $\sigma(2+\sqrt{2}) = 2-\sqrt{2}$. Then clearly $\sigma(\sqrt{2}) = -\sqrt{2}$. Now since $\alpha\beta = \sqrt{2}$, $\sigma(\alpha\beta) = -\alpha\beta$ which implies $\sigma(\beta) = -\alpha$. We see from this that $\sigma^2(\alpha) = -\alpha$ and thus σ cannot have order 2. Hence σ has order 4 and thus G is cyclic of order 4.

Now G has only one proper subgroup, which is $\langle\sigma^2\rangle$. The corresponding intermediate field is $\mathbb{Q}(\sqrt{2}) = \mathrm{Fix}\langle\sigma^2\rangle$.

For any splitting field K of a separable polynomial f over F, it is possible to visualize the Galois group as a subgroup of the permutation group of the roots of f in K; indeed, this is how Galois originally thought about it. This is because any automorphism in $\mathrm{Gal}(K/F)$ must permute these roots and is determined by its action on these roots. For instance, in the previous example one notes that σ acts as a 4-cycle on the roots of $x^4 - 4x^2 + 2$, which is another way of seeing that it is cyclic of order 4.

18.4. **Cyclotomic extensions.** The set $\mu_n = \{e^{2m\pi i/n} | 0 \leq m \leq n-1\}$ consists of the n distinct nth roots of 1 in $\mathbb{C}$. This can also be described as the set of roots in $\mathbb{C}$ of $x^n - 1 \in \mathbb{Q}[x]$. The group μ_n is a subgroup of the multiplicative group $\mathbb{C}^\times$.

The map $(\mathbb{Z}_n, +) \to \mu_n$ given by $\overline{m} \mapsto e^{2m\pi i/n}$ is easily seen to be an isomorphism of groups. Thus μ_n is cyclic, and any generator of μ_n is called a *primitive* nth root. From our knowledge of cyclic groups we know that the generators of $\mathbb{Z}_n$ are $\{\overline{m} | \gcd(m,n) = 1\}$ and so the primitive nth roots are $\{e^{2m\pi i/n} | \gcd(m,n) = 1\}$. There are $\varphi(n)$ of them, where φ is the Euler φ-function.

Definition 18.24. Let $P_n = \{e^{2m\pi i/n} | \gcd(m,n) = 1\}$ be the set of primitive nth roots of 1. Let

$$\Phi_n(x) = \prod_{\alpha \in P_n} (x - \alpha) \in \mathbb{C}[x].$$

$\Phi_n(x)$ is called the nth *cyclotomic polynomial*. Clearly $\deg \Phi_n(x) = \varphi(n)$.

If $\alpha \in \mu_n$, then α has order d in $\mathbb{C}^\times$ for some $d|n$, and then α is a primitive dth root. Thus the following formula is clear.

Lemma 18.25. *Let $n \geq 1$. Then $x^n - 1 = \prod_{d|n} \Phi_d(x)$.*

Using this lemma we can compute the polynomials $\Phi_n(x)$ inductively. This is easy to do by hand if n is small.

Example 18.26. 1 is the only primitive 1st root, so $\Phi_1(x) = x - 1$. Similarly, -1 is the lone primitive 2nd root and $\Phi_2(x) = x + 1$. By Lemma 18.25 we have $x^3 - 1 = \Phi_3(x)\Phi_1(x)$ and so $\Phi_3(x) = (x^3 - 1)/(x - 1) = x^2 + x + 1$. Next,

$$\Phi_4(x) = (x^4 - 1)/(\Phi_2(x)\Phi_1(x)) = (x^4 - 1)/((x + 1)(x - 1)) = (x^4 - 1)/(x^2 - 1) = x^2 + 1.$$

Of course, $x^2 + 1 = (x - i)(x + i)$, and the roots of $\Phi_4(x)$ are the two primitive 4th roots of unity, i and $-i$. We leave it to the reader to check that $\Phi_6(x) = x^2 - x + 1$.

If p is prime then $\Phi_p(x) = (x^p - 1)/(x - 1) = x^{p-1} + \cdots + x + 1$. We proved this polynomial is irreducible in $\mathbb{Q}[x]$ using the Eisenstein criterion with substitution. In fact, all of the polynomials $\Phi_n(x)$ lie in $\mathbb{Z}[x]$ and are irreducible over $\mathbb{Q}$. We give the proof of irreducibility now, though as it is a bit technical the reader might wish to simply assume this fact and move on.

Theorem 18.27. *For any $n \geq 1$, $\Phi_n(x) \in \mathbb{Z}[x]$, it is monic, and it is irreducible in $\mathbb{Q}[x]$.*

Proof. By definition it is obvious that $\Phi_n(x)$ is monic. We have $x^n - 1 = \Phi_n(x) \prod_{d|n, d<n} \Phi_d(x)$ by Lemma 18.25. We prove that $\Phi_n(x) \in \mathbb{Z}[x]$ by induction on n. Thus we can assume that $\Phi_d(x) \in \mathbb{Z}[x]$, for all divisors d of n with $d < n$, by the induction hypothesis. Then $g = \prod_{d|n, d<n} \Phi_d(x)$ is also monic. By Gauss's lemma, there is $\lambda \in \mathbb{Q}$ such that $\lambda^{-1}\Phi_n(x)$ and λg are in $\mathbb{Z}[x]$. Since g and $\Phi_n(x)$ are monic, this forces λ and $\lambda^{-1} \in \mathbb{Z}$, so $\lambda = \pm 1$ and $\Phi_n(x) \in \mathbb{Z}[x]$ already.

Let f be one of the irreducible factors of $\Phi_n(x)$ in $\mathbb{Q}[x]$ and write $\Phi_n(x) = fg$. By Gauss's Lemma again, we can choose this factorization so f and g are in $\mathbb{Z}[x]$ and are monic. Now pick any prime p with $\gcd(p, n) = 1$. The idea of the proof is to consider this factorization of $\Phi_n(x)$ modulo p.

Let ζ be a primitive nth root of 1; since $\gcd(p, n) = 1$, we also have ζ^p is a primitive nth root. Thus ζ and ζ^p are roots of $\Phi_n(x)$ and each is a root of either f or g. Suppose that $f(\zeta) = 0$, while $g(\zeta^p) = 0$.

Now $g(\zeta^p) = 0$ means that ζ is a root of $g(x^p) \in \mathbb{Z}[x]$. Since f is irreducible, $f = \text{minpoly}_{\mathbb{Q}}(\zeta)$. Hence $f | g(x^p)$ in $\mathbb{Q}[x]$, say $g(x^p) = fh$. As above, h is monic and by Gauss's lemma we have $h \in \mathbb{Z}[x]$.

Let $\phi : \mathbb{Z}[x] \to (\mathbb{Z}/p\mathbb{Z})[x] = \mathbb{F}_p[x]$ be the reduction mod p homomorphism which sends each coefficient $a \in \mathbb{Z}$ to $\overline{a} = a + p\mathbb{Z}$. For $f \in \mathbb{Z}[x]$ write $\overline{f}$ for $\phi(f)$. Now applying ϕ we have $\overline{g}(x^p) = \overline{f}\,\overline{h}$. If we write $\overline{g}(x) = \sum_{i=0}^m a_i x^i$, with $a_i \in \mathbb{F}_p$, then since the pth power map is a ring homomorphism of $\mathbb{F}_p[x]$, we have

$$\overline{g}(x^p) = \sum_{i=0}^m a_i x^{ip} = \sum_{i=0}^m a_i^p x^{ip} = \sum_{i=0}^m (a_i x^i)^p = (\sum_{i=0}^m a_i x^i)^p = (\overline{g}(x))^p,$$

where we have used that $a^p = a$ for all $a \in \mathbb{F}_p$ by Fermat's little theorem.

