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. $\square$

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 $|xH| = |H|$. $\square$

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. $\square$

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 $a'b' = ah_1bh_2 = abh_3h_2 \in abH$. 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 $G = D_{2n} = \{1, a, a^2, \dots, a^{n-1}, b, ab, \dots, a^{n-1}b\}$, 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 $H = \{1, a, a^2, \dots, a^{n-1}\}$ 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 $bH = \{b, ab, \dots, a^{n-1}b\}$ 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

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

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 $K \subseteq N_G(H)$. $\square$

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 $H \le G$ and $K \le G$.*

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

(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 $HK \le G$ 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 $\lambda, 1, 1, \dots, 1$ 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 $\lambda, 1, 1 \dots, 1$, 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. $\square$

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 $\operatorname{Inn}(G) \leq \operatorname{Aut}(G)$, so we just need to prove normality. Let $\theta_g \in \operatorname{Inn}(G)$ and let $\rho \in \operatorname{Aut}(G)$. 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 $\rho\theta_g\rho^{-1} = \theta_{\rho(g)} \in \operatorname{Inn}(G)$ and so $\operatorname{Inn}(G)$ is normal in $\operatorname{Aut}(G)$. $\square$

Because of the lemma, it makes sense to define the factor group $\operatorname{Out}(G) = \operatorname{Aut}(G)/\operatorname{Inn}(G)$, 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 $G = \{a^i b^j | 0 \leq i \leq 3, 0 \leq j \leq 1\}$, with $a^4 = 1, b^2 = 1$, and $ba = a^{-1}b$. Then $H = \{1, a^2, b, a^2 b\}$ 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 $aba^{-1} = a^2 b \notin K$.

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 $\sigma \in \operatorname{Aut}(G)$, $\sigma(H) = H$. We write $H \operatorname{char} G$ in this case.

Clearly from the remarks above, characteristic subgroups are normal.

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

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

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 $K \operatorname{char} H$, 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. $\square$

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 $(h_1, k_1)(h_2, k_2) = (h_1h_2, k_1k_2)$ 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 $G_1, G_2, \dots, G_k$ in an analogous way, as the set of all k-tuples $(g_1, g_2, \dots, g_k)$ 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 $\{a^ib^j | 0 \leq i \leq n-1, 0 \leq j \leq 1\}$ 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:

$$\begin{tikzcd} G \arrow[r, dashed, "\exists! \widehat{f}"] & H \\ X \arrow[u, "i"] \arrow[ur, "f"'] & \end{tikzcd}$$

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 $\mathrm{red}(w)$ (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 $v * w = \mathrm{red}(vw)$, 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 $v * w = \mathrm{red}(vw) = v'w'$. 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. $\square$

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 $\mathrm{Sym}(G) \cong S_n$. $\square$

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 $\{1, 2, \dots, n\}$.

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. $\square$

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 Lagranges theorem, and $|G : G_x| = |G|/|G_x|$ is also a divisor of $|G|$. $\square$

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 $|HK| = |H||K|/|H \cap K|$. $\square$

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 $\mathrm{Cl}(a^i b) = \{b, ab, \dots, a^{n-1} b\}$ is the set of all reflections. If n is even, on the other hand, then the reflections break up into two conjugacy classes $\{b, a^2 b, \dots, a^{n-2} b\}$ and $\{ab, a^3 b, \dots, a^{n-1} b\}$, 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 $C_G(a^i) = \{1, a, \dots, a^{n-1}\}$ 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, that 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

$(yH)^n = y^n H = 1H = H$. 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 $|y^\ell| = p$. $\square$

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. $\square$

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. $\square$

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 $n_p = |G : N_G(P)| = |G : G| = 1$. $\square$

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 $\{1\}$ 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 $X = \{1, 2, \dots, n\}$. 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 $a_1, a_2, \dots, a_k$ 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 $(a_1 a_2 \dots a_k)$ 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 $(a_1 a_2 \dots a_k)$ and $(b_1 b_2 \dots b_l)$ 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 since it can be found in any beginning abstract algebra book.

