MATH 200 LECTURE NOTES

DAN ROGALSKI

1. Rings: definitions, examples, and basic properties

A ring is an object that captures many of the properties familiar to us from the systems of numbers, such as the integers and the rational numbers, that students first learn in school. 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. Thus examples of rings are everywhere throughout modern mathematics.

Definition 1.1. 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 tried in his algebra book to introduce the amusing term "rng" for a ring without identity, but it didn't catch on.) We will seldom encounter a need to use non-unital 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 1.2. 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 1.3. 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 1.4. The subset $2\mathbb{Z}$ of even integers in $\mathbb{Z}$, under the usual addition and multiplication, is a ring without identity.

Example 1.5. 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 of these notes if he forgets to do that consistently. It is best simply to assume all rings are nonzero.

Example 1.6. 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 1.7. Let R be a ring, for example any of the familiar number systems in Example 1.3, 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, where the proofs of the basic properties, such as associativity of multiplication, are exactly the same as in a linear algebra course.

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.)

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 1.8. 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 first 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 1.3 are all integral domains. We can ask what the zero-divisors are in some of our other examples so far.

Example 1.9. 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 $\overline{b} = \overline{0}$.

Example 1.10. Let F be any field (we will define fields in a moment, or the reader can just think of the explicit case where $F = \mathbb{R}$ is the real numbers). In the matrix ring $M_n(F)$, it is easy to see that a matrix is a zero-divisor if and only if it is singular.

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 1.11. 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 $U(R)$ of all units in a ring is easily seen to be a group under the multiplication operation of the ring. It is called the *units group* of R. Another common notation for this group is $R^{\times}$.

Definition 1.12. 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.* (In fact, an older term for division ring is *skew field.*) By convention the zero ring is not considered a field.

Let us investigate units in our previous examples.

Example 1.13. $U(\mathbb{Z}) = \{-1, 1\}$, while $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are fields.

Example 1.14. We claim that the units in $\mathbb{Z}_n$ are $U(\mathbb{Z}_n) = \{\overline{a} | \gcd(a, n) = 1\}$; this was already called the group of units modulo n earlier in the course. Note that if $\gcd(a, n) = 1$, then there are integers b, k such that $ka + bn = 1$. Then $\overline{1} = \overline{k}\,\overline{a} + \overline{b}\,\overline{n} = \overline{k}\,\overline{a}$ since $\overline{n} = \overline{0}$, so $\overline{k} = \overline{a}^{-1}$. Conversely, if $\overline{k}\overline{a} = \overline{1}$ then $ak - 1 = bn$ for some integer b, so $ak - bn = 1$ forcing $\gcd(a, n) = 1$.

In particular, when $n = p$ is a prime number, then $\mathbb{Z}_p$ is a field, since $U(\mathbb{Z}_p) = \mathbb{Z}_p - \{\overline{0}\}$.

Example 1.15. If F is a field, then the units in $M_n(F)$ are exactly the invertible matrices. In other words, the units group $(M_n(F))^{\times}$ is the general linear group $\mathrm{GL}_n(F)$. Since we saw above the that the singular (i.e. non-invertible) matrices are zero-divisors, every element in $M_n(F)$ is either a zero-divisor or a unit.

Division rings which are not fields exist in abundance, but it is less obvious how to construct examples. The simplest and most famous example is the ring of quaternions $\mathbb{H}$, discovered by William Rowan Hamilton in 1843.

Example 1.16. 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 1.31.

Note that $\mathbb{H}$ contains the subset $\{\pm 1, \pm i, \pm j, \pm k\}$ which is isomorphic to the quaternion group Q under multiplication; this is how the quaternion group got its name.

There are various constructions which produce new rings from a given ring or rings. We already saw the example of the matrix ring $M_n(R)$ for a ring R. Here are some further examples.

Example 1.17. Let $\{R_\alpha | \alpha \in A\}$ be an indexed collection of rings. The *direct product* is the ring $\prod_{\alpha \in A} R_\alpha$ which is the Cartesian product of these sets, which 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.

Recall from our study of groups that there is also the *restricted product*, the additive subgroup of $\prod_{\alpha \in A} R_\alpha$ consisting of those elements such that all but finitely many coordinates are equal to 0. This is sometimes called the *direct sum* of the rings and is written $\bigoplus_{\alpha \in A} R_\alpha$. If the index set A is infinite, then this is a ring without unit under the coordinatewise operations.

Example 1.18. 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 $\{a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m + \dots \mid a_i \in R\}$. Note that no convergence is expected or implied, and we don't even think of these as functions in the variable x; an element of $R[[x]]$ is simply determined by the countable sequence of coefficients $(a_0, a_1, a_2, a_3, \dots)$, and the powers of x can be viewed as placeholders to help explain the multiplication rule. In fact, formally as an abelian group we can identify R with $\prod_{i=0}^{\infty} R$, the product of a countable number of copies of R.

We write an element of $R[[x]]$ as $\sum_{n=0}^{\infty} a_n x^n$. The addition and multiplication are as expected for power series; namely, $(\sum a_n x^n) + (\sum b_n x^n) = \sum (a_n + b_n) x^n$, and

$$(\sum a_n x^n)(\sum b_n x^n) = \sum_{n=0}^{\infty} [\sum_{i=0}^{n} a_i b_{n-i}] x^n$$

(note that only finite sums of elements in R are needed to define each coefficient of the product).

Example 1.19. Actually more important 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 $a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$. 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. Now $R[x]$ is is a ring under the same operations as for the power series ring, in other words $R[x]$ is a subring of $R[[x]]$ in the sense to be defined in the next section.

The next example gives an interesting link between group theory and ring theory.

Example 1.20. 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 15 of Isaacs or Chapter 18 of Dummit and Foote.

Keeping with our theme, we can ask what the zero-divisors and units look like for the examples above.

Example 1.21. 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 a zero-divisor in R_α or 0 (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 1.22. 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 $a_1, \dots, a_m$ are nilpotent in R. This is most easily proved after we have seen a bit more theory (see Exercise 4.10). *McCoy's Theorem* states that $\sum_{n=i}^m a_i x^i$ 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 1.30).

Example 1.23. 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 1.27). 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 1.24. Let G be a finite group and consider the group algebra $\mathbb{C}G$. Assume we can find an explicit isomorphism $\phi : \mathbb{C}G \to M_{n_1}(\mathbb{C}) \times \dots M_{n_s}(\mathbb{C})$. Then if $phi(x) = (A_1, \dots, A_s)$, x is a unit if and only if each A_i is an invertible matrix and x is a zerodivisor otherwise, i.e. if at least one A_i is singular. On the other hand, the structure of the unit group of RG, or its set of zero-divisors, for a general commutative ring R (even a field) and arbitrary, possibly infinite group G is a complicated subject about which there are many open questions.

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 1.25. *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 $fg \neq 0$. □

1.1. **Exercises.**

Exercise 1.26. 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 1.27. 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 1.28. 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 $\mathcal{K} = \{k_1, \dots, k_m\}$ 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 $\mathcal{K}_1, \dots, \mathcal{K}_r$ 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 $Z(RG) = \{a_1K_1 + \cdots + a_rK_r | a_i \in R \text{ for all } 1 \leq i \leq r\}$. In other words, the center consists of all R-linear combinations of the K_i.

Exercise 1.29. 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 1.30. Prove *McCoy's Theorem*: If $f = a_0 + a_1x + \cdots + a_mx^m \in R[x]$ 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 $g = b_0 + b_1x + \cdots + b_nx^n$ with $b_n \neq 0$. Suppose that $a_ig = 0$ for all i; then $a_ib_j = 0$ for all i, j and so $b_nf = 0$ and we are done. Thus some $a_ig \neq 0$ and we can take j maximal such that $a_jg \neq 0$. Then $f(a_jg) = 0$ but $\deg(a_jg) < \deg g$.)

Exercise 1.31. Let $\mathbb{H}$ be the ring of Hamilton's quaternions as in Example 1.16.

(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 $\mathbb{H}$.

2. BASIC RING TECHNOLOGY

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 2.1. 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 2.2. $\mathbb{Z}$ 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 2.3. 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.

Example 2.4. 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 2.5. 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 2.6. 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 the addition and multiplication operations of 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 2.7. 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, $\phi(0) = 0$ 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. 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. Note that the inclusion map of a non-unital subring R of a ring S is an example of a non-unital homomorphism.

Example 2.8. 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 2.9. 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 2.10. 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 2.11. 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 $e_1 + e_2 = 1$. 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 $f_1 = (1/2)(1+a)$ and $f_2 = (1/2)(1-a)$ 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 2.12. Let $\phi : R \to S$ be a homomorphism of rings. The *kernel* of ϕ is $\ker \phi = \{r \in R | \phi(r) = 0\}$ and the *image* of ϕ is $\phi(R)$.

Definition 2.13. 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 2.14. 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 right but not left ideal. If I is an ideal of R, then $L = \{\left(\begin{smallmatrix} r & s \\ t & u \end{smallmatrix}\right) | r, s, t, u \in I\}$ is an ideal of S.

Example 2.15. 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, the sum of any finite number of ideals is an ideal.

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 2.16. 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 2.17. 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 all ideals of $\mathbb{Z}$ as a ring.

Example 2.18. 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.

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 2.19. *Let R be a ring with ideal I. Let R/I be the factor group of $(R, +)$ by its subgroup $(I, +)$. Thus $R/I = \{r + I | r \in R\}$ is the set of additive cosets of I, with addition operation $(r + I) + (s + I) = (r + s) + I$. Then R/I is also a ring, with multiplication $(r + I)(s + I) = rs + I$ and unit element $1 + I$. The surjective map $\phi : R \to R/I$ given by $\phi(r) = r + I$ is a homomorphism of rings.*

Proof. The main issue is to make sure the claimed multiplication rule is well defined. Let $r + I = r' + I$ and $s + I = s' + I$, so $r - r' \in I$ and $s - s' \in I$. Then $rs - r's' = r(s - s') + (r - r')s' \in I$ (note that we use that I is closed under both left and right multiplication by elements in R) and so $rs + I = r's' + I$. Having shown the multiplication is well defined, the ring axioms for R/I follow immediately from the axioms for R, and the fact that ϕ is a homomorphism follows directly from the definition. □

Example 2.20. 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 2.21. *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 2.22. 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 2.23. 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 2.24. 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 of it, 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, so the usual rules for exponents become the following 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 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 number such as $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ have characteristic 0. Here is another basic fact about the characteristic.

Lemma 2.25. *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 1.9. □

Remark 2.26. 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: A subring R of S must contain 1_S, and an ideal I of S that contains 1_S is all of S, because $s1 = s \in I$ for all $s \in S$. An ideal I is closed under addition and multiplication, however, so one can view it as a subring "without unit" (it could even have a unit element different from 1_S and be a non-unital subring of S). 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 2.27. *Let R be a ring with ideal I. There is a $1-1$ 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 we have $(R/I)/(J/I) \cong R/J$.

The ring-theoretic version of the 2nd isomorphism theorem exists, though it is not used so often. It is better interpreted in terms of our later study of modules, so we omit it here.

2.1. **Exercises.**

Exercise 2.28. Check the claims in Example 2.23 using the 1st isomorphism theorem.

Exercise 2.29. 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 2.30. 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 2.31. 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 $n \geq 2$.

3. Prime and Maximal Ideals

We begin this section with some important notational concepts for ideals.

Let R be a 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

$$(X) = \{r_1x_1s_1 + \cdots + r_nx_ns_n | x_i \in X, r_i, s_i \in R \text{ for all } i,\ n \geq 1\}.$$

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.

If R is a commutative ring this simplifies to

$$(X) = \{r_1x_1 + \cdots + r_nx_n | x_i \in X, r_i \in R \text{ for all } i,\ n \geq 1\}$$

and in this case 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 have $(x) = \{rx | r \in R\}$, which we also write 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.*

Note that if R is noncommutative, then an ideal (x) generated by one element must still be described in the more complicated way as

$$(x) = \{r_1xs_1 + \cdots + r_nxs_n | r_i, s_i \in R \text{ for all } i\},$$

as there is no way to combine the terms of the sum.

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 $\{x_1y_1 + \cdots + x_ny_n | x_i \in X, y_i \in Y\}$. In this notation, for any subset X of a ring we have $(X) = RXR$, and a principal ideal Rx in a commutative ring follows this notational rule as well. If I and J are ideals of a ring R, it is easy to check that the product IJ is a again an ideal. It is an additive subgroup by definition of a product, and closure under multiplication on either side follows from the fact that I is a left ideal and J is a right ideal.

We focus now on commutative rings for a bit and discuss special kinds of ideals. We call an ideal I of a ring R *proper* if $I \neq R$.

Definition 3.1. 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, even though it trivially satisfies the condition in the definition above.

We first make the following simple observation about the ideals of fields.

Lemma 3.2. *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 3.3. *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 2.27, 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 3.2.

(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 3.4. *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 3.5. 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 in one-to-one correspondence to the positive prime numbers, and they are all maximal ideals.