We now see that $\overline{f} \,|\, (\overline{g})^p$ in $\mathbb{F}_p[x]$. This implies that $\overline{f}$ and $\overline{g}$ have a common irreducible factor in $\mathbb{F}_p[x]$. But since $\overline{\Phi_n} = \overline{f}\,\overline{g}$ this means that $\overline{\Phi_n}$ has a repeated irreducible factor in $\mathbb{F}_p[x]$, and so it is not a separable polynomial. On the other hand, $\overline{\Phi_n}$ divides $\overline{x^n - 1}$, which is a separable polynomial in $\mathbb{F}_p[x]$: its derivative is $\overline{n}x^{n-1} \neq 0$ (because $\gcd(p, n) = 1$), and so $\gcd(x^n - 1, \overline{n}x^{n-1}) = 1$ in $\mathbb{F}_p[x]$. This is a contradiction.

The contradiction implies that for all roots ζ of f and all primes p with $\gcd(p, n) = 1$, ζ^p must also be a root of f. Now if $\gcd(i, n) = 1$ for some integer i, then factorizing $i = p_1 p_2 \dots p_k$ where each p_j is prime, then $\gcd(p_j, n) = 1$ for all j and so by induction we get for any root ζ of f that ζ^i is also a root of f. However, any root ζ of f is by definition a generator of the group μ_n of nth roots of 1, and the other generators are equal to ζ^i for $0 < i < n$ with $\gcd(i, n) = 1$. Hence every primitive nth root of 1 is a root of f. This shows that $\Phi_n(x) = f$ and hence $\Phi_n(x)$ is irreducible over $\mathbb{Q}$. □

Theorem 18.28. *Let $n \geq 1$. Consider the nth cyclotomic field $K = \mathbb{Q}(\zeta)$ where ζ is a primitive nth root of 1. Then $[K : \mathbb{Q}] = \varphi(n)$, and $K/\mathbb{Q}$ is Galois with $\mathrm{Gal}(K/\mathbb{Q}) \cong \mathbb{Z}_n^*$, where $\mathbb{Z}_n^*$ is the multiplicative group of units mod n.*

Proof. We know that K is the splitting field of $x^n - 1$ over $\mathbb{Q}$ by Example 17.31. We also know from Theorem 18.27 that $\Phi_n(x)$ is irreducible over $\mathbb{Q}$; hence $\Phi_n(x) = \mathrm{minpoly}_{\mathbb{Q}}(\zeta)$, which implies $[\mathbb{Q}(\zeta) : \mathbb{Q}] = \deg \Phi_n(x) = \varphi(n)$. $K/\mathbb{Q}$ is Galois since it is a splitting field of a separable polynomial.

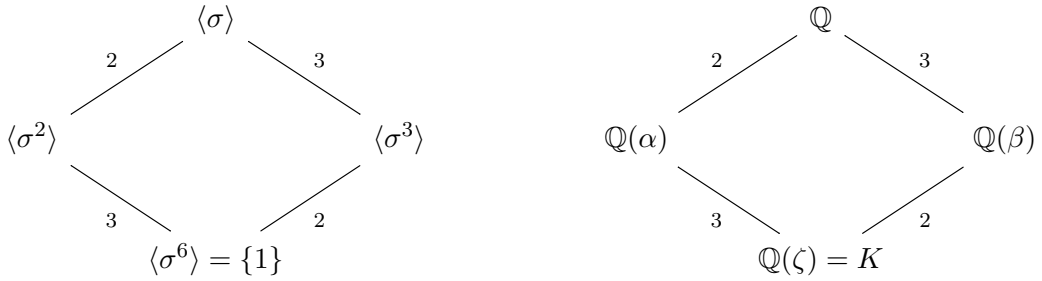
Now any $\sigma \in G = \mathrm{Gal}(K/\mathbb{Q})$ is determined by its action on ζ, and $\sigma(\zeta)$ must be another root of $\Phi_n(x)$. Since this polynomial has $\varphi(n)$ roots and $|G| = \varphi(n)$, all possibilities occur. For each $0 \leq i \leq n-1$ with $\gcd(i, n) = 1$, we have an automorphism $\sigma_i \in G$ where $\sigma_i(\zeta) = \zeta^i$, so $G = \{\sigma_i \,|\, 0 \leq i \leq n-1,\ \gcd(i, n) = 1\}$. Now define a function $\phi : G \to \mathbb{Z}_n^*$ by $\phi(\sigma_i) = \overline{i}$. Since $\sigma_i\sigma_j(\zeta) = \sigma_i(\zeta^j) = \sigma_i(\zeta)^j = (\zeta^i)^j = \zeta^{ij} = \zeta^m$, where m is unique integer with $0 \leq m \leq n-1$ such that $m \equiv ij \mod n$, we have $\sigma_i\sigma_j = \sigma_m$ where $\overline{ij} = \overline{m}$. This implies that ϕ is a homomorphism of groups, and it is clearly bijective. □

Recall from our study of groups that the structure of the group $\mathbb{Z}_n^*$ is well-understood. First, if $n = p_1^{e_1} \dots p_m^{e_m}$ for distinct primes p_i, then $\mathbb{Z}_n^* \cong \prod_i \mathbb{Z}_{p_i^{e_i}}^*$. If p is an odd prime, then $\mathbb{Z}_{p^e}^*$ is cyclic of order $\varphi(p^e) = p^{e-1}(p-1)$; while $\mathbb{Z}_{2^e}^* \cong \mathbb{Z}_2 \times \mathbb{Z}_{2^{e-2}}$.

Example 18.29. Let $n = 9$ and consider the splitting field K of $x^n - 9$ over $\mathbb{Q}$. We know that $[K : Q] = \varphi(9) = 6$ and if $G = \mathrm{Gal}(K/\mathbb{Q})$ then $G \cong \mathbb{Z}_9^*$, which is cyclic order 6.

Let ζ be a primitive 9th root of unity, so $K = \mathbb{Q}(\zeta)$. One may check that $\overline{2}$ is a generator of $\mathbb{Z}_9^*$. It follows that the automorphism $\sigma \in G$ with $\sigma(\zeta) = \zeta^2$ generates G. So $G = \langle \sigma \rangle$, and by the stucture of cyclic groups of order 6, the subgroups of G are $\langle \sigma^2 \rangle$, $\langle \sigma^3 \rangle$, and the trivial subgroup.

We have a diagram of subgroups and and corresponding diagram of fixed subfields as follows, where the numbers indicate the index of one subgroup in another (on the left) or the degree of one subfield over another (on the right).



Let us find elements α and β which generate the indicated extensions on the right. Note that since ζ is a primitive 9th root, ζ^3 is a primitive 3rd root. Thus K also contains the cyclotomic extension $\mathbb{Q}(\zeta^3)$, where the minimal polynomial of ζ^3 is $x^2 + x + 1$. Thus $\mathbb{Q}(\zeta^3)$ is a field of degree 2 over $\mathbb{Q}$, so we can take $\alpha = \zeta^3$ in the picture, as there is only one subfield of degree 2 over $\mathbb{Q}$. For the other extension we note that $\zeta + \sigma^3(\zeta) = \zeta + \zeta^8 = \zeta + \zeta^{-1}$ is fixed by σ^3, so $\zeta + \zeta^{-1} \in \mathrm{Fix}(\langle \sigma^3 \rangle)$. $\zeta + \zeta^{-1}$ is not in $\mathbb{Q}$ (for if it was, $\zeta + \zeta^{-1} = q \in \mathbb{Q}$ would give $\zeta^2 + 1 = q\zeta$ and hence ζ would satisfy a degree 2 polynomial in $\mathbb{Q}[x]$, while we know that $\deg \mathrm{minpoly}_{\mathbb{Q}}(\zeta) = 6$). Thus we can take $\beta = \zeta + \zeta^{-1}$ in the picture.

18.5. **More on finite fields.** Recall that we have seen that there is a unique field $\mathbb{F}_{p^n}$ with p^n elements up to isomorphism, for any prime p and $n \geq 1$. It can be defined as the splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$. We can now easily calculate the Galois group of this field as an extension of $\mathbb{F}_p$.