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 $\{1, g, g^2, \dots, g^{d-1}\}$ are d distinct elements of g, so that for any $x \in G$, the elements $\{x, gx, g^2x, \dots, g^{d-1}x\}$ are also distinct. Moreover, since the action of g on the left satisfies $g \cdot g^i x = g^{i+1}x$ for $0 \le i \le d-2$ 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 $H = \phi^{-1}(A_n) \le G$. 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 $\rho \in \operatorname{Aut}(G)$, 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 $\operatorname{Cl}_{S_n}(\sigma)$ 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 $|\mathcal{O}_\sigma| = |\operatorname{Cl}_{S_n}(\sigma)|/2$ by the calculations above. If this happens, because $\operatorname{Cl}_{S_n}(\sigma)$ is a union of A_n-orbits, the only possibility is that $\operatorname{Cl}_{S_n}(\sigma)$ 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 $|\mathcal{O}_\sigma| = |\operatorname{Cl}_{S_n}(\sigma)|$, so that $\mathcal{O}_\sigma = \operatorname{Cl}_{S_n}(\sigma)$.

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 $\operatorname{Cl}_{A_n}(\sigma)$. 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 $\mathcal{K} = \operatorname{Cl}_{S_n}(\sigma)$, 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 $N = A_5$. $\square$

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 $H_1, \dots, H_k$ are finite groups, with no assumed relationship to each other, we define $H_1 \times H_2 \times \dots \times H_k$ to be the cartesian product of sets, $\{(h_1, h_2, \dots, h_k) | h_i \in H_i\}$, with the product

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

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 $1 = (1, 1, \dots, 1)$ and $(h_1, h_2, \dots, h_k)^{-1} = (h_1^{-1}, h_2^{-1}, \dots, h_k^{-1})$. This group is called the *external direct product* of the groups $H_1, H_2, \dots, H_k$.

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 $|G| = |H_1||H_2| \dots |H_k|$, since this is true of the cartesian product of sets. If $(h_1, \dots, h_k) \in H_1 \times \dots \times H_k$, then $(h_1, \dots, h_k)^n = (h_1^n, \dots, h_k^n)$, which immediately implies that $|(h_1, \dots, h_k)| = \mathrm{lcm}(|h_1|, \dots, |h_k|)$ if all the $|h_i|$ are finite.

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

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

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

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

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

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 $H_1, H_2, \dots, H_k$. We say that G is the *internal direct product* of the subgroups $H_1, H_2, \dots, H_k$ if

(i) $H_i \trianglelefteq G$ for all $1 \le i \le k$;
(ii) $H_1 H_2 \dots H_k = G$; and
(iii) $H_i \cap H_2 \dots H_{i-1} H_{i+1} \dots H_k = \{1\}$ for all $1 \le i \le k$.

The comments made before the definition show that the external direct product $H_1 \times \dots \times H_k$ is the internal direct product of the subgroups $\overline{H_1}, \dots, \overline{H_k}$. 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

$$\begin{aligned}\phi((g_1, \dots, g_k)(h_1, \dots, h_k)) = \phi((g_1 h_1, \dots, g_k h_k)) &= g_1 h_1 g_2 h_2 \dots g_k h_k = g_1 g_2 \dots g_k h_1 h_2 \dots h_k \\ &= \phi((g_1, \dots, g_k))\phi((h_1, \dots h_k))\end{aligned}$$

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 6.6.

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 $(h_2,k_2) \in H \times K$,

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

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 $(h_1, k_1) * (h_2, k_2) = (h_1(k_1 \cdot h_2), k_1k_2)$.

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 $\phi(k)(h_1h_2) = \phi(k)(h_1)\phi(k)(h_2)$, or equivalently $k \cdot (h_1h_2) = (k \cdot h_1)(k \cdot h_2)$, 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

$$((h_1, k_1) * (h_2, k_2)) * (h_3, k_3) = (h_1(k_1 \cdot h_2), k_1k_2) * (h_3, k_3) = (h_1(k_1 \cdot h_2)((k_1k_2) \cdot h_3), k_1k_2k_3)$$

and

$$(h_1, k_1) * ((h_2, k_2) * (h_3, k_3)) = (h_1, k_1) * (h_2(k_2 \cdot h_3), k_2k_3) = (h_1k_1 \cdot (h_2(k_2 \cdot h_3)), k_1k_2k_3).$$

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

$$k_1 \cdot (h_2(k_2 \cdot h_3)) = (k_1 \cdot h_2)(k_1 \cdot (k_2 \cdot h_3)) = (k_1 \cdot h_2)((k_1k_2) \cdot h_3)$$

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 $\bar{i}$ 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 $\theta(\bar{i}) = \overline{3}^i$.