Example 3.6. Let F be a field and let $I = (x) \subseteq F[x]$. We saw in Example 2.24 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 3.7. 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 2.27, 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 3.8. *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 $IJ = (P + (x))(P + (y)) \subseteq P + (x)(y) = P + xRyR = P + xyR = P$. 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 3.9. 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. If R is an arbitrary ring, an ideal P is called prime if it satisfies the condition in Lemma 3.8: If $IJ \subseteq P$ for ideals I, J, then $I \subseteq P$ or $J \subseteq P$. An ideal is said to be maximal just as before, if it is maximal under inclusion among proper ideals.

A ring is called *prime* if 0 is a prime ideal; similarly as in Proposition 3.3, an ideal P is prime if and only if R/P is a prime ring. However, a prime ring is not necessarily a domain. 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. It is still true that maximal ideals are prime.

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 division ring or even a domain (Exercise 2.31).

3.1. **Exercises.**

Exercise 3.10. 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 3.11. 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 3.12. 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 ideals of R, and (x, y) is a maximal ideal.

4. Zorn's Lemma and applications

Given a ring, 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 4.1. 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 4.2. 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 4.3. 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 $x \in P$ 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 4.4. 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 4.5. *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 at least one 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. It is not terribly difficult to prove Zorn's lemma from the axiom of choice. You can find a proof in Chapter 11 in Isaacs' book.

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 not maximal element is acheived, we get a set $S = \{x_i | i \geq \mathbb{N}\}$ 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 over again, etc. 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 4.6. *Let R be a nonzero 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 4.4. 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 $r, s \in R$ and $x \in J$, again we have $x \in I_\alpha$ for some α, and so $rxs \in I_\alpha \subseteq J$. We see that J is again an ideal (note that it was important that J is the union of a chain of ideals, and not an arbitrary union of ideals).

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 $I_1 < I_2 < I_3 < \dots I_n < \dots$, 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, an element $x \in R$ is *nilpotent* if $x^n = 0$ for some $n \geq 1$. If R is commutative, then we claim that the set N of all nilpotent elements of R is an ideal. Note that certainly $0 \in N$ so that N is nonempty. If $x \in N$ and $r \in R$, then $x^n = 0$ for some $n \geq 1$, and so $(xr)^n = x^n r^n = 0$. Finally, if $x, y \in N$, where $x^m = 0$ and $y^p = 0$, then using the binomial theorem $(x - y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i (-y)^{n-i}$ (see Exercise 2.29), we see that $(x - y)^{m+p-1} = 0$, as every term in the sum either has $i \geq m$ or $j \geq p$. Thus N is an ideal as claimed. The ideal N is called the *nilradical* of R, and it has the following interesting alternative characterization.

Proposition 4.7. *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. Suppose that $x \in N$. Since $x^n = 0$ for some $n \geq 1$, for any prime ideal I we have $x^n \in I$. Now 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 hypothesis 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 4.6. 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. Suppose that $yz \in M$, but that $y \notin M$ and $z \notin M$. Then consider the ideals $M + yR$ and $M + zR$, which satisfy $M \subsetneq M + yR$ and $M \subsetneq M + zR$. By the maximality of M in $\mathcal{P}$, $M + yR \notin \mathcal{P}$, so there is $x^m \in M + yR$ for some $m \geq 0$. Similarly, $x^n \in M + zR$ for some $n \geq 0$. But we have $(M + yR)(M + zR) \subseteq M + xyR \subseteq M$, so $x^{m+n} \in M$, contradicting that $M \cap S = \emptyset$. Thus M is a prime ideal, as claimed. Moreover, M does not contain the element 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 4.8. 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 4.7 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 4.6 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 4.9.

4.1. **Exercises.**

Exercise 4.9. Show that $G = (\mathbb{Q}/\mathbb{Z}, +)$ has no maximal subgroups. (Hint: Suppose that M is a maximal proper subgroup of G. Since every element of G has finite order, show that if $x \in G - M$ then $N = M + \langle x \rangle$ satisfies $|N : M| < \infty$; but $N = G$ by maximality. However, if $|G : M| = n$, show that for any $y \in G$ there is $x \in G$ with $nx = y$, which implies $y = nx \in M$. So $M = G$, contradicting properness.)

Exercise 4.10. Let R be a commutative ring and let $S = R[x]$. Show that $f = a_0 + a_1 x + \cdots + a_m x^m$ 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 1.29. 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 4.11. 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 4.12. 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 4.11. 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 4.13. 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 4.14. Use the steps below to prove the following theorem: Let R be a commutative ring. If every prime ideal of R is finitely generated, then all ideals of R are finitely generated.

(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 infinitely 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.

5. 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 5.1. 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 5.2. *Let $I_1, I_2, \dots, I_n$ be ideals of a commutative ring R and assume that the I_j are pairwise comaximal. Then*

(1) $I_1 I_2 \dots I_n = I_1 \cap I_2 \cap \dots \cap I_n$.

(2) $R/(I_1 \cap I_2 \cap \dots \cap I_n) \cong R/I_1 \times R/I_2 \times \dots \times R/I_n$ *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 $t - r = ry + sx - r = r(y-1) + sx = -rx + sx \in I$ and $t - s = ry + sx - s = ry + s(x-1) = ry - sy \in J$. 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 4.6. Since M is maximal, it is a prime ideal. Now $I_2I_3 \dots I_n \subseteq M$ in particular. If $I_j \nsubseteq M$ for all $2 \leq j \leq n$, then choosing $x_j \in I_j \setminus M$ for all j, we get $x_2x_3 \dots x_n \in M$. But as M is prime, this implies that $x_j \in M$ for some j, a contradiction. Hence for some j, $I_j \subseteq M$. But now $I_1 + I_j \subseteq M$, contradicting that I_1 and I_j are comaximal. This proves the claim. (One could also use the characterization of primeness using ideals given in Lemma 3.8 to see that $I_2I_2 \dots I_n \subseteq M$ implies $I_j \subseteq M$ for some j).

By the case of 2 ideals, we get that $I_1(I_2I_3 \dots I_n) = I_1 \cap (I_2I_3 \dots I_n)$. Since $I_2I_3 \dots I_n$ is a product of a smaller number of pairwise comaximal ideals, we see that $I_1I_2 \dots I_n = I_1 \cap (I_2 \cap \dots \cap I_n)$ by induction on the number of ideals. This proves (1) in general.

Again applying the two ideal case, we have $R/(I_1 \cap I_2 \cap \dots \cap I_n) = R/(I_1 \cap (I_2I_3 \dots I_n)) \cong R/I_1 \times R/(I_2 \dots I_n)$. 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. □

Remark 5.3. The explicit connection of this result to internal direct products of groups is the following: essentially the proof shows that $R/(I_1 \cap I_2 \dots \cap I_n)$ is the internal direct product (as additive groups) of the additive subgroups $J_i/(I_1 \cap I_2 \dots \cap I_n)$, where $J_i = I_1 \cap I_2 \cap \dots I_{i-1} \cap I_{i+1} \cap \dots I_n$ for $1 \leq i \leq n$. Moreover, $J_i/(I_1 \cap I_2 \dots \cap I_n)$ is isomorphic to R/I_i as groups. But the fact that we get an isomorphism of rings, as well as the hypothesis that pairwise maximality of the ideals I_i is enough, are new aspects to the ring case separate from what is going on with the underlying groups.

Corollary 5.4. *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 $a\mathbb{Z} \cap b\mathbb{Z} = \mathrm{lcm}(a, b)\mathbb{Z}$. 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). □

Example 5.5. 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, of course, stated in the language of congruence). 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 5.2, there is an isomorphism $\phi : R/(I \cap J) \to R/I \times R/J$. In this case $I \cap J$ consists of integers which are multiples of m and n, and hence $I \cap J = mn\mathbb{Z}$ since $\operatorname{lcm}(m, n) = mn$. We seek an element x such that $\phi(x+mn\mathbb{Z}) = (x+m\mathbb{Z}, x+n\mathbb{Z}) = (a+m\mathbb{Z}, b+n\mathbb{Z})$. This equation shows that the element x we seek is unique only up to multiples of mn, which is not surprising.

The proof of Theorem 5.2 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 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 (as we will define soon).

5.1. **Exercises.**

Exercise 5.6. 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 5.7. 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 5.5. (Hint: one way is to find x' satisfying the first two congruences, then solve the pair of congruences $x \equiv x' \mod 77$, $x \equiv 3 \mod 13$.)

6. Localization

The familiar set of rational numbers $\mathbb{Q}$ consists of fractions a/b where $a, b \in \mathbb{Z}$ and b is nonzero. This seems like a simple construction on the surface, but hides the fact that 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.

We often face the same problem for 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 larger ring of functions which may be defined only locally on a neighborhood rather than globally.

There is a version of localization for a noncommutative ring, but it is more complicated and outside the scope of this course. So we focus on commutative rings. Given a commutative ring R, 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.

We now state our main result. We concentrate on domains here, where the proof is the most intuitive. This case will suffice for our applications to field theory later. The general case is quite important though, and will be studied more in Math 200C.

Theorem 6.1. *Let R be a commutative ring with multiplicative system X. Assume that R is a domain and that $0 \notin X$. There exists a ring RX^{-1}, called the* localization of R along X, *and a ring homomorphism $\phi : R \to RX^{-1}$, with the following properties:*

(1) *$\phi(x)$ is a unit in RX^{-1} for all $x \in X$.*

(2) *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 $\theta \circ \phi = \psi$.*

Proof. 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 $r_1x_2 = x_1r_2$. 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 $r_2x_3 = x_2r_3$. Then $x_1x_2r_3 = x_1r_2x_3 = r_1x_2x_3$, and thus $x_2(x_1r_3 - r_1x_3) = 0$. Since $x_2 \in S$ we have $x_2 \neq 0$, so $x_1r_3 - r_1x_3 = 0$ as R is a domain. Then $r_1/x_1 \sim r_3/x_3$ and so $\sim$ is transitive. 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 $[r_1/x_1] = [t_1/y_1]$ and $[r_2/x_2] = [t_2/y_2]$, then $r_1y_1 = x_1t_1$ and $r_2y_2 = x_2t_2$. Thus

$$(r_1x_2 + r_2x_1)(y_1y_2) = r_1x_2y_1y_2 + r_2x_1y_1y_2 = x_1t_1x_2y_2 + x_1y_1x_2t_2 = (t_1y_2 + y_1t_2)(x_1x_2).$$

Thus $[(r_1x_2 + r_2x_1)/(x_1x_2)] = [(t_1y_2 + y_1t_2)/(y_1y_2)]$ and addition is well-defined. Showing that multiplication is well-defined is even easier and left to the reader. Now it is routine to check that the ring axioms are satisfied by RX^{-1}, where the identity for addition is $[0/1]$ and the identity for multiplication is $[1/1]$. The reader should check the details.

We can now define the map $\phi : R \to RX^{-1}$ by $\phi(r) = [r/1]$. It is obvious 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 $[x/1]^{-1} = [1/x]$.

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 $[r_1/x_1] = [r_2/x_2]$, then $r_1x_2 = x_1r_2$ implies $\psi(r_1)\psi(x_2) = \psi(x_1)\psi(r_2)$, and hence $\psi(r_1)\psi(x_1)^{-1} = \psi(r_2)\psi(x_2)^{-1}$. Then 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 the unique such function: If θ' is any homomorphism with $\theta'\phi = \psi$, since $[r/x] = [r/1][1/x] = [r/1][x/1]^{-1}$ in RX^{-1}, and 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 $\theta' = \theta$. □

Note that in fact the homomorphism $\phi : R \to RX^{-1}$ constructed in the proof above is injective: if $\phi(r) = [r/1] = [0/1]$, then $r = 0$, so $\ker \phi = 0$. Then we can identify R with the subring $\phi(R)$ of RX^{-1}, so we can think of RX^{-1} as a larger ring in which the elements in X have become units. The second statement of the theorem shows that the ring RX^{-1} satisfies a certain universal property. You can think of it as saying that RX^{-1} is the smallest or most efficient overring of R inside of which the elements of X have become units.

When the localization RX^{-1} is used in practice, one tends to write its elements as fractions 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.

Example 6.2. 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 $r/1 = \phi(r)$.

The ring RX^{-1} really is a field (if it weren't, it would have a terrible name). This is easy to see, since if r/x is a nonzero element of this ring, then since $0/x = 0/1$ is the zero element, 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.

Remark 6.3. Almost everything we have done above goes through even if R is not a domain. Theorem 6.1 is still true if one removes the second sentence with the hypothesis that R is a domain and X does not contain 0. In this general case, the main difference is that the homomorphism $\phi : R \to RX^{-1}$ satisfying the universal property will no longer be an injective homomorphism, so we can no longer think of R as a subring of its ring of fractions. It is not hard to see why we must allow ϕ to be noninjective in general. If $x \in X$ is a zerodivisor, say $xt = 0$ where $x \neq 0$ and $t \neq 0$, then if $\psi : R \to D$ is any homomorphism for which $\phi(x)$ becomes a unit in D, then $\phi(x)\phi(t) = 0$ implies $\phi(t) = 0$, by multiplying by $\phi(x)^{-1}$.

The main change to the proof of Theorem 6.1 is that the equivalence relation on fractions needs to be defined differently, and so verification of the various steps requires slightly more work. The reader may wish to work through the details in Exercise 6.7 below.