Theorem 18.30. *Let K be a field with $|K| = p^n$ for a prime p and $n \geq 1$. Then $\mathbb{F}_p \subseteq K$ and $K/\mathbb{F}_p$ is Galois with $[K : \mathbb{F}_p] = n$ and $G = \mathrm{Aut}(K) = \mathrm{Gal}(K/\mathbb{F}_p) \cong \mathbb{Z}_n$. In particular, G is generated as a group by the Frobenius automorphism $\sigma : K \to K$ given by $\sigma(a) = a^p$.*

Proof. We know that K has characteristic p, and we have seen that the additive subgroup of K generated by 1 is a copy of the field $\mathbb{F}_p$. Any automorphism of K sends 1 to itself and so will fix the elements in $\mathbb{F}_p$, which are sums of 1. Thus $\mathrm{Aut}(K) = \mathrm{Gal}(K/\mathbb{F}_p)$.

We know that $K^\times$ is a cyclic group of order $p^n - 1$. Let $\gamma \in K$ be a generator of this group. Then $\gamma^{p^n-1} = 1$ and so $\gamma^{p^n} = \gamma$. Conversely, if $\gamma^j = \gamma$ for some $j > 1$ then $\gamma^{j-1} = 1$ and hence $j \geq p^n$ since γ has order $p^n - 1$.

Now the Frobenius map σ is an automorphism of K, as in Example 17.50. We have $\sigma^i(\gamma) = \gamma^{p^i}$ and this cannot equal γ for $0 < i < n$; so $\sigma^i \neq 1_K$. On the other hand $\sigma^n(\gamma) = \gamma$ and since $K = \mathbb{F}_p(\gamma)$, $\sigma^n = 1$. It follows that σ is an element of order n in G. Since $|G| = [K : \mathbb{F}_p] = n$, the Frobenius map σ must generate G and hence $G \cong (\mathbb{Z}_n, +)$. □

Corollary 18.31. *Let K be a field with $|K| = p^n$ for some prime p and $n \geq 1$. Then K has a unique subfield with p^d elements for each positive divisor d of n. These are the only subfields of K.*

Proof. $G = \mathrm{Gal}(K/\mathbb{F}_p) = \langle\sigma\rangle$, where σ, the Frobenius map, has order n in G. Since G is cyclic of order n, for each divisor d of n there is a unique subgroup $H_d = \langle\sigma^d\rangle$ with $[G : H_d] = d$. Then the fields E_d where $\mathbb{F}_p \subseteq E_d = \mathrm{Fix}(H_d) \subseteq K$ are the only intermediate fields of the extension $K/\mathbb{F}_p$, where $[E_d : \mathbb{F}_p] = d$ and hence $|E_d| = p^d$. In fact these E_d are all of the subfields of K, because every subfield of K must contain the prime subfield $\mathbb{F}_p$. There is one for each divisor d of n. □

We can describe the subfield E_d of order p^d inside a field K of order p^n more explicitly: Since E_d has order p^d, all $a \in E_d$ must satisfy $a^{p^d} = a$, because we saw in our original study of finite fields that the elements of E_d are all roots of $x^{p^d} - x$. Since that polynomial only has p^d roots, the elements in E must be all of its roots. Thus $E = \{a \in K | a^{p^d} = a\}$.

One elegant consequence of our results so far is the following description of the factorization of $x^{p^n} - x$ over $\mathbb{F}_p$.

Proposition 18.32. *$x^{p^n} - x \in \mathbb{F}_p[x]$ is the product of all monic irreducible polynomials of degree d over $\mathbb{F}_p$, for all divisors d of n.*

Proof. If $f \in \mathbb{F}_p[x]$ is monic and irreducible of degree d, where $d|n$, then $K = \mathbb{F}_p[x]/(f)$ is a field with p^d elements. We know then by Corollary 18.31 that the field $\mathbb{F}_{p^n}$ has a subfield isomorphic to this field. Since every element of $\mathbb{F}_{p^n}$ satisfies $a^{p^n} = a$, the same must be true of the elements of K. In particular, $(x + (f))^{p^n} = x^{p^n} + (f) = x + (f)$ in K, which means $x^{p^n} - x \in (f)$. In other words, f divides $x^{p^n} - x$.

Conversely, if $g \in \mathbb{F}_p[x]$ is any monic irreducible factor of $x^{p^n} - x$, then we know that g splits over $\mathbb{F}_{p^n}$ since this field is the splitting field of $x^{p^n} - x$. If $\alpha \in \mathbb{F}_{p^n}$ is a root of g, then $\mathbb{F}_p(\alpha) \subseteq \mathbb{F}_{p^n}$ where $[\mathbb{F}_p(\alpha) : \mathbb{F}_p] = \deg g$ because $g = \text{minpoly}_{\mathbb{F}_p}(\alpha)$. It follows that $|\mathbb{F}_p(\alpha)| = p^{\deg g}$. Since we have seen that all subfields of a field with p^n elements have p^d elements for some divisor d of n, $\deg g = d$ for some divisor d of n.

Finally, we know that $x^{p^n} - x$ is separable over $\mathbb{F}_p$, as we showed that its roots in $\mathbb{F}_{p^n}$ are the p^n distinct elements of $\mathbb{F}_{p^n}$. It follows that $x^{p^n} - x$ is a product of distinct irreducible polynomials. By the arguments above the irreducibles occurring are exactly those of degree d where $d|n$. $\square$

The proposition can be used to give a precise count of the number of irreducible polynomials of each degree n over $\mathbb{F}_p$, by induction. We omit the exact formula here, but demonstrate the idea in an example.

Example 18.33. Consider $x^{81} - x \in \mathbb{F}_3[x]$. Here $81 = 3^4$. By the proposition, $x^{81} - x$ is the product of all monic irreducible polynomials in $\mathbb{F}_3[x]$ of degree $1, 2$, or 4. We know the degree 1 monic polynomials are $(x-1)$, $(x-2)$, and $(x-4)$. The degree 2 monic irreducibles are those without a root in $\mathbb{F}_3$; these are x^2+1, x^2+2x+2, and x^2+x+2 by direct calculation. The product of these 6 polynomials has degree 9. That means there is a polynomial of degree $81 - 9 = 72$ left over in the factorization of $x^{81} - x$, which is a product of all distinct monic degree 4 irreducibles. There are thus $72/4 = 18$ distinct such irreducibles over $\mathbb{F}_3$.

18.6. **Root Extensions.** A very common kind of field extension is "adding an nth root of an existing element". We have seen many examples of this already. In other words, one has a field F and an element $a \in F$ such that $f = x^n - a \in F[x]$ does not split already over F. In a splitting field K for f there is a root $\alpha \in K$ of f and we can consider the extension $F \subseteq F(\alpha)$ inside K. Since $\alpha^n = a \in F$, we think of α as an nth root of a, and might loosely write $\alpha = \sqrt[n]{a}$, though this notation is not uniquely defined, as a may have as many as n different nth roots. We are going to see in the next results that extensions of this kind are closely related to cyclic Galois groups. Technically, extensions of this form are simplest when $x^n - 1$ already splits in the base field with distinct roots, and we will concentrate on that case.

We first see that adding an nth root gives a cyclic Galois group (when the base field already has enough roots of 1).

Proposition 18.34. *Let F be a field such that $x^n - 1 \in F[x]$ splits with distinct roots in F. Suppose that $F \subseteq K$ is a field extension and $\alpha \in K$ is a root of $f = x^n - a \in F[x]$. Then $F(\alpha)/F$ is a Galois extension and $\text{Gal}(F(\alpha)/F)$ is cyclic of order d for some divisor d of n.*

Proof. The set of roots of $x^n - 1$ in F is a finite multiplicative subgroup of $F^\times$. Since we are assuming this polyonomial splits with distinct roots in F, this is a subgroup of order n. We have seen that a finite subgroup of the multplicative group of a field is always cyclic in Corollary 17.53, so this group is cyclic generated by some ζ, say; thus $\{1, \zeta, \zeta^2 \ldots, \zeta^{n-1}\}$ is the set of roots of $x^n - 1$ in F.