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 $\sigma_3 : \bar{i} \mapsto \overline{3i}$ 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

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

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 $\lambda = \bar{j}$ 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 $\{v_1, \dots, v_m\}$ 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 $(p^m - 1)(p^m - p) \dots (p^m - p^{m-1})$ for the number of ordered bases of $\mathbb{F}_p^m$, and hence this is the size of $|\operatorname{GL}_m(\mathbb{F}_p)|$. $\square$

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 6.6. 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 $P = \langle a \rangle = \{1, a, a^2, \dots, a^{p-1}\}$, 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 $Q = \langle b \rangle = \{1, b, b^2, \dots, b^{q-1}\}$, 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 6.6, 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 $Q = \{1, b, b^2, \dots, b^{12}\}$ 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 $\{a^i b^j | 0 \le i \le 2, 0 \le j \le 2\}$. 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 $p_1^{e_1}, \dots, p_m^{e_m}$ 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

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

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. $\square$

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) *$G' \operatorname{char} G$.*
(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 $G \geq G^{(1)} \geq G^{(2)} \geq \dots \geq G^{(n)} \geq \dots$ 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 $\{1\} = H_0 \trianglelefteq H_1 \trianglelefteq \dots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$ 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 $\{1\} = K_n \trianglelefteq K_{n-1} \trianglelefteq \dots \trianglelefteq K_1 \trianglelefteq K_0 = G$, 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 $\{1\} = G^{(n)} \trianglelefteq G^{(n-1)} \trianglelefteq \dots \trianglelefteq G^{(1)} \trianglelefteq G^{(0)} = G$ 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. $\square$

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. $\square$

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 $[G, K] \subseteq H$. $\square$

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 $Z_0 \leq Z_1 \leq Z_2 \dots$ of G which we call the upper central series.

Proposition 7.17. *Let G be a group and let $Z_0 \leq Z_1 \leq Z_2 \leq \dots$ 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, $G^n \leq H_n = \{1\}$. $\square$

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 $G = P_1 \times P_2 \times \dots P_n$, 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 $G_{i+1} \not\subseteq H$.

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 $G_{i+1} \not\subseteq H$, 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 $N \trianglelefteq G$). $\square$

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 $T_1, \dots, T_{n-1}$ and $U_1, \dots, U_{n-1}$ are equivalent. Then since $T_n = G/G_{n-1} = G/H_{n-1} = U_n$ also, we see that $T_1, \dots, T_n$ and $U_1, \dots, U_n$ 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, $V_1, V_2, \dots, V_p, L/(K \cap L) \cong U_m$. 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, $V_1, V_2, \dots, V_{n-2}, U_m$ is equivalent to $T_1, \dots, T_{n-1}$. 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 $V_1, \dots, V_{n-2}, T_n$. 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, $U_1, \dots, U_{n-1}$ is equivalent to $V_1, V_2, \dots V_{n-2}, T_n$.

Finally, since $T_1, \dots, T_{n-1}$ is equivalent to $V_1, V_2, \dots, V_{n-2}, U_n$, then $T_1, \dots, T_{n-1}, T_n$ is equivalent to $V_1, V_2, \dots, V_{n-2}, U_n, T_n$. Similarly, since $U_1, \dots, U_{n-1}$ is equivalent to $V_1, V_2, \dots V_{n-2}, T_n$, we have $U_1, \dots, U_n$ is equivalent to $V_1, V_2, \dots, V_{n-2}, T_n, U_n$. But obviously $V_1, V_2, \dots, V_{n-2}, U_n, T_n$ is equivalent to $V_1, V_2, \dots, V_{n-2}, T_n, U_n$. So $T_1, \dots, T_n$ and $U_1, \dots, U_n$ 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 $\mathbb{Z}_{p_1}, \mathbb{Z}_{p_2}, \dots, \mathbb{Z}_{p_k}$ in some order, where $p_1, p_2, \dots, p_k$ 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. $\square$