6.1. **Exercises.**

Exercise 6.4. 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 6.5. 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 $PX^{-1} = \left\{ \frac{r}{x} \middle| r \in P, x \in X \right\}$.

(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 6.6. 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 $I \cap X = \emptyset$.

Exercise 6.7. This exercise asks you to prove Theorem 6.1, removing the hypothesis that R is a domain and X does not contain 0. Again you should define RX^{-1} as a set of equivalence classes of formal fractions r/x with $r \in R, x \in X$, except now one takes the equivalence relation $r/x \sim s/y$ if

$(ry - xs)t = 0$ for some $t \in X$. (Note that if R is a domain and $0 \notin X$, then this latter condition forces $ry = xs$ and we get the same equivalence relation as before). Addition and multiplication are defined exactly as previously.

(a). Check the details that this gives a well defined ring RX^{-1} and that the homomorphism $\phi : R \to RX^{-1}$ given by $\phi(r) = [r/1]$ satisfies the required universal property.

(b). Prove that $\ker \phi = \{t \in R | tx = 0 \text{ for some } x \in X\}$. Thus ϕ is injective if and only if every element of X is a non-zero-divisor.

(c). What happens when $0 \in X$?

7. 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 7.1. Let R be an integral domain. We say that R is a *Euclidean domain* if there is a function $d : R \to \mathbb{N} = \{0, 1, 2 \dots\}$, with $d(0) = 0$, such that for any $a, b \in R$ with $b \neq 0$, there exist q, r such that $a = qb + r$ with either $d(r) < d(b)$ or $r = 0$.

The function d is called the *norm function* for the Euclidean domain.

Example 7.2. 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 certainly $r = |r| < |b|$.

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 second most important example of a Euclidean domain is the ring of polynomials over a field.

Example 7.3. Let F be a field and let $R = F[x]$. For $0 \neq f \in F[x]$ define $d(f) = \deg(f)$, and set $d(0) = 0$. We claim that R is a Euclidean domain with respect to this norm function. Given $f, g \in F[x]$ with $g \neq 0$, we need $q, r \in F[x]$ such that $f = qg + r$, with $d(r) < d(g)$ or $r = 0$. The elements q and r can be found explicitly by polynomial long division, but here we just give an abstract existence proof. Let $S = \{f - tg | t \in F[x]\}$. Let r be an element of S with minimal value of $d(r)$ among elements of S. If $r = 0$, that is fine. Otherwise, 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 either $r = 0$ or $d(r) < d(g)$. Since $r = f - qg$ for some $q \in F[x]$, we now have $f = qg + r$ with $r = 0$ or $d(r) < d(g)$, as required.

It is sometimes useful to note that in this case q and r are actually unique. To see this, suppose that $f = q'g + r'$ as well 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 1.25. 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.

Having the two separate allowable conclusions $r = 0$ or $d(r) < d(b)$ in the definition of Euclidean domain seems a bit awkward, but it is convenient, as we see from the example of $F[x]$ above. In that example, all nonzero scalars $\lambda \in F$ have $d(\lambda) = 0$, so for any $f \in F[x]$ and $0 \neq g = \lambda \in F$ we get $f = qg + r$ with $q = (\lambda^{-1} f)$ and $r = 0$, but $d(r) = 0 = d(g)$. We could instead change d to a different function to avoid this (see the exercises below) but it is not worth the trouble.

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 $D = -1$ (we exclude $D = 1$, which is sometimes taken to be squarefree by convention). 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 $(a+b\sqrt{D})(c+d\sqrt{D}) = (ac + dbD + (ad+bc)\sqrt{D})$, 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 $a^2 - b^2 D \neq 0$ so that $(a^2 - b^2 D)^{-1}(a - b\sqrt{D}) = x^{-1}$ in $\mathbb{Q}(\sqrt{D})$.

The norm is also multiplicative:

$$N((a + b\sqrt{D})(c + d\sqrt{D})) = N((ad + bcD) + (bc + ad)\sqrt{D})$$
$$= (ac + bdD)^2 - (bc + ad)^2 D = (a^2 - b^2 D)(c^2 - d^2 D) = N(a + b\sqrt{D})N(c + d\sqrt{D}).$$

In fact, when $D < 0$ so that $\sqrt{D}$ is imaginary, then $a - b\sqrt{D} = \overline{a + b\sqrt{D}}$ and $N(x) = ||x||^2$ where $||\ ||$ is the complex norm.

Definition 7.4. 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 can also uniformly 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 $\mathbb{O}_{\mathbb{Q}(\sqrt{D})}$ comes from number theory. The ring $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is the *integral closure* of $\mathbb{Z}$ inside $\mathbb{Q}(\sqrt{D})$. Explicitly, this means that $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is the set of all $\alpha \in \mathbb{Q}(\sqrt{D})$ such that α is a root of a *monic* polynomial $f = x^m + a_{m-1}x^{m-1} + \cdots + a_0 \in \mathbb{Z}[x]$, that is, a polynomial whose leading coefficient is 1. Such rings and their factorization theory (which we will study soon) are relevant to the study of certain diophantine equations. Integral closures are usually studied more extensively in Math 200C.

Note that if $x = a + b\sqrt{D}$ is a unit in $\mathbb{Z}[\sqrt{D}]$, then $1 = N(1) = N(x)N(x^{-1})$. Since $N(y) \in \mathbb{Z}$ for all $y \in \mathbb{Z}[\sqrt{D}]$, we conclude that $N(x) = \pm 1$. Conversely, since $N(a + b\sqrt{D}) = N(a - b\sqrt{D})$, if $N(x) = 1$ then $x = a + b\sqrt{D}$ is a unit in $\mathbb{Z}[\sqrt{D}]$ with inverse $x^{-1} = a - b\sqrt{D}$. Thus the units group of $\mathbb{Z}[\sqrt{D}]$ is $\{x \in \mathbb{Z}[\sqrt{D}] | N(x) = 1\}$.

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 7.5. The 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 $||(z - q)||^2 = ||(e + fi) - (s + ti)||^2 = (e - s)^2 + (f - t)^2 \leq 1/4 + 1/4 = 1/2$. 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 $N(r) < N(y)$, as required. □

Note that in this case the choice of q and r are not necessarily unique; this is already clear from the fact that 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 $\mathbb{O}_{\mathbb{Q}(\sqrt{D})}$ are Euclidean domains for a finite number of small values of D, but most of these rings are not Euclidean domains (or even PIDs in the sense we will study shortly). They are all what is known as *Dedekind Domains*, which is a more general class of rings is studied in Math 200C.

7.1. **Exercises.**

Exercise 7.6. Let R be a Euclidean domain with respect to the function $d : R \to \mathbb{N}$.

(a) Suppose that $d(x) = 0$ for $x \in R$ implies that $x = 0$. Show that for all $a, b \in R$, we can find $q, r \in R$ such that $a = qb + r$ with $d(r) < d(b)$, so we don't need to separate out the possibility $r = 0$ in the definition of Euclidean domain.

(b) Define $d'(x) = d(x) + 1$ for all $0 \neq x$, and $d'(0) = 0$. Show that R is also a Euclidean domain with respect to d', but now $d'(x) = 0$ implies $x = 0$.

Exercise 7.7. 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 7.8. Recall that when D is a squarefree integer, then the *ring of integers* in the field $\mathbb{Q}(\sqrt{D}) = \{x + y\sqrt{D} | x, y \in \mathbb{Q}\}$ is the subring $\mathcal{O} = \{a + b\omega | a, b \in \mathbb{Z}\}$ of $\mathbb{Q}(\sqrt{D})$, where $\omega = \sqrt{D}$ if D is congruent to 2 or 3 modulo 4, while $\omega = (1 + \sqrt{D})/2$ if D is 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. $N(z_1 z_2) = N(z_1)N(z_2)$ for $z_1, z_2 \in \mathbb{Q}(\sqrt{D})$.

(a). Consider the ring of integers $\mathcal{O}$ in $\mathbb{Q}(\sqrt{D})$. Suppose that for every $z \in \mathbb{Q}(\sqrt{D})$, there exists an element $y \in \mathcal{O}$ such that $|N(z - y)| < 1$. Prove that $\mathcal{O}$ is a Euclidean domain with respect to the function $d : \mathcal{O} \to \mathbb{N}$ given by $d(x) = |N(x)|$. (Hint: follow the method of proof we used to show that $\mathbb{Z}[i]$ is a Euclidean domain).

(b). Show that the ring of integers $\mathcal{O}$ is a Euclidean domain when $D = -2, 2, -3, -7$, or -11. (In each case show that part (a) applies).

8. 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 8.1. 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 3.5. More generally, we have the following result.

Proposition 8.2. *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). □

Example 8.3. 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 2.18 that it makes perfect sense to evaluate 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 8.2(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 8.4. 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 $\mathbb{Z}[i]/(2 - i) \cong \mathbb{Z}_5$.

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 8.5. 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 $a, b \in R$, not both 0, 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, it is still true that every common divisor divides 0 so that $\gcd(0, 0) = 0$ according to our definition.

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, $\gcd(a, b)$ 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 are exactly the other elements d' such that $(d') = (d)$. Let us tease out further exactly how this can happen.

Definition 8.6. 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 8.7. *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 $(a) = (b)$. □

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.

Let us return to PIDs now.

Proposition 8.8. *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 $d \in (a, b)$. □

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 11.3.

Since a Euclidean domain is a PID, gcd's always exist in a Euclidean domain. In fact in this case there is 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 8.9. 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 $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1) = x - 1$.

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 most 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.

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.

8.1. **Exercises.**

Exercise 8.10. 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 8.11. 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.

(c). Prove that R is a PID if and only if R is both a UFD and a Bezout domain. (Hint: If R is a UFD and Bezout, given a nonzero ideal I, choose $0 \neq a \in I$ with a minimal number of irreducibles in its factorization. Given an arbitrary $b \in I$ consider the ideal (a, b).)

Exercise 8.12. Use the calculation in Example 8.9 to write find $u(x), v(x) \in \mathbb{Q}[x]$ such that $\gcd(x^5 - x^2 + 5x - 5, x^4 - 1) = u(x)(x^5 - x^2 + 5x - 5) + v(x)(x^4 - 1)$.

9. 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 9.1. 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 9.2. 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 9.3. 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 9.4. 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 9.5. 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 9.6. $\mathbb{Z}$ 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_1p_2 is also the product of irreducibles p_1', p_2' were $p_1' = up_1$, $p_2' = u^{-1}p_2$ for any unit u, so this kind of ambiguity cannot be avoided. It should be clear then that 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. First, here are some basic properties of prime and irreducible elements.

Lemma 9.7. *Let R be an integral domain.*

(1) *If $a \in R$ 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*.

The other 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; it will appear frequently in Math 200C as well.

Definition 9.8. 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 $I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots \subseteq I_n \subseteq \dots$, 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.

The term noetherian honors Emmy Noether, a German mathematician who 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 9.9. *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. □

We pause here to prove several different characterizations of the noetherian property, all of which are useful and interesting. We don't technically need to know all of this to prove the theorem that PIDs are UFDs, but given how important noetherian rings are, it is good to get a head start in understanding the intuition behind this condition.

Proposition 9.10. *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 9.9; indeed, that proof could have been subsumed into this result. 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 1 element, is generalized to the weaker condition that every ideal must be generated by some finite set of elements. Knowing that every ideal will have a finite generating set in a noetherian ring is often very useful in proofs.

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 9.11. *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 9.10) and so S has a maximal element, say (a). Now a is not itself irreducible (note that we consider a single irreducible to be a "product" of 1 irreducible) and so we can write $a = bc$ where b and c are both not units. Then $(a) \subsetneq (b)$, for if $(a) = (b)$, then c would be forced to be a unit. Similarly, $(a) \subsetneq (c)$. Since (a) is a maximal element of S, neither (b) nor (c) belongs to S, and neither b nor c is zero or a unit. Thus b and c are both finite products of irreducibles. But then $a = bc$ is a finite product of irreducibles as well, a contradiction. It follows that $S = \emptyset$ and so every nonzero nonunit element of R is a finite product of irreducibles.

Now suppose that $p_1 p_2 \dots p_m = q_1 q_2 \dots q_n$, where each p_i and q_j is irreducible, and hence also prime by hypothesis. Note that we allow the case that $m = 0$ or $n = 0$, so that one or the other product is empty and by convention equal to 1. We prove by induction on m that $m = n$, and after relabeling the q_j we have p_i is an associate of q_i for all i. If $m = 0$ then we have $1 = q_1 q_2 \dots q_n$; since a product of irreducibles cannot be a unit, this implies $n = 0$ and there is nothing further to show. Now we assume $m \geq 1$; similarly, this forces $n \geq 1$. Since p_1 is prime, the definition of prime extends by induction to prove that since $p_1 | q_1 q_2 \dots q_n$, we have $p_1 | q_i$ for some i. Relabel the q's so that q_i becomes q_1. Now $p_1 | q_1$ means $q_1 = p_1 x$, but since q_1 is irreducible, either p_1 or x is a unit. The element p_1 is irreducible and hence not a unit, so x is a unit and p_1, q_1 are associates.