Now $\alpha^n = a$ and so clearly $(\alpha\zeta^i)^n = a$ also for all i; thus $x^n - a$ has the n distinct roots $\{\alpha, \alpha\zeta, \ldots, \alpha\zeta^{n-1}\}$ in K. Since by assumption $\zeta \in F$, we see that all of these roots are contained in $F(\alpha)$ already. Thus $F(\alpha)$ is the splitting field of f over F; and f has distinct roots and so is a separable polynomial. Thus $F(\alpha)/F$ is a Galois extension.

Now consider any $\sigma \in G = \mathrm{Gal}(F(\alpha)/F)$. Since σ permutes the roots of $x^n - a \in F[x]$, we have $\sigma(\alpha) = \alpha\zeta^i$. This allows us to define a map $\phi : G \to (\mathbb{Z}_n, +)$ defined by $\phi(\sigma) = \bar{i}$, where i is any integer such that $\sigma(\alpha) = \alpha\zeta^i$. Note that if we express $\sigma(\alpha)$ as $\alpha\zeta^j$ for some other j, we will have $\zeta^i = \zeta^j$ and thus since ζ has order n, $\bar{i} = \bar{j} \in \mathbb{Z}_n$, so that ϕ is well-defined.

Now to see that ϕ is a homomorphism, we just note that if $\sigma, \tau \in G$, with $\sigma(\alpha) = \alpha\zeta^i$ and $\tau(\alpha) = \alpha\zeta^j$, then

$$\tau\sigma(\alpha) = \tau(\alpha\zeta^i) = \tau(\alpha)\tau(\zeta)^i = \tau(\alpha)\zeta^i = \alpha\zeta^j\zeta^i = \alpha\zeta^{i+j},$$

where we use that $\zeta \in F$ and that $\tau \in \mathrm{Gal}(F(\alpha)/F)$ fixes F. This shows that $\phi(\tau\sigma) = \overline{i+j} = \bar{i} + \bar{j} = \phi(\tau) + \phi(\sigma)$ and thus ϕ is a homomorphism.

Since any $\sigma \in G$ is determined by where it sends α, it follows that ϕ is injective. Thus G is isomorphic to a subgroup of the cyclic group $\mathbb{Z}_n$, and thus from our classification of subgroups of cyclic groups, we conclude that G is cyclic of order d for some divisor d of n. □

Example 18.35. One really can get a proper subgroup of $\mathbb{Z}_n$ in the previous theorem, because there is no requirement that $f = x^n - a$ be irreducible over F. (If f does happen to be irreducible, then $[F(\alpha) : F] = n = |G|$ and this does force $G \cong \mathbb{Z}_n$).

Here is an explicit example. Let K be the splitting field of $f = x^8 - 2$ over $\mathbb{Q}$. Let ζ be a primitive 8th root of 1 in $\mathbb{C}$, and let $\alpha = \sqrt[8]{2}$ be the positive 8th root of 2. Then the roots of f in $\mathbb{C}$ are $\{\alpha, \alpha\zeta, \ldots, \alpha\zeta^7\}$ and $K = \mathbb{Q}(\zeta, \alpha)$.

Of course $\mathbb{Q}$ does not contain 8 distinct roots of $x^8 - 1$, but $F = \mathbb{Q}(\zeta)$ does, and so Proposition 18.34 applies to the extension $F \subseteq F(\alpha) = \mathbb{Q}(\zeta, \alpha) = K$. By that proposition, $G = \mathrm{Gal}(K/F)$ is cyclic of order dividing 8.

Let us now calculate $[K : \mathbb{Q}]$. Explicitly we have $\zeta = e^{2\pi i/8} = \sqrt{2}/2 + (\sqrt{2}/2)i \subseteq \mathbb{Q}(\sqrt{2}, i)$, and it is easy to see that $[\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}] = 4$. On the other hand, we know that $[\mathbb{Q}(\zeta) : \mathbb{Q}] = \varphi(8) = 4$ by

the theory of cyclotomic extensions. (In fact one may easily calculate that $\text{minpoly}_{\mathbb{Q}}(\zeta) = \Phi_8(x) = x^4 + 1$). This shows that $\mathbb{Q}(\zeta) = \mathbb{Q}(\sqrt{2}, i)$.

We do know that $x^8 - 2$ is irreducible over $\mathbb{Q}$ by the Eisenstein criterion, and so $x^8 - 2 = \text{minpoly}_{\mathbb{Q}}(\alpha)$ and hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 8$. Now

$$K = \mathbb{Q}(\alpha, \zeta) = \mathbb{Q}(\sqrt[8]{2}, i, \sqrt{2}) = \mathbb{Q}(\sqrt[8]{2}, i)$$

from which it is straightforward to see that $[K : \mathbb{Q}] = 16$, since $i \notin \mathbb{Q}(\sqrt[8]{2})$. This implies that $[K : F] = 4$ and so $\text{Gal}(K/F) \cong \mathbb{Z}_4$.

In retrospect, it is easy to see explicitly that $f = x^8 - 2$ is not irreducible over $F = \mathbb{Q}(\zeta)$. Since $\sqrt{2} \in F$, f factors as $(x^4 - \sqrt{2})(x^4 + \sqrt{2})$ in $F[x]$.

The perhaps more surprising fact is that when a Galois extension has a cyclic Galois group, and the base field has enough roots of 1 already, then the extension must be given by adjoining an nth root. The standard proof relies on a result known as "linear independence of group characters", a result with other applications in field theory which we would certainly present as well if we had more time. Here, for simplicity we give a proof which relies on some of the techniques of modules over PIDs we already have on hand.

Proposition 18.36. *Let $F \subseteq K$ be a an extension with $[K : F] < \infty$. Assume that $x^n - 1$ splits in F with distinct roots. Suppose that K/F is Galois with $\text{Gal}(K/F)$ cyclic of order dividing n. Then $K = F(\alpha)$ for some $\alpha \in K$ such that $\alpha^n \in F$.*

Proof. By assumption $G = \text{Gal}(K/F)$ is cyclic of order d where d divides n. Since K/F is Galois, $d = [K : F]$ also. Let σ be a generator of G, so σ has order d. Note that σ preserves addition and if $\lambda \in F, a \in K$, then $\sigma(\lambda a) = \sigma(\lambda)\sigma(a) = \lambda\sigma(a)$, since σ fixes F. It follows that σ is an F-linear transformation of the n-dimensional vector space K over F. As such, there is an associated $F[x]$-module structure on K, where x acts as σ, and we now investigate this module using our results on modules over a PID.

Since $\sigma^d = 1$, σ satisfies the polynomial $x^d - 1$. Thus the minimal polynomial of σ divides $x^d - 1$. Since d divides n and F has n distinct nth roots of 1, F has d distinct dth roots of 1 already as well. Thus $x^d - 1$ factors as $(x - 1)(x - \rho) \dots (x - \rho^{d-1})$ for some primitive dth root ρ of 1 in F, where $1, \rho, \dots \rho^{d-1} \in F$ are all distinct.

Let $a_1, \dots, a_m \in F[x]$ be the invariant factors of K as an $F[x]$-module, so that

$$K \cong F[x]/(a_1) \oplus \cdots \oplus F[x]/(a_m)$$

as $F[x]$-modules, where $a_i | a_{i+1}$ for all i. We have seen that the largest invariant factor a_m is the minimal polynomial of σ. Now a_m divides $x^d - 1$ and thus a_m factors in $F[x]$ as a product of distinct linear factors. Then the same is true of all a_i. Now the elementary divisors of the module are found by factoring each invariant factor as a product of powers of distinct irreducibles. So in this case we see that all elementary divisors have degree 1. It follows that σ has a Jordan form over F, and moreover this Jordan form is diagonal, with diagonal entries which are dth roots of 1. Since σ has order d and not smaller order, one of the diagonal entries has to be a primitive dth root of 1. Without loss of generality we can assume ρ is one of the entries. This tells us that σ has an eigenvector in K, say $\alpha \in K$, with eigenvalue ρ. Thus $\sigma(\alpha) = \rho\alpha$.