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 $1 = H_0 \trianglelefteq H_1 \trianglelefteq \dots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$ 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 $1 = H_0 \trianglelefteq H_1 \trianglelefteq \dots \trianglelefteq H_{n-1} \trianglelefteq H_n = G$ 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. $\square$

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 $\{a_0 + a_1x + a_2x^2 + \cdots + a_mx^m + \ldots \,|\, a_i \in R\}$. 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 $\sum_{n=0}^{\infty} a_nx^n$. The addition and multiplication are as expected for power series; namely, $(\sum a_nx^n) + (\sum b_nx^n) = \sum(a_n + b_n)x^n$, and

$$(\sum a_nx^n)(\sum b_nx^n) = \sum_{n=0}^{\infty}[\sum_{i=0}^{n} a_ib_{n-i}]x^n$$

(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 $n_1, \dots, n_s$ 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 $\mathbb{Z}_n^\times = \{\overline{a} \mid \gcd(a, n) = 1\}$. 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 $fg \neq 0$. $\square$

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. $\square$

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. □

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 $I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots I_n \subseteq \dots$, 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 $I = (r_1, \dots, r_n)$ 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 $I = (p_1 + xq_1, \dots, p_n + xq_n)$, where $p_i \in P, q_i \in R$. Let $K = (p_1, \dots p_n)$ 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 $I_1, I_2, \dots, I_n$ 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) $I_1 I_2 \dots I_n = I_1 \cap I_2 \cap \cdots \cap I_n$.

(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. $\square$

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). $\square$

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 $\mathrm{lcm}(m, n) = mn$. 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 R be an integral domain and 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 $PX^{-1} = \{r/x \mid r \in P, x \in X\}$.

(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. 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 $D = \pm p_1 p_2 \ldots p_m$ for some nonempty set of distinct primes $p_1, \ldots, p_m$, 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 $f = x^m + a_{m-1}x^{m-1} + \dots + a_0 \in \mathbb{Z}[x]$, 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 10.9), 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. **Exercises.**

Exercise 10.7. 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.

Exercise 10.8. 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 10.9. 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 congruent 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. $N(z_1z_2) = N(z_1)N(z_2)$ 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).

11. 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 11.1. 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 11.2. *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). $\square$

Example 11.3. 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 11.2(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 11.4. 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 11.5. 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 11.6. 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 11.7. *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 $(a) = (b)$. $\square$

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 11.8. *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 $d \in (a, b)$. $\square$

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 11.27.

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.

11.0.1. *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 $a_1, a_2, a_3, \dots, a_n$ 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 11.9. 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 $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1) = x - 1$.

11.0.2. *Exercises.*

Exercise 11.10. 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 11.11. 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 11.12. Use the calculation in Example 11.9 to write find $u(x), v(x) \in \mathbb{Q}[x]$ such that $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1) = u(x)(x^5 - x^2 + 5x - 5) + v(x)(x^4 - 1)$.

11.1. **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 11.13. 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 11.14. 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 11.15. 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 11.16. 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 11.17. 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 11.18. $\mathbb{Z}$ 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. Thus 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 11.19. *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.*

11.1.1. *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 11.20. 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 $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots \subseteq I_n \subseteq \dots$, 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 11.21. *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 11.22. *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 $I_1 \subsetneq I_2 \subsetneq I_3 \subsetneq \dots \subsetneq I_n \subsetneq \dots$, 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 11.21; 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.

11.1.2. *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 11.23. *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 11.22) 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 $p_1p_2 \dots p_m = q_1q_2 \dots q_n$, 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 $1 = q_1q_2 \dots q_n$; 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 $p_1 | q_1q_2 \dots q_n$, 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 $q_1 = p_1x$, 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 11.21, and that irreducible elements are prime in a PID in Lemma 11.19. Thus (1) applies and shows that a PID is a UFD. □

11.1.3. *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 11.24. *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 11.19.

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 11.25. *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. $\square$

11.1.4. *Examples.* There are many examples of integral domains which are not UFDs. We think the following example is one of the simplest.

Example 11.26. Let F be a field. Let

$$R = \{f \in F[x] | f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m \text{ with } a_1 = 0\}.$$

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 11.27. 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 11.24). 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).