Since we are in a domain, We may now cancel p_1 from both sides to get $p_2 p_3 \dots p_m = (x q_2) q_3 \dots q_n$ (some product could be empty). Since x is a unit and q_2 is irreducible, $x q_2$ 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 $x q_2$ 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 9.9, and that irreducible elements are prime in a PID in Lemma 9.7. Thus (1) applies and shows that a PID is a UFD. □

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 9.12. *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 9.7.

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 $a = up_1^{e_1}p_2^{e_2} \dots p_m^{e_m}$ for some $e_i \geq 1$, where p_i and p_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 p_i with associates and changing the unit u. Note that the unit u cannot be removed in general; in $\mathbb{Z}$ we have $-36 = (-1)(2^2)(3^2)$, and replacing 2 by -2 or 3 by -3 does not remove the unit in front.

Now we can also easily get that gcd's exist in a UFD.

Lemma 9.13. *Let R be a UFD. Then for every pair of elements $a, b \in R$, $\gcd(a, b)$ exists.*

Proof. If $a = 0$ then $\gcd(0, b) = b$, and similarly if $b = 0$. 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. We claim that $d = p_1^{g_1}p_2^{g_2} \dots p_m^{g_m}$ is a gcd of a and b. We leave the details to the reader. □

9.1. **Exercises.**

Exercise 9.14. Finish the proof of Lemma 9.13.

Exercise 9.15. 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.

10. Polynomial extensions

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 10.1. 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. 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 10.2. 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.

Definition 10.3. Let $f \in R[x]$ for a UFD R. Write $f = a_0 + a_1x + a_2x^2 + \dots + a_mx^m$ with $a_m \neq 0$. The *content* of f is $C(f) = \gcd(a_0, a_1, \dots, a_m) \in R$. As usual this is defined only up to associates.

For example, if $f = 12x^2 + 15x - 6 \in \mathbb{Z}[x]$, then $C(f) = 3$ (or -3).

Since a lot of things will hold "up to associates" in this section, we use the notation $a \sim b$ to indicate that elements a, b are associates in the ring R. If we need to emphasize in which ring R the elements are associates, we write $a \sim_R b$.

Lemma 10.4. *Let R be a UFD and let $f, g \in R[x]$. Let $a \in R$.*

(1) *$C(af) \sim_R aC(f)$.*
(2) *$C(fg) \sim_R C(f)C(g)$.*

Proof. (1) It is easy to verify fact that for $a_1, \dots, a_n, b \in R$, $\gcd(ba_1, ba_2, \dots, ba_n) = b\gcd(a_1, \dots, a_n)$. The formula in (1) is an immediate consequence.

(2) We may assume that $f \neq 0$ and $g \neq 0$, since otherwise the statements are trivial. Write $f = a_0 + a_1x + \dots + a_mx^m$ and $g = b_0 + b_1x + \dots + b_nx^n$, where $a_m \neq 0$ and $b_n \neq 0$. Since $C(f) = \gcd(a_0, a_1, \dots, a_m)$ divides every coefficient a_i, we can write $f = C(f)\widetilde{f}$ where $\widetilde{f} \in R[x]$ has content $C(\widetilde{f}) \sim 1$. Similarly, $g = C(g)\widetilde{g}$ for $\widetilde{g} \in R[x]$ with $C(\widetilde{g}) \sim 1$. Now $fg = C(f)C(g)\widetilde{f}\widetilde{g}$ and so using (1), $C(fg) \sim C(f)C(g)C(\widetilde{f}\widetilde{g})$. We see that it is enough to prove that the product of two polynomials with content 1 also has content 1.

Switch back to the original notation and assume $C(f) = C(g) = 1$. 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 $\widetilde{\phi}(f) = \overline{f} = \widetilde{\phi}(a_0 + a_1 x + \cdots + a_m x^m) = \overline{a_0} + \overline{a_1} x + \cdots + \overline{a_m} x^m$. 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 $C(f) = 1$, 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 $C(g) = 1$, $\overline{g} \neq 0$ in $R/(p)[x]$. But now note that since p is irreducible, it is a prime element by Lemma 9.12 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 tha some coefficient of fg is not divisible by p. Since p was arbitrary, $C(fg) = 1$ as desired. □

We are now ready to prove Gauss's Lemma.

Lemma 10.5 (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 $a = t_1 t_2 \dots t_m$ 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 $abf = (ag)(bh)$ and since $f \in R[x]$, $abf \in R[x]$. Applying Lemma 10.4 we have $abC(f) \sim C(abf) \sim C(ag)C(bh)$ up to associates. Now $ag = C(ag)g'$ for some $g' \in R[x]$ with $C(g') \sim 1$. similarly, $bh = C(bh)h'$ for $h' \in R[x]$ with $C(h') \sim 1$, and $f = C(f)f'$ with $C(f') \sim 1$. We now have $abC(f)f' = (ag)(bh) = C(ag)C(bh)g'h'$. Since $abC(f) \sim C(ag)C(bh)$, cancelling gives a unit $u \in R$ such that $f' = ug'h'$ or $f = C(f)g'uh'$. Let $g'' = C(f)g' \in R[x]$ and $h'' = uh \in R[x]$. We now have precisely that $f = g''h''$ 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 $\lambda \in F$. □

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

Corollary 10.6. *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 $C(f) = 1$.*

(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. $C(f) \sim 1$ up to associates. 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 suppose that $g = a \in R$ 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 $C(g) \sim 1$. Thus $s \sim r$ and hence λ is a unit in R. So f and g are associates in $R[x]$. The converse is trivial. □

We are now ready to prove the main theorem.

Theorem 10.7. *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 $r = p_1p_2\dots p_m$ 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 $r = p_1p_2\dots p_m$ 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 $C(f') = 1$, by Corollary 10.6, f' is also reducible over $F[x]$, so $f' = gh$ for $g, h \in F[x]$ with $\deg g < \deg f$ and $\deg g < \deg h$. 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, 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 $p_1p_2\dots p_mg_1g_2\dots g_n = q_1q_2\dots q_sh_1h_2\dots h_t$, 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 10.6. Taking contents of both sides we thus get $p_1p_2\dots p_m \sim q_1q_2\dots q_s$. 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 $g_1 g_2 \dots g_n \sim h_1 h_2 \dots h_t$. Each g_i and h_i is also irreducible in $F[x]$, by Corollary 10.6. 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 10.6(2), g_i is an associate of h_i in $R[x]$ for all i as well, so we are done. □

10.1. **Exercises.**

11. Examples

In this section we review some of the standard examples to keep in mind to help ground your understanding of factorization in commutative rings.

First, we can easily see that there are a whole slew of UFDs that are not PIDs.

Lemma 11.1. *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 10.7. 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 not a unit. □

There are also many examples of integral domains which are not UFDs at all. We think the following example is the simplest one to demonstrate the possible failure of uniqueness in a factorization into irreducibles.

Example 11.2. Let F be a field. Let

$$R = \{f \in F[x] | f(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m \text{ with } a_1 = 0\}.$$

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 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 11.3. Let $R = \mathcal{O}_{\mathbb{Q}(\sqrt{-5})}$. Thus $R = \mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} | a, b \in \mathbb{Z}\}$. In this ring we have the norm $N(a + b\sqrt{-5}) = a^2 + 5b^2$. Since an element is a unit if and only if it has norm 1, it is clear that $R^\times = \{\pm 1\}$.

Note that $9 = (3)(3) = (2 + \sqrt{-5})(2 - \sqrt{-5})$ in R. We claim that 3, $2 + \sqrt{-5}$, and $2 - \sqrt{-5}$ are all irreducibles in R. Because we know the units in R it is clear that none of them are associates of each other, so this will then imply that factorization in R is not unique.

Since $N(3) = 9$, if $3 = xy$ with $x, y \in R$ both nonunits, since $N(3) = N(x)N(y)$ we must have $N(x) = N(y) = 3$. But $a^2 + 5b^2 = 3$ has no solutions. So 3 is irreducible in R. Similarly, $N(2 + \sqrt{-5}) = 9 = N(2 - \sqrt{-5})$ and it follows that $2 + \sqrt{-5}$ and $2 - \sqrt{-5}$ are also irreducible as claimed.

We now see that R is not a UFD. We can also see that R has irreducible elements which are not prime (which also implies that R is not a UFD, by Lemma 9.12). We already saw that 3 is irreducible and that $3|(2 + \sqrt{-5})(2 - \sqrt{-5})$. Suppose that 3 is prime. Then $3|(2 + \sqrt{-5})$ or $3|(2 - \sqrt{-5})$. But if $2 + \sqrt{-5} = 3x$ for $x \in R$ then $N(x) = 1$ and so x is a unit, which is clearly not the case. Similarly $2 - \sqrt{-5}$ cannot be a multiple of 3 and so we have a contradiction. Thus 3 is irreducible but not prime.

Finally, using the same idea we can also give an example of a pair of elements in an integral domain which has no greatest common divisor. Let $a = 9$ and $b = 3(2 + \sqrt{-5})$. One may check that both principal ideals (3) and $(2 + \sqrt{-5})$ contain (a, b) and are minimal among principal ideals containing it. Thus there is no uniquely minimal principal ideal containing (a, b).

11.1. **Exercises.**

Exercise 11.4. Let n be a squarefree integer with $n > 3$ and let $R = \mathbb{Z}[\sqrt{-n}] = \{a + b\sqrt{-n} | a, b \in \mathbb{Z}\}$. (Note this is different from the ring of integers $\mathcal{O}_{\mathbb{Q}(\sqrt{-n})}$ when $n \equiv 1 \mod 4$).

(a). Prove that 2, $\sqrt{-n}$, $1 + \sqrt{-n}$, and $1 - \sqrt{-n}$ are all irreducible in R.

(b). Show that R is not a UFD.

(c). Find an element in R which is irreducible and not prime.

Exercise 11.5. Consider the ring $R = \mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} | a, b \in \mathbb{Z}\}$, in other words the ring of integers $\mathcal{O}_{\mathbb{Q}(\sqrt{-5})}$.

(a). Consider the ideals $I_2 = (2, 1 + \sqrt{-5})$, $I_3 = (3, 2 + \sqrt{-5})$, $I_3' = (3, 2 - \sqrt{-5})$. Show that $R/I_2 \cong \mathbb{Z}_2$, and $R/I_3 \cong R/I_3' \cong \mathbb{Z}_3$. Conclude that all three ideals are maximal ideals.

(b). Show that $R/(3) \cong \mathbb{Z}_3 \times \mathbb{Z}_3$ as rings. (Hint: Chinese Remainder theorem).

(c). Is $R/(2) \cong \mathbb{Z}_2 \times \mathbb{Z}_2$?

Exercise 11.6. This problem continues the investigations of the ring R in the previous problem.

(a). Prove that I_2, I_3, I_3' are all not principal ideals of R.

(b). Prove that $I_2^2 = (2)$, $I_2 I_3 = (1 - \sqrt{-5})$, $I_2 I_3' = (1 + \sqrt{-5})$, and $I_3 I_3' = (3)$. In particular, this gives multiple examples showing that a product of nonprincipal ideals can be principal.

(c). Consider the equality of products of principal ideals $(2)(3) = (1 + \sqrt{-5})(1 - \sqrt{-5})$. Show that expressing each of the ideals in this equation as a product of maximal ideals, one gets the same result on both sides of the equation up to rearrangement of the ideals.

Remark. The ring R is an example of a Dedekind domain. *Although unique factorization fails in the sense that R is not a UFD, there is a different kind of unique factorization: every nonzero ideal is a product of maximal ideals in a unique way up to the order of the factors. This is demonstrated by part (c): even though the element* 6 *factors in two essentially different ways (hence R is not a UFD), the equality of products of principal ideals* $(2)(3) = (1 + \sqrt{-5})(1 - \sqrt{-5})$ *leads to the same answer once everything is expressed in terms of products of maximal ideals. Dedekind domains are important in algebraic geometry and number theory and we will study them in more detail in Math 200c.*

Exercise 11.7. Suppose that R is a UFD with field of fractions F. A polynomial f is *monic* if it has leading coefficient 1; in other words $f(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n$.

(a). Suppose that $f \in R[x]$ factors as $f = gh$ with $g, h \in F[x]$. Show that the product of any coefficient of g with any coefficient of h is in R.

(b). Suppose that f, g, and h are as in part (a) and that moreover g and h are monic. Show that $g \in R[x]$ and $h \in R[x]$.

(c). Show that the ring $S = \mathbb{Z}[2\sqrt{2}] = \{a + b2\sqrt{2} | a, b \in \mathbb{Z}\}$ is not a UFD by finding $f \in S[x], g, h \in F[x]$, where F is the field of fractions of S, which violate the results above.

12. Modules

12.1. **Definition and first examples.**

Let R be a ring. In our initial study of modules, we will not assume that R is commutative. The concept of a left R-module is a "linearization" of the concept of a left group action on a set. We saw that studying group actions had a lot of consequences for the structure of the groups themselves. Similarly, to get a deeper understanding of rings, modules are essential.