Up until now we have not used that σ is an automorphism of K, i.e. that σ preserves multiplication. Now we note that $\sigma(\alpha^i) = (\sigma(\alpha))^i = \rho^i \alpha^i$. Thus $\alpha^i \in K$ is an eigenvector of σ with eigenvalue ρ^i, and we conclude that all powers of ρ, and hence all dth roots of 1, are eigenvalues of σ. Also, since $1, \alpha, \alpha^2, \ldots \alpha^{d-1}$ are eigenvectors with distinct eigenvalues, they are linearly independent over F. Since $[K : F] = d$, these powers of α form a basis of K over F. With respect to this basis σ has diagonal entries $1, \rho, \ldots, \rho^{d-1}$ and so the minimal polynomial of σ is in fact $x^d - 1$. (Note that the minimal polynomial of a vector space map is not necessarily irreducible, unlike the minimal polynomial of an algebraic element in a field extension.)

Since the powers of α give a basis of K over F, certainly $F(\alpha) = K$. And since $\sigma(\alpha^d) = \rho^d \alpha^d = \alpha^d$, we have $\alpha^d \in \mathrm{Fix}\langle\sigma\rangle = \mathrm{Fix}\, G = F$ since the extension is Galois. Certainly then $\alpha^n \in F$ as well. □

Putting together the previous two results we get the following very nice theorem.

Theorem 18.37. *Let $F \subseteq K$ be a field extension and suppose that $x^n - 1$ has n distinct roots in F. Then the following are equivalent:*

(1) *K/F is Galois with $\mathrm{Gal}(K/F)$ cyclic of order dividing n.*
(2) *$K = F(\alpha)$ for some $\alpha \in K$ with $\alpha^n \in F$.*

We now single out those extensions that can be formed by iterating the procedure of adjoining a root.

Definition 18.38. A field extension $F \subseteq K$ is called a *root extension* if there is a chain of subfields

$$F = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K$$

such that for all $i \geq 0$ there is $\alpha_i \in K_{i+1}$ and $n_i \geq 1$ such that $K_{i+1} = K_i(\alpha_i)$ and $\alpha_i^{n_i} \in K_i$. A polynomial $f \in F[x]$ is *solvable by radicals* if there exists a root extension $F \subseteq K$ such that f splits in $K[x]$.

In other words, at each step in a root extension, we get the next field by adjoining some root of an element we already have. Speaking loosely, the elements in a root extension K of F are those that can be expressed using only elements in F, field operations, and nested root symbols. Thus a polynomial is solvable by radicals if all of its roots can be expressed in such a way. For example, $\sqrt[3]{(1/10) + \sqrt{2}}/\sqrt[5]{3}$ lies in a root extension of $\mathbb{Q}$, namely

$$\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2}) \subseteq \mathbb{Q}(\sqrt{2})(\sqrt[3]{(1/10) + \sqrt{2}}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt[3]{(1/10) + \sqrt{2}})(\sqrt[5]{3}).$$

A major preoccupation of mathematicians in the Renaissance period in Europe was to find solutions for a general polynomial equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

Really, what was desired was a formula, or method, that would give the solutions to any equation in terms of manipulations of its coefficients, including algebraic operations and taking roots. Solutions of this type for quadratic equations had been long known (although not exactly in the form of the quadratic formula as we know it today). Formulas for solving polynomial equations of degree 3 and 4 were eventually successfully developed. These were first published by Cardano in 1545, but the solutions are now attributed to Tartaglia and del Ferro (for the cubic) and Ferrari (for the quartic). Interestingly, complex numbers arise naturally in these formulas even when one is only seeking real roots, and this probably helped spur the eventual acceptance of complex numbers in mathematics. No solution to the general degree 5 (quintic) equation could be found in teh following years, and eventually, hundreds of years later in 1824, Abel proved that no solution of this kind could exist (building on earlier work of Ruffini). Galois's theory, which came just a bit later in 1830, then put this theorem into a more general context.

Technically, what Abel and Galois proved is that there can exist no formula for the solution of a quintic which depends only on field operations and nested root signs. In our terminology, we can state the key theorem as follows:

Theorem 18.39. *There exists a polynomial $f \in \mathbb{Q}[x]$ of degree 5 such that the splitting field K of f in $\mathbb{C}$ is not a root extension of $\mathbb{Q}$.*

Thus the roots of f in $\mathbb{C}$ are "not expressible in terms of radicals". In fact Galois's work showed a much stronger result, which today we state in terms of solvable groups although that concept wasn't formalized at the time.

Theorem 18.40. *Let F be a field of characteristic 0. Then $f \in F[x]$ is solvable by radicals if and only if $\mathrm{Gal}(K/F)$ is a solvable group, where K is the splitting field of f over F.*

Then to give the required example in Theorem 18.39, it just suffices to find a particular polynomial whose splitting field has a non-solvable Galois group. Since the splitting field of a polynomial of degree d has degree at most $d!$ over $\mathbb{Q}$, this explains why no such example could exist when $d \leq 4$, since all groups of order at most 24 are solvable, and thus all polynomials of degree at most 4 in $\mathbb{Q}$ are solvable by radicals. On the other hand, we will see that there does exist a polynomial of degree 5 such that the Galois group of its splitting field is the non-solvable group S_5.

We have decided to omit the complete proof of Theorem 18.40, which is a bit technical. The result itself does not have the importance it once did, as solving polynomial equations explicitly is no longer a central topic in algebra. But from the results we have already presented the main idea of Theorem 18.40 is easy to grasp, and so we briefly discuss this. First, recall that in our study of groups we defined a group G to be solvable if it has a series of subgroups

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq H_2 \trianglelefteq \cdots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$$

such that H_{i+1}/H_i is abelian for all i. However, if G is finite and solvable, then in fact it has such a series where each H_{i+1}/H_i is cyclic. This follows easily from the fact that a finite abelian group is a direct product of cyclic groups, which allows one to add additional terms to any series as above in order to make the factor groups cyclic. Thus finite solvable groups are exactly those groups which are built out of cyclic groups in this sense. By definition, root extensions are field extensions which are built out of extensions where one adjoins a single root. Finally, Theorem 18.37 showed that an extension where one adds a root has Galois group which is cyclic (under certain hypotheses). These facts, together with the fundamental theorem of Galois theory, already suggest that root extensions should correspond roughly to solvable groups under the Galois correspondence.

In order to write down a precise proof of Theorem 18.40, one has to deal with some additional technical details. First, in Theorem 18.40 there is no assumption that F contains any roots of 1, so to successfully apply Theorem 18.37 one first has to adjoin a bunch of roots of 1 to F, and show this doesn't change the property of being a root extension. Second, a root extension is not necessarily Galois, since at each stage we just add one root of a polynomial, so one may need to pass to a larger root extension which is Galois in order to apply the fundamental theorem.

From now on we simply assume Theorem 18.40. We will, however, show how to use it to prove Theorem 18.39.

Proof of Theorem 18.39. Let $f = 2x^5 - 10x + 5 \in \mathbb{Q}[x]$. We claim that f is not solvable by radicals over $\mathbb{Q}$.

Let K be the splitting field of f over $\mathbb{Q}$. We claim that $\mathrm{Gal}(K/\mathbb{Q}) \cong S_5$. Once this proved, then since S_5 is not solvable, we will know that f is not solvable by radicals by Theorem 18.40.