11.1.5. *Exercises.*

Exercise 11.28. Finish the proof of Lemma 11.25.

Exercise 11.29. 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.

12. Polynomial extensions

12.1. **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 12.1. 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 12.2. 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 11.25 directly to the case of finitely many elements.

Definition 12.3. 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 $C(f) = \gcd(a_0, a_1, \ldots, a_m) \in R$. 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 12.4. *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 $a_1, \ldots, a_n, b \in R$, $\gcd(ba_1, ba_2, \ldots, ba_n) = b\gcd(a_1, \ldots, a_n)$. 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 11.24 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 $C(f) = \gcd(a_0, a_1, \ldots, a_m)$ 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 $C(\widetilde{f}\widetilde{g}) \sim 1$. $\square$

We are now ready to prove Gauss's Lemma.

Lemma 12.5 (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 $a = t_1t_2 \dots t_m$ 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 12.4(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 $\lambda \in F$. □

12.2. **Factorization in $R[x]$.** Gauss's Lemma allows us to understand the irreducibles in $R[x]$ in terms of those of $F[x]$.

Corollary 12.6. *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. $\square$

We are now ready to prove the main theorem.

Theorem 12.7. *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 12.6, 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 12.6. 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 12.6. 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 12.6(2), g_i is an associate of h_i in $R[x]$ for all i as well, so we are done. $\square$

The main result of this section implies that there are many examples of rings that are UFDs and not PIDs.

Lemma 12.8. *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 12.7. 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. $\square$

Example 12.9. Given a ring R, we can define inductively a ring of polynomials in n variables over R by $R[x_1, \dots x_n] = (R[x_1, \dots, x_{n-1}])[x_n]$. If R is a UFD, then our main theorem gives that $R[x_1, \dots, x_n]$ is also a UFD for any n. In particular, if F is a field then $F[x_1, \dots, x_n]$ is a UFD. These rings play an important role in commutative algebra.

Rather than an inductive definition, one can also define $S = R[x_1, \dots, x_n]$ directly as follows. Let S be the set of all sums of the form $\sum_{(i_1,i_2,\dots,i_n)\in\mathbb{N}^n} r_{(i_1,i_2,\dots,i_n)} x_1^{i_1} x_2^{i_2} \dots x_n^{i_n}$, where $r_{(i_1,\dots,i_n)} \in R$ is 0 except for finitely many n-tuples $(i_1, \dots, i_n)$. (Recall that by our convention $0 \in \mathbb{N}$.) In other words, S consists of finite R-linear combinations of *monomials* $x_1^{i_1} x_2^{i_2} \dots x_n^{i_n}$. 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.

12.3. **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 12.10. *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. $\square$

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 12.11. *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 12.12. 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 12.13. *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 12.14. *Let R be a UFD with field of fractions F. Let $f = a_0 + a_1x + \cdots + a_mx^m \in R[x]$. 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 $0 = a_0t^m + a_1st^{m-1} + \cdots + a_{m-1}s^{m-1}t + a_ms^m$. This equation implies $s|a_0t^m$. Since $\gcd(s,t) = 1$, we get $s|a_0$. Similarly, the equation implies $t|a_ms^m$ and since $\gcd(s,t) = 1$ we have $t|a_m$. $\square$

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 12.15. 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 $s = \pm1, \pm2, \pm5, \pm10$ and $t = \pm1, \pm3$. 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 12.16. 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 12.17 (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 12.5), 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 12.18. $f(x) = 5x^7 + 3x^6 - 9x^3 + 6$ 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 12.6).

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 12.19. Let $f = x + x^2 y^{n-1} + y^n \in F[x, y] = (F[x])[y]$, 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 12.6 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 12.20. 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

$$g(z) = f(z+1) = ((z+1)^p - 1)/z = (z^p + \binom{p}{p-1} z^{p-1} + \cdots + \binom{p}{1} z + 1 - 1)/z = z^{p-1} + \binom{p}{p-1} z^{p-2} + \cdots + \binom{p}{1},$$

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 $\mathbb{Q}[x]$. $\square$

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 12.21 (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 12.17), 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 $\overline{R}[x]$. $\square$

Example 12.22. 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 12.16, 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 12.21.

Remark 12.23. 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 $f(x) = x^4 + 1$.