Definition 12.1. Let R be a ring. A *left R-module* is an abelian group $(M, +)$ together with a left action of R on M, that is, a function $f : R \times M \to M$ where we write $f(r, m) = r \cdot m$, such that for all $r, s \in R$ and $m, n \in M$,

(i) $r \cdot (s \cdot m) = (rs) \cdot m$;

(ii) $1 \cdot m = m$;

(iii) $r \cdot (m + n) = r \cdot m + r \cdot n$;

(iv) $(r + s) \cdot m = r \cdot m + s \cdot m$.

Notice that axioms (i) and (ii) are the same as for the action of a group on a set (although R is just a monoid, not a group, under multiplication). However, the set being acted on in this case is assume to be an abelian group, and the other two axioms are kind of generalized distributive laws. Namely, (iii) shows that each element of R acts linearly on M, and (iv) shows that the additive structure of the ring R is compatible with the action.

It is easy to check that the module axioms also force $0 \cdot m = 0$ and $-1 \cdot m = -m$ for all $m \in M$. When the module under discussion is clear, we often just write rm instead of $r \cdot m$ for the action of $r \in R$ on $m \in M$.

As usual, there is also a notion of a right R-module, defined using a function $f' : M \times R \to M$ given by $f'(m, r) = m \cdot r$ and with the obvious right-sided versions of axioms $(i)-(iv)$. In particular, axiom (i) becomes $(i)' : (m \cdot s) \cdot r = m \cdot (sr)$.

In group theory, recall that left and right actions on a set are essentially equivalent concepts, and there is a natural way to turn any left action into a right action; namely, if G acts on X on the left, then there is a right action $*$ with $x * g = g^{-1}x$. In module theory, on the other hand, left and right modules are more distinct concepts. There is something of a way to relate them, however.

Definition 12.2. Given a ring R, its *opposite ring* R^{op} is the ring with the same underlying abelian group as R, but with new product $*$ defined by $r * s = sr$.

It is easy to check that the opposite ring is a ring. Now if M is a left R-module, then we can define a right R^{op}-module structure on the same abelian group M by $m \cdot r = rm$ (where we identify the underlying sets of R and R^{op}). The main thing to observe is that for axiom $(i)'$, we get $(m \cdot s) \cdot r = (sm) \cdot r = r(sm) = (rs)m = m \cdot (rs) = m \cdot (s * r)$ as required. Note that this would not work if we did not use the opposite multiplication $*$.

The rings R and R^{op} are not isomorphic for a general ring, and so left and right modules are distinct concepts that might behave quite differently. When R is commutative, however, of course

$R = R^{op}$. In this case, given a left R-module M, we can freely turn it into a right R-module by acting on the other side, i.e. $(m \cdot r = rm)$.

We now give some important examples of modules.

Example 12.3. Let F be a field. A left F-module consists of an abelian group V together with an action of F on V which we call scalar multiplication in this case. Examining the module axioms, we see that the F-module V is exactly the same as a vector space over the field F.

Example 12.4. For any ring R, R is a left module over itself by left multiplication, i.e. with $r \cdot s = rs$. In this case module axiom (i) is the actual associativity of multiplication in R, and module axioms (iii) and (iv) are the actual distributive laws in the ring.

Similarly, R is a right module over itself by right multiplication.

Example 12.5. Let F be a field and let $V = F^n$ be the set of column vectors $\left\{ \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} | a_i \in F \right\}$. Then V is a left module over the $n \times n$-matrix ring $R = M_n(F)$, where $A \cdot v = Av$.

Similarly, the set of length n row vectors with entries in F is a right $M_n(F)$-module by right matrix multiplication.

Example 12.6. Let M be any abelian group. Given $n \in \mathbb{Z}$, recall that for $m \in M$ we have defined the nth multple of m by

$$nm = \begin{cases} \overbrace{m + m + \cdots + m}^{n} & n > 0 \\ 0 & n = 0 \\ \overbrace{(-m) + (-m) + \cdots + (-m)}^{|n|} & n < 0 \end{cases} \tag{12.7}$$

where these integer multiples of m are the additive analogs of the powers of an elements in a group. This defines a natural action of $\mathbb{Z}$ on M, and one can easily check that the module axioms hold. Conversely, given a $\mathbb{Z}$-module M, of course the underlying set of M is an abelian group, and the module axioms imply that the action of $n \in \mathbb{Z}$ on m must be given by (12.7).

In conclusion, a $\mathbb{Z}$-module is nothing more than an abelian group; the $\mathbb{Z}$-action comes for free and is uniquely determined. This is very useful because the theory of abelian groups will be subsumed into module theory, and our theorems about modules will have interesting applications to abelian groups.

Example 12.8. Let $\phi : R \to S$ be any ring homomorphism. Suppose that M is a left S-module. Then we can make M into an R-module by "restriction of scalars": for $m \in M$, $r \in R$, define

$r \cdot m = \phi(r)m$. The module axioms are immediate. This is called restriction of scalars since in the case where ϕ is the inclusion homomomorphism of a subring R of S, then we really are just restricting the action to a smaller ring.

This raises the question of whether given a left R-module, there is a natural way to make it into an S-module using the homomorphism ϕ. The answer is yes, but will require the theory of tensor products we develop later.

12.2. **Basic module technology.** As with any new algebraic structure, we want to have a basic theory of functions that preserve the structure, definitions of substructures and factor structures, and so on. We make these definitions for left modules, but there are obvious counterparts for right modules.

Definition 12.9. Let M and N be left R-modules. A function $f : M \to N$ is a *homomorphism of modules* if f is a homomorphism of abelian groups, and $f(rm) = rf(m)$ for all $r \in R$ and $m \in M$. If a homomorphism f is bijective it is called an *isomorphism.*

Example 12.10. Let $R = F$ be a field. If V and W are F-modules, that is vector spaces over F, then a function $f : V \to W$ is a homomorphism of F-modules if and only if it is a linear transformation of vector spaces over F.

Example 12.11. Let R be a left module over itself by left multiplication. For any fixed $x \in R$, the function $\phi_x : R \to R$ given by $\phi_x(r) = rx$ is a homomorphism of left R-modules. It is a homomorphism of abelian groups by one of the distributive laws in R, and for $s \in R$, $\phi_x(sr) = (sr)x = s(rx) = s\phi_x(r)$, so ϕ_x preserves the left R-action.

The map ϕ_x is called "right multiplication by x" for obvious reasons. Note that left multiplication by x will not be a left module homomorphism in general (unless R is commutative, or more generally if x is in the center of the ring R).

Example 12.12. We saw above that a $\mathbb{Z}$-module is just an abelian group with its canonical $\mathbb{Z}$-action. If $f : M \to P$ is a homomorphism of abelian groups, it is automatically a homomorphism of $\mathbb{Z}$-modules. For $n \in \mathbb{Z}$ and $m \in M$, the fact that $f(nm) = nf(m)$ follows from the properties of homomorphisms of groups.

Definition 12.13. Let M be a left R-module. A subset $N \subseteq M$ is a *submodule* of M is N is a subgroup of M under $+$, and for all $r \in R, x \in N$, we have $rx \in N$.

Thus a submodule of M is closed under $+$ and under the left R-action. Clearly a submodule N of M is an R-module in its own right under the same R-action restricted to N. Also, the inclusion map $i : N \to M$ is an R-module homomorphism.

Example 12.14. Let M be a left R-module. Then both $\{0\}$ and M are submodules of M. $\{0\}$ is called the *trivial submodule* and is usually just written as 0.

Example 12.15. Let $f : M \to P$ be a homomorphism of left R-modules. Then it is a homomorphism of groups and so we have the kernel defined as $\ker f = \{m \in M | f(m) = 0\}$ like usual. Then $\ker f$ is a submodule of M: it is an additive subgroup of M by group theory, and if $m \in \ker f$ and $r \in R$, then $f(rm) = rf(m) = r0 = 0$.

The image of f, $f(M) = \{x \in P | x = f(m) \text{ for some } m \in M\}$ is a submodule of P, since if $x = f(m)$, then $rx = rf(m) = f(rm)$.

Example 12.16. Let R be a left module over itself by left multiplication. Then a submodule of R is an additive subgroup I of R such that $rx \in I$ for all $r \in R$ and $x \in I$. This is what we called a *left ideal* when we studied rings. So left ideals of R are the same as submodules of R as a left module over itself.

Example 12.17. If F is a field and V is an F-module, in other words a vector space over F, then a submodule of V is the same as a subspace of V as defined in linear algebra.

Example 12.18. Let $V = F^n$ be the set of length n column vectors over the field F, considered as a left module over the ring $R = M_n(F)$ by left matrix multiplication. We claim that the only R-submodules of V are 0 and V. To see this, suppose that $0 \neq v \in V$. It is a standard result of linear algebra that given any vector $w \in V$, there is some matrix $A \in M_n(F)$ such that $Av = w$. Indeed, if the ith entry of v is nonzero we can find such an A which is 0 except along its ith column. It follows that any submodule of V which contains v will be all of V. Since v was an arbitary nonzero vector, this proves the claim.

Definition 12.19. A left R-module M is called *simple* or *irreducible* if its only submodules are 0 and M.

We just saw an example of a simple module over $M_n(F)$, where F is a field.

Definition 12.20. Let N be a submodule of a left R-module M. Then the quotient of abelian groups M/N is an R-module via the action $r \cdot (m + N) = rm + N$. M/N is called a *quotient module* or *factor module*.

As usual, one must check that the definition of the R-action on M/N makes sense. We know that every subgroup of an abelian group is normal, so M/N is certainly a well-defined additive abelian group. Now if $m + N = m' + N$, then $(m - m') \in N$, so $r(m - m') \in N$ since N is a submodule. But then $rm - rm' \in N$ and so $rm + N = rm' + N$. Thus the R-action is well-defined.

We also note that the quotient map $\pi : M \to M/N$ given by $\pi(m) = m + N$ is a homomorphism of modules with kernel N.

One nice aspect of module theory is that the substructures of a modules which are modules in their own right, submodules, are also the things that you can mod out by to get a factor module.

Example 12.21. Let R be a ring with a left ideal I. Then I is a submodule of R, considered as a left module via multiplication. Thus we have a factor module R/I with action $r \cdot (s + I) = rs + I$.

Note if I is just a left ideal (not an ideal), then R/I is not in general a ring, it is only a left R-module.

There are versions for modules of all of the basic homomorphism theorems. Here is the 1st isomorphism theorem.

Theorem 12.22. *Let $f : M \to N$ be a homomorphism of left R-modules, and let $P = \ker f$. Then there is an isomorphism of R-modules $\overline{f} : M/P \to f(M)$ given by $\overline{f}(m + P) = f(m)$.*

Proof. The 1st isomorphism theorem for groups tells us that $\overline{f}$ is well-defined and an isomorphism of abelian groups. We just need to check that $\overline{f}$ is a homomorphism of modules. But this is easy, since $\overline{f}(r(m + P)) = \overline{f}(rm + P) = f(rm) = rf(m) = r\overline{f}(m + P)$. □

Similarly, there are versions of the 2nd, 3rd, and 4th isomorphism theorems. For each one, one takes the corresponding isomorphism theorem for abelian groups and simply notes that everything works at the level of R-modules. We omit the statements here but will freely use these results when we need them.

12.3. **Module structure on Hom.**

Definition 12.23. Let R be a ring and let M and N be left R-modules. We define

$$\operatorname{Hom}_R(M, N) = \{f : M \to N | f \text{ is a homomorphism of modules over } R\}.$$

A priori, $\operatorname{Hom}_R(M, N)$ is just a set of functions. However, it naturally has additional structure, and this is very useful.

First, we note that $\operatorname{Hom}_R(M, N)$ is always again an abelian group. For this, given $f, g \in \operatorname{Hom}_R(M, N)$ we define a function $f + g \in \operatorname{Hom}_R(M, N)$ by $[f + g](m) = f(m) + g(m)$. This is sometimes called *pointwise* addition of functions, since for each element $m \in M$ (a "point") we simply define the sum of functions at that point by summing the images of that point under the two functions, using that N is an abelian group. Note that $f+g$ is again an R-module homomorphism, since

$$[f+g](m_1+m_2) = f(m_1+m_2) + g(m_1+m_2) = f(m_1) + f(m_2) + g(m_1) + g(m_2)$$
$$= f(m_1) + g(m_1) + f(m_2) + g(m_2) = [f+g](m_1) + [f+g](m_2)$$

and

$$[f+g](rm) = f(rm) + g(rm) = rf(m) + rg(m) = r(f(m) + g(m)) = r[f+g](m).$$

The identity element of $\mathrm{Hom}_R(M,N)$ is the identically zero function 0, and $-f$ is the function with $[-f](m) = -f(m)$. The group axioms for $\mathrm{Hom}_R(M,N)$ are immediate, and $\mathrm{Hom}_R(M,N)$ is again abelian since N is.

Now suppose that R is commutative. We claim that in this case $\mathrm{Hom}_R(M,N)$ even has an R-module structure. It is an abelian group as above, and we define for $r \in R$, $f \in \mathrm{Hom}_R(M,N)$ the function $rf \in \mathrm{Hom}_R(M,N)$ by $[rf](m) = rf(m)$, using the R-module structure of N. It is routine to check that rf respects addition, and note that $[rf](sm) = rf(sm) = rsf(m) = srf(m) = s[rf](m)$, so $rf \in \mathrm{Hom}_R(M,N)$. We have used here that R is commutative. The module axioms for $\mathrm{Hom}_R(M,N)$ are routine to check.