The polynomial f is irreducible over $\mathbb{Q}$ by an application of the Eisenstein criterion with prime 5. Thus if $\alpha \in K$ is a root of f, we will have $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 5$. In particular, if $G = \mathrm{Gal}(K/\mathbb{Q})$ then $|G| = |K : \mathbb{Q}|$ is divisible by 5. Thus G has an element of order 5 by Cauchy's theorem.

Next, if $\alpha_1, \ldots, \alpha_5$ are the roots of f in K (they are distinct since we are in characteristic 0), then every $\sigma \in G$ sends each root of f to a root of f, and so gives a permutation of these roots. Since these roots also generate K over $\mathbb{Q}$, σ is determined by how it permutes the roots of f. In this way we get an injective homomorphism from G to S_5, which maps $\sigma \in G$ to the corresponding permutation of the roots of f. Now think of G as a subgroup of S_5. The only elements of order 5 in S_5 are 5-cycles, so G contains a 5-cycle.

Now f was chosen to have exactly 3 real roots. This fact can be easily verified using calculus. Namely, we have $f' = 10x^4 - 10$ which has real zeroes at the values ± 1. By the mean value theorem, between any two real roots $r_1 < r_2$ of f there must be $r_1 < s < r_2$ such that $f'(s) = 0$. Since f' has only two real roots we conclude that f has at most 3 real roots. On the other hand, $f(-2) < 0$, $f(-1) > 0$, $f(1) < 0$, and $f(2) > 0$, so an application of the intermediate value theorem shows that f has at least 3 real roots. So f has exactly 3 real roots. Now we have, say, that $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$, while $\alpha_4, \alpha_5 \notin \mathbb{R}$. Since the non-real roots of a polynomial with real coefficients come in conjugate pairs, we must have $\alpha_5 = \overline{\alpha_4}$.

Now the complex conjugation map $z \mapsto \overline{z}$ is an automorphism of $\mathbb{C}$. It permutes the roots α_i of f and thus it restricts to an automorphism of K, that is an element in $G = \mathrm{Gal}(K/\mathbb{Q})$. This automorphism obviously acts on the roots as the 2-cycle (45).

To complete the proof, one checks that any subgroup of S_5 which contains a 5-cycle and a 2-cycle is equal to all of S_5. This is an exercise in group theory which we leave to the reader. This proves the claim that $G \cong S_5$ and finishes the proof. □

18.7. **Algebraically closed fields and algebraic closures.** We have used throughout our study of field theory that the field $\mathbb{C}$ of complex numbers is algebraically closed. In this section we will prove that result, as well as studying algebraically closed fields more generally.

Recall the following definition:

Definition 18.41. A field K is algebraically closed if for all $f \in K[x]$, f splits in $K[x]$, that is $f = c(x - \alpha_1) \dots (x - \alpha_n)$ for some $\alpha_1, \dots, \alpha_n \in K$.

This definition can formulated in a number of slightly different ways.

Lemma 18.42. *Let K be a field. The following are equivalent:*

(1) *K is algebraically closed, that is, every polynomial in $K[x]$ splits in $K[x]$.*

(2) *Every nonconstant polynomial $f \in K[x]$ has a root in K.*

(3) *If $K \subseteq L$ is an algebraic extension, then $L = K$.*

(4) *If $K \subseteq L$ is an algebraic extension with $[L : K] < \infty$, then $L = K$.*

Proof. (1) $\implies$ (2) is obvious.

For (2) $\implies$ (3), supppose that L/K is algebraic. Given $\alpha \in L$, let $f = \text{minpoly}_K(\alpha) \in K[x]$. Then f has a root in K, but f is also irreducible over K. This forces $\deg f = 1$ and hence $\alpha \in K$. Thus $K = L$.

(3) $\implies$ (4) is also immediate.

(4) $\implies$ (1): Let $f \in K[x]$. Let L be a splitting field for f over K, so $[L : K] < \infty$ by definition. Then we have $L = K$, so f splits in $K[x]$ already. □

Definition 18.43. If $F \subseteq K$ is a field extensions, K is called an *algebraic closure* of F if K/F is algebraic and K is algebraically closed.

We are going to show shortly that every field has an algebraic closure, and the algebraic closure is unique up to isomorphism.

Here is an alternative way of thinking about the algebraic closure.

Lemma 18.44. *Let $F \subseteq K$ be an algebraic extension. Then K is an algebraic closure of F if and only if every $f \in F[x]$ splits in $K[x]$.*

Proof. One direction is easy: if K is an algebraic closure, then K is algebraically closed. Since every $f \in F[x]$ is also in $K[x]$, of course f splits in $K[x]$ by definition.

Conversely, suppose that every $f \in F[x]$ splits in $K[x]$. Suppose that $K \subseteq L$ is an algebraic extension. Since $F \subseteq K$ is algebraic, we have that $F \subseteq L$ is algebraic, by Theorem 17.26. For $\alpha \in L$, consider $f = \text{minpoly}_F(\alpha)$. Then f splits in $K[x]$. Since α is a root of f, $\alpha \in K$. Thus $K = L$. Now K is algebraically closed by Lemma 18.42. □

While we have only defined the splitting field of a single polynomial, given a set S of polynomials in $F[x]$, one could define an algebraic extension $F \subseteq K$ to be a *splitting field of the set* S if every polynomial in S splits in K and K is generated over F by the roots of all polynomials in S. In this point of view, Lemma 18.44 is telling us that an algebraic closure of F is essentially a splitting field for the set of *all* polynomials in $F[x]$.

Let us note that to find an algebraic closure, it suffices to find some algebraically closed field containing a given one.

Lemma 18.45. *Let $F \subseteq L$ be a field extension and assume that L is algebraically closed. Let $K = \{\alpha \in L | \alpha$ is algebraic over $F\}$. Then K is an algebraic closure of F.*

Proof. We have seen previously in Corollary 17.20 that K is a subfield of L such that K/F is algebraic. If $f \in F[x]$, then f splits in $L[x]$ since L is algebraically closed. But the roots of f in L are algebraic over F and so they belong to K. Thus f already splits over K. Now K is an algebraic closure of F by Lemma 18.44. □

We previously defined the field of algebraic numbers $\overline{\mathbb{Q}}$ as the set of elements in $\mathbb{C}$ that are algebraic over $\mathbb{Q}$. We see by Lemma 18.45 that $\overline{\mathbb{Q}}$ is an algebraic closure of $\mathbb{Q}$. It is much smaller than $\mathbb{C}$, for it is not hard to prove that an algebraic closure of a countable field F is again countable, for there are countably many polynomials in $F[x]$, and each has finitely many roots.

We are now ready to prove the first main result of this section.

Theorem 18.46. *Any field F has an algebraic closure.*

Proof. By Lemma 18.45, it suffices to find any algebraically closed field L containing F. Then the set of elements in L that are algebraic over F will be the desired algebraic closure.

All proofs of this result depend on some version of the axiom of choice. The following elegant proof, due to Emil Artin, just uses the fact that any commutative ring has a maximal ideal (which we proved as a consequence of Zorn's lemma). Recall the idea of Lemma 17.28: Given a irreducible polynomial $f \in F[x]$, we can create a larger field containing F in which f has a root by taking $F[x]/(f)$. The idea of this proof is to do this for all polynomials in $F[x]$ at once, adding one new variable for each one. The resulting ring is not a field, but we can just pass to some factor ring that is a field.

For each nonconstant polynomial $f \in F[x]$ we define an intedeterminate x_f. Now we define $R = F[x_f | f \in F[x], \deg(f) > 0]$ to be a polynomial ring generated over F by these (infinitely many)

variables. Let I be the ideal of R generated by the set of polynomials $\{f(x_f)|f \in F[x], \deg(f) > 0\}$. Each $f(x_f)$ is a polynomial involving only one of the variables.