When R is not commutative, in general $\mathrm{Hom}_R(M,N)$ has no natural additional structure beyond being an abelian group. There are ways to give it a module structure when M or N is a *bimodule*; we will explore this in a homework exercise.

Suppose now that R is an arbitrary ring again. When $M = N$ we call a homomorphism of R-modules $f : M \to M$ an *endomorphism*. We may write $\mathrm{Hom}_R(M,M)$ as $\mathrm{End}_R(M)$. In this special case $\mathrm{End}_R(M)$ again has additional structure besides its abelian group structure: it is naturally a ring, called the *endomorphism ring* of M. The product is defined by composition: for $f, g \in \mathrm{End}_R(M)$ we let $fg = f \circ g$. It is obvious that a composition of two module endomorphisms of M is again an endomorphism. The ring axioms follow routinely. For the sake of example, let's check one of the distributive laws $(f+g)h = fh + gh$ for $f, g, h \in \mathrm{End}_R(M)$. Since it is two functions that are being claimed equal, we check by applying them to an arbitrary element of M.

We have

$$[(f+g)h](m) = [(f+g)\circ h](m) = (f+g)(h(m)) = f(h(m)) + g(h(m))$$
$$= (f\circ h)(m) + (g\circ h)(m) = [f\circ h + g\circ h](m) = [fh+gh](m)$$

and so $(f+g)h = fh+gh$. The reader should check the other ring axioms to convince themselves that nothing complicated is going on.

If R is commutative and M is an R-module, then $\mathrm{End}_R(M)$ is both an R-module and a ring, by the constructions above. This is a structure called an *R-algebra* which will be defined later.

Example 12.24. Let F be a field and let V be an n-dimensional vector space over F. Then $\mathrm{End}_F(V)$ consists of all F-linear transformations from V to itself. As we saw above, $\mathrm{End}_F(V)$ is a ring. Suppose we fix a basis $v_1, v_2, \dots, v_n$ for V. Given any $f \in \mathrm{End}_F(V)$, we have scalars $a_{ij}^f \in F$ defined by $f(v_j) = \sum_{i=1}^n a_{ij}^f v_i$. These form a matrix $(a_{ij}^f) \in M_n(F)$. This gives a map $\psi : \mathrm{End}_F(V) \to M_n(F)$, where $\psi(f) = (a_{ij}^f)$. One may check that ψ is an isomorphism of rings. This isomorphism does depend on the choice of fixed basis; there is no canonical or preferred isomorphism between the two rings.

Example 12.25. Let R be a left module over itself by left multiplication. We will show that $\mathrm{End}_R(R)$ is isomorphic as a ring to the ring R^{op}.

Define a map $\phi : \mathrm{End}_R(R) \to R^{op}$ by $\phi(f) = f(1)$, where we identify the underlying sets of R and R^{op}, so that $f(1) \in R = R^{op}$.

Then we claim that ϕ is a homomorphism of rings. The map ϕ is clearly additive, since $\phi(f+g) = [f+g](1) = f(1) + g(1) = \phi(f) + \phi(g)$. Now $\phi(fg) = [f\circ g](1) = f(g(1))$. Since f is a module homomorphism, $f(r) = f(r\cdot 1) = rf(1)$ for all r. Thus $f(g(1)) = g(1)f(1) = \phi(g)\phi(f) = \phi(f) * \phi(g)$, where $*$ is the multiplication in R^{op}. This proves the claim.

Finally, if $\phi(f) = 0$, then $f(1) = 0$ and so $f(r) = rf(1) = 0$ for all r, and $f = 0$, so ϕ is injective. If $r \in R$, then we have the "right multiplication by r" map $f : s \mapsto sr$ which we have seen is an element of $\mathrm{End}_R(R)$; and $\phi(f) = f(1) = r$. So ϕ is surjective. Thus ϕ is an isomorphism of rings.

12.4. **Modules as maps to endomorphism rings.** Recall from our study of groups that there were two ways of thinking about a (left) action of a group on a set X. In the definition, one focuses on the action of $g \in G$ on one element $x \in X$ at a time. The other point of view thinks of how g acts on all of X at once as a permutation of X, and puts the whole action together into a homomorphism of groups $G \to \mathrm{Sym}(X)$.

A module is like a linearization of a group action, and in fact a module can also be thought of in terms of a single homomorphism (of rings, in this case).

Theorem 12.26. *Let R be a ring and let M be a fixed abelian group. There is a bijective correspondence*

$$\{\textit{left } R-\textit{module structures on } M\} \xrightarrow{\Phi} \{\textit{(unital) ring homomorphisms } \theta : R \to \operatorname{End}_{\mathbb{Z}}(M)\}$$

where given an R-module structure on M, Φ sends it to the map $\theta : R \to \operatorname{End}_{\mathbb{Z}}(M)$ where $[\theta(r)](m) = r \cdot m$.

We hope the reader sees the similarity between this result and the corresponding result for group actions. The main difference is that the ring R has an underlying abelian group structure, the set M to be acted on is assumed to already be an abelian group, and the action is assumed to be compatible with these linear structures. Note that because an abelian group is automatically a $\mathbb{Z}$-module, it does make sense to consider the endomorphism ring $\operatorname{End}_{\mathbb{Z}}(M)$.

Proof. There are many details to check in this result, but they are all routine steps that follow from definitions. We will check that the function Φ makes sense, but leave the rest for the reader to verify.

Assume that M is an R-module, where the action of r on m is written as $r \cdot m$. Then as in the statement we define $\theta : R \to \operatorname{End}_{\mathbb{Z}}(M)$ where $[\theta(r)](m) = r \cdot m$.

First, why is does θ land in $\operatorname{End}_{\mathbb{Z}}(M)$? For this we need $\theta(r)$ to be a $\mathbb{Z}$-module homomorphism from M to itself, in other words a homomorphism of abelian groups. But $[\theta(r)](m_1 + m_2) = r \cdot (m_1 + m_2) = r \cdot m_1 + r \cdot m_2 = [\theta(r)](m_1) + [\theta(r)](m_2)$ by module axiom (iii), so this is fine.

Next, why is θ a unital homomorphism of rings? First, to see that θ respects addition, we want $\theta(r + s) = \theta(r) + \theta(s)$. We check this by applying to an arbitrary $m \in M$. So $[\theta(r + s)](m) = (r + s) \cdot m = r \cdot m + s \cdot m = [\theta(r)](m) + [\theta(s)](m) = [\theta(r) + \theta(s)](m)$, where we have used module axiom (iv) and the pointwise definition of addition in the endomorphism ring.

To see that θ respects multiplication, we want $\theta(rs) = \theta(r) \circ \theta(s)$, since the multiplication in $\operatorname{End}_{\mathbb{Z}}(M)$ is composition. Using module axiom (i), we check that $[\theta(rs)](m) = (rs)\cdot m = r\cdot(s\cdot m) = r \cdot [\theta(s)](m) = [\theta(r)]([\theta(s)](m)) = [\theta(r) \circ \theta(s)](m)$, as required.

Finally, to see that θ is unital, we want $\theta(1)$ to be the identity element of $\operatorname{End}_{\mathbb{Z}}(M)$, which is the identity function. We have $[\theta(1)](m) = 1 \cdot m = m$ for all $m \in M$, by module axiom (ii).

We have checked that Φ makes sense; i.e. given a module M there is a homomorphism of rings $\theta : R \to \mathrm{End}_{\mathbb{Z}}(M)$ defined by $[\theta(r)](m) = r \cdot m$. Notice that we used all of the module axioms (i)-(iv).

To show that Φ is a bijection, one may directly construct an inverse function Ψ. Given a homomorphism of rings $\theta : R \to \mathrm{End}_{\mathbb{Z}}(M)$, we let $\Psi(\theta)$ be the R-module struture on M, where $r \cdot m = [\theta(r)](m)$. We leave it to the reader to check the axioms of a module; the argument is basically already contained in the work above, which related each module axiom to some aspect of the homomorphism θ.

The fact that Ψ and Φ are inverse functions is then clear from their definitions. □

Next, we show how one may describe the structure of a module over a polynomial ring over a field. Fix a field F, and let $R = F[x]$ be the ring of polynomials in one variable over F. Suppose that V is a left R-module. Since we can identify F with the subring of $F[x]$ given by constant polynomials, by restricting scalars the $F[x]$-module module V is also an F-module. Now define $\phi : V \to V$ by $\phi(v) = x \cdot v$, where $\cdot$ is the action of R on V. Clearly ϕ respects sums. Note that $ax = xa$ in $F[x]$ for all scalars $a \in F$. Thus $\phi(av) = x \cdot (a \cdot v) = xa \cdot v = ax \cdot v = a \cdot (x \cdot v) = a\phi(v)$. In other words, $\phi : V \to V$ is a linear transformation, or alternatively an element of $\mathrm{End}_F(V)$.

The argument above shows that an $F[x]$-module V leads to an F-vector space and a choice of linear transformation of V. In fact, conversely, a vector space together with a choice of linear transformation uniquely determines an $F[x]$-module. We formalize this as follows.

Proposition 12.27. *Let F be a field. An $F[x]$-module is the same thing as an F-vector space V together with a choice of F-linear transformation $\phi : V \to V$.*

Proof. Let V be an $F[x]$-module with action $\cdot$. We saw above that V is an F-vector space by restriction of scalars, and that $\phi : V \to V$ defined by $\phi(v) = x \cdot v$ is a linear transformation of V. Now by the module axioms, $x^2 \cdot v = x \cdot (x \cdot v) = \phi(\phi(v)) = \phi^2(v)$. By induction we get $x^n \cdot v = \phi^n(v)$ for all $n \geq 0$ (where we define $\phi^0 = 1_V$.) It follows that the action of $F[x]$ on V can be described by the formula

$$(\sum_{i\geq 0} a_i x^i) \cdot v = \sum_{i \geq 0} a_i \phi^i(v). \tag{12.28}$$

Conversely, suppose we are given a vector space V and a linear transformation $\phi : V \to V$. Then we define an action of $F[x]$ on V by (12.28). It is routine to check the module axioms, so that this does make V into an $F[x]$-module.

If we start with an $F[x]$-module V, it determines a vector space structure on V and a linear transformation ϕ. If we use this data to define an $F[x]$ action on V using (12.28), we have already seen that the original $F[x]$-module action must be given by this formula. Conversely, if we start with a V and a ϕ and use it to determine an $F[x]$ action on V via (12.28), clearly restricting the action to F gives the original V, and (12.28) gives $x \cdot v = \phi(v)$, so we recover V and ϕ. Thus we have proved there is a bijection between $F[x]$ modules and choices of (V, ϕ) where V is an F-vector space and $\phi : V \to V$ is linear. □

There is another point of view on Proposition 12.27 that uses Theorem 12.26, which we would like to describe. We first note the following result about homomorphisms from a polynomial ring. We leave the proof as an exercise.

Lemma 12.29. *Let $\phi : R \to T$ be a homomorphism of rings. Suppose that $t \in T$ commutes with every element of $\phi(R)$. Then there exists a unique homomorphism of rings $\widetilde{\phi} : R[x] \to T$ such that $\widetilde{\phi}(r) = \phi(r)$ for $r \in R$ and $\widetilde{\phi}(x) = t$. Conversely, given any homomorphism $\psi : R[x] \to T$, $\psi(x)$ commutes with every element of $\psi(R)$.*

The result shows that to define a homomorphism from $R[x]$ to another ring T, it is equivalent to define a homomorphism from R to T, and choose any element in T which commutes with the image of R to send x to. This is a freeness property of the polynomial ring with respect to ring homomorphisms.

Now let us consider $F[x]$-modules again, where F is a field. We know that an $F[x]$-module action on an abelian group V can be described by a ring homomorphism $\widetilde{\theta} : F[x] \to \mathrm{End}_{\mathbb{Z}}(V)$, using Theorem 12.26. By Lemma 12.29, such a homomorphism is equivalent to a choice of ring homomorphism $\theta : F \to \mathrm{End}_{\mathbb{Z}}(V)$ and a choice of $\phi \in \mathrm{End}_{\mathbb{Z}}(V)$ which commutes with $\theta(F)$. By Theorem 12.26 again, θ corresponds to an F-module action, i.e. vector space structure, on V. The fact that ϕ commutes with $\theta(F)$ means that $[\phi \circ \theta(a)](v) = \phi(av)$ is the same as $[\theta(a) \circ \phi](v) = a\phi(v)$, in other words that ϕ respects the scalar multiplication. We see from this that an $F[x]$-module amounts to a choice of F-vector space V and an F-linear transformation $\phi : V \to V$, recovering Proposition 12.27.

Because an $F[x]$-module encodes a choice of linear transformation of a vector space, we are going to derive applications to linear algebra by proving later that modules over PIDs such as $F[x]$ have a tightly restricted struture.

Example 12.30. Let F be a field. Define an $F[x]$-module structure on the vector space $V = F^2$ of length-2 column vectors, where x acts by the linear transformation ϕ given by left multiplication by the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Explicitly, we have $x \cdot \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right] = \left[\begin{smallmatrix} a+b \\ b \end{smallmatrix}\right]$. What are the $F[x]$-submodules of V?

To answer this question, a little thought shows that an $F[x]$-submodule of V is a subset closed under the action of both F and x, in other words an F-subspace W such that $x \cdot W \subseteq W$, or $\phi(W) \subseteq W$. So $F[x]$-submodules are subspaces which are stable under the action of the linear transformation ϕ. In this case, it is not hard to work out that the only such subspaces are 0, V, and the 1-dimensional F-subspace $W = \{\left[\begin{smallmatrix} a \\ 0 \end{smallmatrix}\right] | a \in F\}$.

12.5. **Generation of modules and cyclic modules.**

Definition 12.31. Let M be an R-module. Given a subset $X \subseteq M$, the R-submodule *generated by* X is the unique smallest R-submodule of M containing X. Since the intersection of an arbitrary collection of submodules is again a submodule, it can be described as the intersection of all R-submodules of M containing X. We say that M is *finitely generated* if there is some finite subset of M which generates M; otherwise we say that M is *infinitely generated.* M is *cyclic* if it is generated by a subset with one element.

More explicitly, if we define $RX = \{r_1x_1 + \dots r_nx_n | n \geq 0, r_i \in R, x_i \in X\}$ to be the set of all finite sums of elements of R acting on elements in X, then it is easy to see that RX is the submodule of M generated by X. If $X = \{x_1, \dots, x_n\}$ is finite, we also write this submodule as $Rx_1 + \dots + Rx_n$. So M is cyclic if $M = Rx$ for some $x \in M$. We also call any submodule of M of the form Rx a cyclic submodule, as it is cyclic considered as a module in its own right.

Example 12.32. If M is a $\mathbb{Z}$-module, we have seen this is the just the canonical $\mathbb{Z}$-action on the abelian group M. Then $\mathbb{Z}$-submodules are the same as subgroups, so the submodule generated by a subset X is the same as the subgroup generated by the subset. Thus a $\mathbb{Z}$-module is cyclic if and only if it is cyclic as a group, i.e. either isomorphic to $\mathbb{Z}$ or to $\mathbb{Z}_n$ for some $n \geq 1$.

Example 12.33. Let R be a commutative ring, and let R be a left R-module by left multiplication. The R-submodule generated by $x \in R$ is Rx, in other words the principal ideal generated by x. If R is a PID, then we know that every ideal is principal, and so every submodule of R is a cyclic submodule of R. The ring R is a noetherian ring if and only if every ideal is finitely generated as an ideal, i.e. if and only if every submodule of R is finitely generated as an R-module.

Example 12.34. If $R = F$ is a field, then the submodule of an F-module (i.e. vector space) V generated by a subset X is just the span of X. So V is finitely generated as an F-module if and only if it is spanned by a finite subset, i.e. if and only if V is finite dimensional as a vector space. The cyclic submodules of V are the 1-dimensional subspaces (and the 0-subspace).

Example 12.35. Let R be an arbitrary ring, and let M be a cyclic left R-module, which is generated by $x \in M$. Then we can define a homomorphism $f : R \to M$ by $f(r) = rx$, where R is the usual left R-module structure on R. It is easy to see that f is an R-module homomorphism. The image of f is Rx, which is M by assumption, so f is surjective. Let $I = \ker f$. Then I is a left ideal of R, since kernels of homomorphisms are submodules. The 1st isomorphism theorem now tells us that $R/I \cong M$ as R-modules.

Conversely, for any left ideal I of R, we can form the factor module R/I, and this is a cyclic module generated by the element $(1 + I)$, since $r + I = r(1 + I)$ for all $r \in R$.

We see that the cyclic left R-modules are exactly the factor modules R/I for left ideals I, up to isomorphism.

Example 12.36. $(\mathbb{Q}, +)$ is an example of an infinitely generated $\mathbb{Z}$-module (i.e, abelian group). We will see in a few lectures that finitely generated abelian groups are easy to completely describe and classify. Infinite abelian groups are much more complicated, and there are still many open questions about their structure, many of which involve sensitive set-theoretic issues.

12.6. **Free modules.**

Definition 12.37. Given an indexed family $\{M_\alpha | \alpha \in I\}$ of left R-modules, the *direct product* is the cartesian product $\prod_{\alpha \in I} M_\alpha$, which is again an R-module under the coordinate-wise operations; in other words it is the direct product of abelian groups, with R-action $r \cdot (m_\alpha) = (r \cdot m_\alpha)$.

The *direct sum* of the family is the submodule of the direct product given by $\bigoplus_{\alpha \in I} M_\alpha = \{(m_\alpha) \in \prod M_\alpha | m_\alpha = 0 \text{ for all but finitely many } \alpha\}$. As an abelian group, we called this the *restricted product* of the groups M_α earlier, but the term direct sum is definitely the standard one in the context of modules.

Note that when we have a finite family $M_1, M_2, \dots, M_n$ of R-modules, the direct product and direct sum of this family are the same. We usually preference the direct sum notation and write this as $M_1 \oplus M_2 \cdots \oplus M_n$.

Free modules are the modules over a ring which are the most like vector spaces over a field. In general, a structure that is called "free" on a subset satisfies a certain universal property; we already saw the example of the free group when we studied group theory. For this reason we will take the universal property as our definition of a free module, and then show what they look like more explicitly.

Definition 12.38. Let F be a left R-module. Let X be a subset of F, and let $i : X \to F$ be the inclusion function. The module F is called *free* on a subset X if given any R-module M and a function $f : X \to M$, there is a unique R-module homomorphism $g : F \to M$ such that $g \circ i = f$. The *rank* of the free module F is the cardinality $|X|$ of the set X.

This universal property can be represented by the following commutative diagram:

$$\begin{tikzcd} X \arrow[r, "i"] \arrow[dr, "f"] & F \arrow[d, dashed, "\exists! g"] \\ & M \end{tikzcd}$$

In other words, given the inclusion function i and the homomorphism f, there exists a unique homomorphism g that completes the diagram to a commutative diagram (so $g \circ i = f$). Here the dashed arrow indicates that that homomorphism exists as a consequence of the property, and the ! indicates uniqueness.

The term "free" is used for properties like this because we may freely choose any elements of M whatsoever to send the elements in X to; then there is a unique homomorphism from F to M which does this to the elements in X.

Free objects are always determined uniquely up to isomorphism as a consequence of their universal properties, and the argument is always basically the same. Here is the result for reference, but it is really no different from our proof in group theory that the free group on a set is uniquely determined up to isomorphism by the cardinality of the set.

Proposition 12.39. *Let F be a free R-module on a subset X, and let G be a free R-module on a subset Y. If $|X| = |Y|$, then $F \cong G$ as R-modules.*

Proof. Let $i : X \to F$ and $j : Y \to G$ be the inclusion functions. Choose a set bijection $h : X \to Y$. Then the function $j \circ h : X \to G$ extends uniquely to a homomorphism $f : F \to G$, i.e. a homomorphism f such that $f|_X = h$. Similarly, $i \circ h^{-1} : Y \to F$ extends uniquely to a homomorphism $g : G \to F$. Now $g \circ f : F \to F$ is a homomorphism which restricts on X to the identity function; the identity homomorphism $1_F : F \to F$ is also such a homomorphism, so by the uniqueness property of the free module F, $g \circ f = 1_F$. Similarly, $f \circ g = 1_G$ since G is free. Thus f is an isomorphism with inverse g. □

Example 12.40. Let R be an R-module by left multiplication. Then R is free on the set $\{1\}$. To see this, let M be any R-module. Given $m \in M$, then $f : R \to M$ defined by $f(r) = rm$ is an R-module homomorphism such that $f(1) = m$. Moreover, it is the unique homomorphism sending 1 to m because $f(1) = m$ forces $f(r) = f(r1) = rf(1) = rm$.

We show now that it is easy to construct free modules explicitly, by extending Example 12.40 to a direct sum of copies of the module R.

Theorem 12.41. *Let I be any index set. Then $F = \bigoplus_{\alpha \in I} R$ is a free left R-module on the subset $X = \{e_\beta | \beta \in I\}$, where $e_\beta = (r_\alpha)_{\alpha \in I}$ with $r_\alpha = 0$ for $\alpha \neq \beta$ and $r_\alpha = 1$.*

You can think of the elements e_β as like "standard basis vectors" in a Euclidean space—they are 1 in exactly one coordinate and 0 elsewhere.

Proof. Let M be a module and let $f : X \to M$ be a function. We define $g : F \to M$ by $g((r_\alpha)) = \sum_\alpha r_\alpha f(e_\alpha)$. Note that this sum makes sense because $r_\alpha = 0$ for all but finitely many α. It is easy to see that g is an R-module homomorphism. Moreover, $g(e_\alpha) = f(e_\alpha)$ for all α, so g extends f. To see that g is unique, note that $(r_\alpha) = \sum_\alpha r_\alpha e_\alpha$, and so since g is an R-module homomorphism the formula $g((r_\alpha)) = \sum_\alpha r_\alpha f(e_\alpha)$ is forced. □

The theorem shows that for any set I, there is a free module on a subset X with cardinality $|I|$; and the free module is uniquely determined by the cardinality of that set by Proposition 12.39. So up to isomorphism, there is exactly one free module of any given rank, namely a direct sum of copies of the module R, over an index set of that rank.

There is another important way of thinking about free modules in terms of the concept of a basis.

Definition 12.42. Let F be an R-module. a subset X of F is called a *basis* of F if (i) X generates F as an R-module, so every $m \in F$ has an expression $m = r_1x_1 + \cdots + r_nx_n$ with $r_i \in R$ and $x_i \in X$; and (ii) whenever $r_1x_1 + \cdots + r_nx_n = 0$ for $r_i \in R$ and distinct $x_1, x_2, \ldots, x_n \in X$, then $r_i = 0$ for all i.

Example 12.43. Let V be a K-module, where K is a field, in other words a vector space over K. Then a subset X of V is a basis in the sense of the definition above if and only if X is a basis in the usual sense in linear algebra—(i) is the property that X spans V, and (ii) is the property that X is linearly independent.

We see that the idea of a basis of an R-module is modelled on the basis concept from linear algebra. We now show that it is precisely free modules that have a basis. So free modules are the objects in general module theory that behave most like vector spaces over a field.

Theorem 12.44. *An R-module F is free on a subset X if and only if X is a basis for F.*

Proof. Suppose that X is a basis for F. Property (i) of the definition of basis shows that an arbitrary $m \in F$ is an R-linear combination of elements in X. Thus m has an expression $m = \sum_{x \in X} r_x x$ with $r_x \in R$, where of course all but finitely many of the r_x are zero. But this expression is uniquely determined, for if also $m = \sum_{x \in X} r'_x x$, then $\sum_{x \in X}(r_x - r'_x)x = 0$, which shows that a finite R-linear combination of distinct elements in X is 0; by property (ii) of the definition of basis we get $r_x - r'_x = 0$ for all x, so $r_x = r'_x$ for all x.

Now there is a unique R-module homomorphism $g : \sum_{x \in X} R \to F$ such that $g(e_x) = x$, where e_x is the "standard basis vector" of the direct sum which is 1 in coordinate x and 0 elsewhere. Explicitly, $g(\sum_{x \in X} r_x e_x) = \sum_{x \in X} r_x x$. The fact that every $m \in F$ has a unique expression of the latter form shows that g is bijective. Thus g is an isomorphism of modules. Since $\sum_{x \in X} R$ is free on the subset $\{e_x | x \in X\}$ by Theorem 12.41, F is free on the subset $\{g(e_x) | x \in X\} = X$.

Conversely, if F is free on a subset X, then by Proposition 12.39 there is an R-module isomorphism $F \to \sum_{x \in X} R$ which sends $x \in X$ to the standard basis vector e_x of the direct sum. Since this is an R-module isomorphism, X is a basis of F if and only if $\{e_x | x \in X\}$ is a basis of $\sum_{x \in X} R$. But it is trivial to see that the latter subset satisfies (i) and (ii) of the definition of a basis. □

It is well-known that every vector space has a basis, so one consequence of the preceding result is that all F-vector spaces are free as modules over F. This is just saying something that should already be familiar to you about vector spaces, which is that to define a linear transformation from one vector space V to another W, it suffices to choose arbitrary destinations in W for the elements in a basis of V.

The reader may well have never seen a completely general proof of the fact that every vector space has a basis, however. Since we have Zorn's Lemma in our toolbox, this is not difficult.

Theorem 12.45. *Let V be a vector space over a field K. Then V has a basis X. Moreover V is a free K-module on X.*

Proof. If $V = \{0\}$ is a vector space of dimension 0, then by convention we consider the empty set as a basis. So this case is fine, and from now on we assume that V is a nonzero vector space.

Let S be the collection of all K-linearly independent subsets of V. S is nonempty since any single nonzero vector is an independent set, and we are assuming that $V \neq 0$. Order S by inclusion. Consider a chain $\{X_\alpha\}$ of elements of S. Let X be the union of all of the sets in the chain. We claim that $X \in S$. To see this, take distinct $v_1, \ldots, v_n \in X$, and suppose that $a_1 v_1 + \cdots + a_n v_n = 0$. Now each v_i belongs to some set in the chain. Since it is a chain, there is a single α such that $v_1, \ldots, v_n \in X_\alpha$. By definition then the set $\{v_1, \ldots, v_n\}$ is linearly independent. This forces $a_i = 0$ for all i. Thus X is also linearly independent. So $X \in S$ as claimed, and clearly X is an upper bound for the chain.