We claim that $I \neq R$. Suppose instead that $I = R$ is the unit ideal. Then $1 \in I$, so $1 = \sum_{i=1}^{n} g_i f_i(x_{f_i})$ for some distinct polynomials $f_1, f_2, \ldots, f_n \in R$ with $\deg f_i > 0$ and some $g_i \in R$. Let K be a splitting field over F of $f_1 f_2 \ldots f_n$. Thus each f_i has a root $\alpha_i \in K$. Now we define a homomorphism $\phi : R \to K$ as follows: We let $\phi(a) = a$ for $a \in F$; $\phi(x_{f_i}) = \alpha_i$ for $1 \leq i \leq n$; and $\phi(x_g) = 0$ for all g's not equal to any f_i. Note that by the universal property of a polynomial ring, given a homomorphism $F \to K$, we can specify a unique homomorphism from $R \to K$ by sending each variable x_f to any element of K we please, so there does exist such a homomorphism ϕ.

Now we have $1 = \phi(1) = \sum_{i=1}^{n} \phi(g_i) f_i(\alpha_i) = \sum_{i=1}^{n} \phi(g_i) 0 = 0$, because α_i is a root of f_i by definition. This is a contradiction. Thus $I \neq R$ as claimed.

Since I is a proper ideal, we can choose a maximal ideal M of R with $I \subseteq M \subseteq R$ (Proposition 9.6). Then $L_1 = R/M$ is a field and we have a homomorphism $\psi : F \to L_1$ which is the composition of the inclusion of F in R followed by the natural homomorphism $R \to R/M$. Since F is a field, ψ is injective and so we can think of F as a subfield of L_1. As such, if $f \in F[x]$ is irreducible, then f has a root $x_f + M \in L_1$, because $f(x_f + M) = f(x_f) + M = 0 + M$, as $f(x_f) \in I \subseteq M$.

To summarize, we have shown that there exists a field extension $F \subseteq L_1$ such that every nonconstant $f \in F[x]$ has a root in L_1. Note that this does not prove that L_1 is an algebraic closure of F, because we do not know that f splits over L_1, only that it has at least one root. However, we can now proceed inductively as follows. By the same argument, there is a field extension $L_1 \subseteq L_2$ such that every nonconstant polynomial in $L_1[x]$ has a root in L_2. Continue in this way, defining a chain of fields $F \subseteq L_1 \subseteq L_2 \subseteq \ldots$. Now let $L = \bigcup_{n \geq 0} L_n$, which as a union of fields is easily seen to be a field. If $g \in L[x]$ is nonconstant, then each coefficient of g lies in some L_i, and so $g \in L_n[x]$ for some n. By construction, g has a root in $L_{n+1} \subseteq L$. Since every nonconstant polynomial in $L[x]$ has a root in L, the field L is algebraically closed by Lemma 18.42. □

Let us also prove now that algebraic closures are unique up to isomorphism. This is similar to our results on uniqueness of splitting fields (as we have already remarked, an algebraic closure is like a splitting field for the set of all polynomials). As in those results, is convenient to work over an isomorphism of base fields rather than a single base field.

Theorem 18.47. *Let $\theta : F \to F'$ be an isomorphism of fields. Let $F \subseteq K$ and $F' \subseteq K'$ be algebraic closures. Then there is an isomorphism $\rho : K \to K'$ such that $\rho|_F = \theta$.*

Proof. Consider the set consisting of triples (E, E', ψ) where E, E' are subfields with $F \subseteq E \subseteq K$, $F' \subseteq E' \subseteq K'$, and $\psi : E \to E'$ is an isomorphism such that $\psi|_F = \theta$. Put a partial order on this set where $(E, E'\psi) \leq (L, L', \rho)$ if $E \subseteq L$, $E' \subseteq L'$, and $\rho|_E = \psi$. In other words, elements of the set are isomorphisms matching up subfields of K and K', and a larger element represents an extension of that isomorphism to one defined on larger subfields. The hypotheses of Zorn's Lemma hold for this set, because given any chain $\{(E_i, E_i', \psi_i) | i \in I\}$ we can can extend the isomorphisms ψ_i to the unions to get an upper bound of the form $(\bigcup_i E_i, \bigcup_i E_i', \psi)$.

By Zorn's Lemma, there is a maximal element (L, L', ρ) in the set. Suppose that $L \neq K$. Then if $\alpha \in K \setminus L$, let $f = \text{minpoly}_L(\alpha)$ and choose any root $\alpha' \in K'$ of $f' = \rho(f)$; such a root exists because K' is algebraically closed. By Lemma 17.33, we can extend the isomorphism ρ to an isomorphism $\delta : L(\alpha) \to L'(\alpha')$, where $\delta|_L = \rho$. But this implies that $(L(\alpha), L'(\alpha'), \delta)$ is a strictly larger element of our set of partial isomorphisms, contradicting that (L, L', ρ) is maximal. We conclude that $L = K$.

Now $\rho : K \to L'$ is an isomorphism. The property of being algebraically closed is preserved by automorphisms, so L' is algebraically closed as well. However, since K'/F' is algebraic, we see that K'/L' is algebraic. Because algebraically closed fields have no proper algebraic extensions by Lemma 18.42, $K' = L'$ and ρ is an isomorphism $K \to K'$ such that $\rho|_F = \theta$, as we wished. □

Of course, taking $\theta = 1_F$ in the result above we get that any two algebraic closures K, K' of F are isomorphic as fields. Given that the algebraic closure is essentially unique by this result, sometimes the algebraic closure of a field F is simply notated $\overline{F}$. The notation $F = \overline{F}$ is used as shorthand to indicate that a field F is itself algebraically closed.

Here is another interesting consequence.

Corollary 18.48. *Let $F \subseteq E$ be an algebraic extension. If $F \subseteq K$ is an algebraic closure, there is an isomorphism $\phi : E \to L$ for some subfield L with $F \subseteq L \subseteq K$, where $\phi|_F = 1_F$.*

In other words, given any algebraic extension, an "isomorphic copy" of it can be found inside any fixed algebraic closure of the base field. Thus when we are studying algebraic extensions of F, we can always fix an algebraic closure of F and make all constructions inside there if we wish.

Proof. Let $E \subseteq K'$ be an algebraic closure of E. Since E/F and K'/F are algebraic, K'/F is also algebraic. Because K' is algebraically closed, we conclude that K' is also an algebraic closure of F. Now choose by Theorem 18.47 an isomorphism $\rho : K' \to K$ such that $\rho|_F = 1_F$. If $L = \rho(E)$, then $\phi = \rho|_E$ is an isomorphism $\phi : E \to L$ with $\phi|_F = 1_F$. □

Example 18.49. Let $\mathbb{F}_p$ be the field with p elements for some prime p. Fix an algebraic closure $\overline{\mathbb{F}_p}$. For each $n \geq 1$ we have a field $\mathbb{F}_{p^n}$ with p^n elements. By the corollary we can find a copy of this field inside the algebraic closure, so we consider $\mathbb{F}_p \subseteq \mathbb{F}_{p^n} \subseteq \overline{\mathbb{F}_p}$.

Now we claim that $\overline{\mathbb{F}_p} = \bigcup_n \mathbb{F}_{p^n}$. To see this, note that if $\alpha \in \overline{\mathbb{F}_p}$, then α is algebraic over $\mathbb{F}_p$, so we can consider $f = \text{minpoly}_{\mathbb{F}_p}(\alpha)$. If f has degree n, then f divides $x^{p^n} - x$ in $\mathbb{F}_p[x]$, so α is a root of $x^{p^n} - x$. The subfield $\mathbb{F}_{p^n}$ must be equal to the set of roots of $x^{p^n} - x$ in $\overline{\mathbb{F}_p}$, so we see that $\alpha \in \mathbb{F}_{p^n}$ for the fixed copy of $\mathbb{F}_{p^n}$. This proves the claim.

This gives a quite explicit picture of $\overline{\mathbb{F}_p}$ as the union of all finite fields of characteristic p. Note, however, that these fields do not form a single chain, as $\mathbb{F}_{p^d}$ is a subset of $\mathbb{F}_{p^n}$ if and only if $d|n$. Rather, we are taking the union of a more complicated partially ordered set.

For our last main result, we will finally prove that the field $\mathbb{C}$ of complex numbers is algebraically closed. Many quite different proofs of this result are known. All use some amount of analysis, which is unavoidable because the real numbers are an analytic object, defined by a limiting process. There is a well-known proof that relies on results in complex analysis, for example.

The proof we give uses Galois theory and reduces the amount of analysis needed to a few elementary facts about the real numbers.

Lemma 18.50. (1) *If $f \in \mathbb{R}[x]$ has odd degree, then f has a root in $\mathbb{R}$.*

(2) *if $g \in \mathbb{C}[x]$ has degree 2, then g splits over $\mathbb{C}$.*

Proof. (1) Let $f = a_n x^n + \cdots + a_0 \in \mathbb{R}[x]$ where $\deg f = n$ is odd. Since we are just trying to show that f has a root in $\mathbb{R}$, without loss of generality we can replace f with $-f$ if necessary and thus assume that $a_n > 0$. Now it is standard that $\lim_{n\to\infty} f(x) = \infty$ and $\lim_{n\to-\infty} f(x) = -\infty$. By the intermediate value theorem, since f is a continuous function f must have a root in $\mathbb{R}$.

(2). If $g = ax^2 + bx + c$ then the quadratic formula tells us that $(-b + \sqrt{b^2 - 4ac})/2a$ is a root of g in $\mathbb{C}$, for any square root $\sqrt{b^2 - 4ac}$ in $\mathbb{C}$. Once g has a root α in $\mathbb{C}$, then $g = (x - \alpha)h$ by the factor theorem, but then h already has degree 1 and so g splits. □

Note that for any complex number $z = re^{i\theta}$ in polar form, with $r \geq 0$, then $\sqrt{r}e^{i\theta/2}$ is a square root of z, where $\sqrt{r}$ is the nonnegative real square root of r. Thus ultimately the existence of square roots in $\mathbb{C}$ is a consequence of the fact that nonegative real numbers have a unique nonnegative square root. This follows easily from the least upper bound property.

The analysis above is all we need to prove our main result.

Theorem 18.51. $\mathbb{C}$ *is algebraically closed.*

Proof. We will show that if $\mathbb{C} \subseteq L$ is a finite degree extension, then $L = \mathbb{C}$. This implies that $\mathbb{C}$ is algebraically closed by Lemma 18.42.

Since $[\mathbb{C} : \mathbb{R}] = 2$, we also have $[L : \mathbb{R}] < \infty$, and of course $L/\mathbb{R}$ is separable since we are in characteristic 0. Thus we can take a Galois closure M of L so that $L \subseteq M$ and $M/\mathbb{R}$ is Galois, by Proposition 18.13.

Now let $G = \mathrm{Gal}(M/\mathbb{R})$. Let P be a Sylow 2-subgroup of G. Let $K = \mathrm{Fix}(P)$. Then $[K : \mathbb{R}] = [G : P]$ is odd. If $\alpha \in K$, then $[\mathbb{R}(\alpha) : \mathbb{R}]$ divides $[K : \mathbb{R}]$ so it is also odd. Then $f = \mathrm{minpoly}_{\mathbb{R}}(\alpha)$ is irreducible and of odd degree. But by Lemma 18.50(1), f has a root in $\mathbb{R}$, and so cannot be irreducible unless it has degree 1, in which case $\alpha \in \mathbb{R}$. This shows that $K = \mathbb{R}$. This implies $P = G$ and so G is a finite 2-group. Also, $[M : \mathbb{R}]$ is a power of 2.

Since $[\mathbb{C} : \mathbb{R}] = 2$ we have $[M : \mathbb{C}]$ is a power of 2 as well. Also, because $M/\mathbb{R}$ is Galois, so is $M/\mathbb{C}$. Suppose that $M \neq \mathbb{C}$. Now $\mathrm{Gal}(M/\mathbb{C})$ is a nontrivial 2-group, so from our earlier results on p-groups we know that it must have a subgroup H of index 2. If $E = \mathrm{Fix}(H)$ then $\mathbb{C} \subsetneq E \subseteq M$ with $[E : \mathbb{C}] = 2$. If $\alpha \in E \setminus \mathbb{C}$, then $\mathrm{minpoly}_{\mathbb{C}}(\alpha)$ has degree 2, but by Lemma 18.50, any degree 2 polynomial in $\mathbb{C}[x]$ splits over $\mathbb{C}$ and cannot be irreducible, which is a contradiction. We conclude that in fact $M = \mathbb{C}$. Then $L = \mathbb{C}$, completing the proof. □

We close with a curious result about the automorphism group of $\mathbb{C}$. In a homework problem, you were asked to show that $\mathrm{Aut}(\mathbb{R}) = 1$, by checking that every automorphism fixes $\mathbb{Q}$ and is continuous. On the other hand, although $\mathbb{C}$ is just a degree 2 extension of $\mathbb{R}$, its automorphism group $\mathrm{Aut}(\mathbb{C})$ is actually very large. This can be seen using the idea of a transcendence basis.

Given any field extension $F \subseteq K$, it is possible to choose a set of elements $\{y_\alpha \in K | \alpha \in I\}$ such that (i) the $\{y_\alpha\}$ are *algebraically independent* in the sense that $T = F(y_\alpha | \alpha \in I)$ is isomorphic to the field of fractions of a polynomial ring $F[y_\alpha | \alpha \in I]$; and (ii) $T \subseteq K$ is algebraic. The set $\{y_\alpha | \alpha \in I\}$ is called a *transcendence basis* for the extension. It can be proved to exist by an application of Zorn's Lemma. The cardinality of a transcendence basis is uniquely determined (though the subfield T it generates isn't). This cardinality is called the *transcendence degree* of the extension.

Consider the particular extension $\mathbb{Q} \subseteq \mathbb{C}$. In this case one can show that the transcendence degree is uncountable. (It this were not the case, but rather the transcendence basis $\{y_\alpha\}$ were countable, then the field of rational functions $\mathbb{Q}(y_\alpha | \alpha \in I)$ would be countable, and then since an algebraic extension of a countable field is countable, $\mathbb{C}$ would be countable, a contradiction.)

Now fix a transcendence basis $\{y_\alpha | \alpha \in I\}$ for $\mathbb{C}$ over $\mathbb{Q}$. Any permutation of the set I gives an automorphism of the polynomial ring $\mathbb{Q}[y_\alpha | \alpha \in I]$ where the variables are permuted in the

same way. This then extends to an automorphism of the field $T = \mathbb{Q}(y_\alpha | \alpha \in I)$ with the same permutation of the variables. Finally, since $\mathbb{C}$ is algebraically closed and $\mathbb{C}/T$ is algebraic, $\mathbb{C}$ is an algebraic closure of T. Thus any automorphism of T extends to an automorphism of $\mathbb{C}$, by Theorem 18.47.

In this way we can see that $\mathbb{C}$ has at least as many automorphisms as the number of elements in the permutation group $\mathrm{Sym}(I)$, where I is the index set of a transcendence basis. Since I is uncountable, the set of permutations of I actually has cardinality even bigger than the cardinality of I (as can be seen by a version of Cantor's diagonal argument). Thus $\mathrm{Aut}(\mathbb{C})$ is huge.

On the other hand, while a transcendence basis for $\mathbb{C}$ over $\mathbb{Q}$ exists, it is impossible to write one down explicitly, and so the automorphisms of $\mathbb{C}$ one gets in this way also do not have any kind of explicit description. And in fact they tend to have bizarre properties. It is possible to show that except for the identity map and complex conjugation, any automorphism of $\mathbb{C}$ is discontinuous and maps $\mathbb{R}$ onto a dense subset of $\mathbb{C}$. So these really are hard to picture.

The concept of a transcendence basis is generally useful in commutative ring theory (not just for creating strange automorphisms). We would cover it in more detail if we had more time. You can find a treatment of it in Chapter 24 of Isaacs' book.