Now by Zorn's Lemma, S has a maximal element X, which is by definition a linearly independent set. Suppose that X does not span V, and let W be the span of X. Then we can pick some $v \in V \setminus W$. Now consider $Y = X \cup \{v\}$. We claim that Y is again linearly independent. Suppose that $a_1 x_1 + \cdots + a_n x_n = 0$ where $x_i \in Y$ are all distinct. If $x_i \in X$ for all i, then $a_i = 0$ for all i since X is independent. Otherwise we can assume that $x_n = v$, with $a_n \neq 0$, and $x_i \in X$ for $i < n$. Then $v = -(a_n)^{-1}(a_1 x_1 + \cdots + a_{n-1} x_{n-1}) \in W$, a contradiction. so Y is independent as claimed, but this contradicts the maximality of X. Thus X must span V, and so X is a basis of V. □

The key step in the proof above is the ability to invert the coefficient $a_n \neq 0$, since K is a field. The same proof would show that for an arbitrary ring R, given a module M, there exists a subset (possibly empty) which is maximal among subsets which are R-linearly independent in the sense of condition (ii) in the definition of basis. The submodule generated by this subset will be a free R-module, but there is no reason for it to equal M.

Free R-modules are certainly useful, but from the point of view of module theory perhaps not the most interesting. For each cardinality, there is a uniquely determined free module of that rank, which is a direct sum of copies of R over an index set of that cardinality. This is a very simple kind of classification result, since we know what they all look like up to isomorphism. We will work starting in the next section on the classification of finitely generated modules over PIDs, and that classification will be harder-earned and will have deeper consequences.

On the other hand, free modules over arbitrary rings R can behave in curious ways that defy the intuition we have from vector spaces. While there exists a unique free module up to isomorphism of each rank, there is no obvious reason that free modules of different ranks cannot be isomorphic. In fact, one can find a ring R such that $R \cong R \oplus R$ as left R-modules, and thus the free modules of rank 1 and rank 2 are isomorphic! For many rings R, however, it is true that two free modules are isomorphic if and only if they have the same rank. This is true for all commutative rings R, for example, which we leave as an exercise.

12.7. **Internal direct sums.** In our study of group theory, we gave conditions for a group G to be an internal direct product of a finite set of subgroups $H_1, \dots, H_n$. This result, applied in the case of abelian groups, extends immediately to modules, as we see in the next theorem.

Definition 12.46. Let M be an R-module and let $\{N_\alpha | \alpha \in I\}$ be an arbitrary collection of submodules of M. The *sum* of these submodules, $N = \sum_{\alpha \in I} N_\alpha$, is the submodule of M generated by all of the elements in the submodules N_α. More explicitly, N consists of all elements of the form $\sum_{\alpha \in I} n_\alpha$ such that $n_\alpha \in N_\alpha$ and $n_\alpha = 0$ for all but finitely many α.

Note that a direct sum $\bigoplus_{\alpha \in I} N_\alpha$ is the sum of its submodules N_α (identifying N_α with its image under the αeth injection i_α). But in general the sum of some collection of submodules of a module is not direct. The case where this does happen is called an internal direct sum.

Because it is the main case we will be concerned with below, we state the theorem on internal direct sums for finite sums only, just as we did for groups. The general case is not really more difficult, it is just notationally more awkward.

Theorem 12.47. *Let M be an R-module. Suppose that M has R-submodules $N_1, \dots, N_m$ with the properties that (i) $N_1 + N_2 + \cdots + N_m = M$ and (ii) $N_i \cap (N_1 + N_2 + \cdots + N_{i-1} + N_{i+1} + \cdots + N_m) = 0$ for all $0 \leq i \leq m$. Then $M \cong N_1 \oplus N_2 \oplus \cdots \oplus N_m$ as R-modules.*

Proof. Conditions (i) and (ii) are precisely the conditions for M to be an internal direct product as groups of the subgroups N_i (when written in additive form). Thus by our earlier study of such internal direct products, conditions (i) and (ii) force the natural map $\phi : N_1 \oplus N_2 \oplus \dots N_m \to M$ given by $\phi(n_1, n_2, \dots, n_m) = n_1 + n_2 + \cdots + n_m$ to be an isomorphism of abelian groups. Now one just notices that ϕ also preserves the R-action and so is an isomorphism of R-modules. □

Here is an application.

Definition 12.48. Let $f : M \to N$ be a homomorphism of left R-modules. Then f is called a *split surjection* if there is a homomorphism $g : N \to M$ such that $f \circ g = 1_N$.

Lemma 12.49. *Suppose that $f : M \to N$ is a split surjection, where $g : N \to M$ is a homomorphism with $f \circ g = 1_N$. Then f is surjective, g is injective, and $M \cong N \oplus K$ as R-modules, where $K = \ker(f)$.*

Proof. The fact that $f \circ g = 1_N$ immediately forces f to be surjective and g to be injective. We claim that M is the internal direct sum of its submodules $N' = g(N)$ and $K = \ker(f)$. Since g is injective, $N' \cong N$ as R-modules, so this will imply the result.

By Theorem 12.47 applied to two submodules, we just have to show that $N' + K = M$ and $N' \cap K = 0$. If $m \in N' \cap K$, then since $m \in N'$, $m = g(x)$ for some $x \in N$, so $f(m) = f(g(x)) = x$. But $m \in K$, so $x = f(m) = 0$. thus $m = g(x) = 0$. So $N' \cap K = 0$. For any $m \in M$, consider $y = m - g(f(m))$. Now $f(y) = f(m) - f(g(f(m))) = f(m) - f(m) = 0$ since $f \circ g = 1_N$. So $y \in \ker(f) = K$. But certainly $g(f(m)) \in g(N) = N'$. Thus $m = g(f(m)) + y \in N' + K$. So $N' + K = M$. □

There is a dual notion of split injection which also leads to a direct sum decomposition; we don't need it at the moment, so we postpone it until a later section where we examine results like this in the context of exact sequences.

One very useful consequence of this result is that surjections onto free modules are split.

Corollary 12.50. *Let $f : M \to F$ be a surjective homomorphism of R-modules, where F is a free R-module. Then f is a split surjection and hence $M \cong F \oplus K$ where $K = \ker(f)$.*

Proof. We just need to find $g : F \to M$ such that $f \circ g = 1_F$. Let F be free on the basis $\{x_\alpha | \alpha \in I\}$. Since f is surjective, for each α we can find an element $m_\alpha \in M$ such that $f(m_\alpha) = x_\alpha$. Now since F is free, there is a unique module homomorphism $g : F \to M$ such that $g(x_\alpha) = m_\alpha$. We have $f(g(x_\alpha)) = x_\alpha$ by definition, for all α. But since F is generated by the elements x_α, and $f \circ g$ is the identity on this subset, $f \circ g = 1_F$. □

13. Classification of modules over PIDs

13.1. Torsion.

In this section, for simplicity we will only consider modules over commutative rings R.

We have seen that when K is a field, then K-modules are vector spaces V. If V is a finitely generated K-module, this is just a finite dimensional vector space. This is a free K-module, and is very easy to describe and understand using a basis.

After fields, the commutative rings which are simplest in some sense are the principal ideal domains (PIDs). The goal of this section is to show that we can completely understand finitely generated modules over a PID R.

Let us first study some definitions that are useful for modules over general integral domains.

Definition 13.1. Let R be an integral domain, and let M be an R-module. An element $m \in M$ is called *torsion* if there is $0 \neq r$ such that $rm = 0$. The subset $\mathrm{Tors}(M) = \{m \in M | m \text{ is torsion}\}$ is called the *torsion submodule* of M. The module M is called a *torsion module* if $M = \mathrm{Tors}(M)$, and M is *torsionfree* if $\mathrm{Tors}(M) = 0$.

Lemma 13.2. *Let R be an integral domain and let M be an R-module.*

(1) $\mathrm{Tors}(M)$ *is an R-submodule of M.*

(2) $M/\mathrm{Tors}(M)$ *is a torsionfree module.*

Proof. (1) If $rm = 0$ and $sm' = 0$ with $0 \neq r, 0 \neq s$ and $m, m' \in \mathrm{Tors}(M)$, then $rs(m - m') = s(rm) - r(sm') = 0$ and $rs \neq 0$ since R is a domain, so $m - m' \in \mathrm{Tors}(M)$. Also, for any $t \in R$, $r(tm) = t(rm) = 0$, so $tm \in \mathrm{Tors}(M)$.

(2) Let $N = \mathrm{Tors}(M)$. Suppose that $m + N \in M/N$ is a torsion element of M/N. Then there is $r \neq 0$ such that $r(m + N) = 0$. This means that $rm \in N$. Then rm is torsion, so there is $s \neq 0$ with $s(rm) = 0$. Since $sr \neq 0$, m is torsion and so $m \in N$. Thus $m + N = 0$ in M/N and so M/N is torsionfree. □

Notice that the proof above works only because R is a domain. One could define torsion elements and modules in the same way over an arbitrary commutative ring, but one typically does not because Lemma 13.2 is what makes these definitions useful.

If M is an R-module, where R is commutative, then for any $m \in M$ we define the *annihilator of* m to be $\mathrm{ann}_R(m) = \{r \in R | rm = 0\}$. It is easy to see that $\mathrm{ann}_R(m)$ is an ideal of R. If R is an integral domain, then clearly m is torsion if and only if $\mathrm{ann}_R(m) \neq 0$. We can also define the *annihilator of* M to be $\mathrm{ann}_R(M) = \{r \in R | rm = 0 \text{ for all } m \in M\}$ which is also equal to $\bigcap_{m \in M} \mathrm{ann}_R(m)$.

Lemma 13.3. *Let M be a finitely generated module over an integral domain R. Then M is torsion if and only if* $\mathrm{ann}_R(M) \neq 0$.

Proof. Suppose that M is generated by $m_1, \dots, m_n$, so $M = Rm_1 + \cdots + Rm_n$. Suppose that M is torsion, and choose $0 \neq r_i$ such that $r_i m_i = 0$. Then $(r_1 r_2 \dots r_n)(s_1 m_1 + \cdots + s_n m_n) = 0$ for all $s_i \in R$ (since R is commutative). So $0 \neq r_1 r_2 \dots r_n \in \mathrm{ann}_R(M)$. Conversely, if $I = \mathrm{ann}_R(M) \neq 0$, then for any $0 \neq r \in I$ we have $rm = 0$ for all $m \in M$, so M is torsion. □

Lemma 13.3 does not necessarily hold for infinitely generated modules. In this case such a module M can be torsion and yet have $\mathrm{ann}_R(M) = 0$.

Example 13.4. Let R be an integral domain. Any free R-module is torsionfree. A cylic R-module is of the form R/I for some ideal I; we have $\mathrm{ann}_R(R/I) = I$ and so R/I is torsionfree if $I = 0$ and otherwise is torsion.

Example 13.5. Let $R = \mathbb{Z}$. A $\mathbb{Z}$-module is torsion if and only if as a group of all of its elements have finite order. A $\mathbb{Z}$-module is torsionfree if and only if every nonzero element of the abelian group has infinite order. A finitely generated torsion $\mathbb{Z}$-module is just a finite abelian group.

13.2. **Classification of modules over PIDs.** Our main goal is prove the following classification theorem.

Theorem 13.6. *Let R be a PID. Let M be a finitely generated R-module. Then*

(1)
$$M \cong \overbrace{R \oplus R \oplus \cdots \oplus R}^{r} \oplus R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \cdots \oplus R/(p_m^{e_m})$$
as R-modules, where the p_i are (not necessarily distinct) primes and $e_i \geq 1$. The number r is called the rank *of M and the prime powers $p_1^{e_1}, \ldots, p_m^{e_m}$ are called the* elementary divisors *of M.*

(2) *The rank and elementary divisors are uniquely determined by M (up to reordering the elementary divisors or replacing the primes by associates). Two modules M and N are isomorphic if and only if they have the same rank and elementary divisors (up to order and associates).*

The theorem shows that every finitely generated module over a PID is a direct sum of finitely many cyclic modules, where the torsion cyclic modules appearing have annihilators which are ideals generated by prime powers. Moreover, this decomposition is unique. It is a very strong structure theorem. If we determine the rank and elementary divisors of a module we essentially know everything we need to know about it.

Let us make some more comments about the theorem before working towards the proof. First, the hypothesis that the module M is finitely generated is essential to Theorem 13.6.

Example 13.7. Consider $\mathbb{Q}$ as a $\mathbb{Z}$-module. It is easy to show that given nonzero $p, q \in \mathbb{Q}$, there are $a, b \in \mathbb{Z}$ such that $ap = bq$. It follows that any two nonzero subgroups of $\mathbb{Q}$ have nonzero intersection. Because of this $\mathbb{Q}$ cannot be an internal direct sum of two $\mathbb{Z}$-submodules. Moreover, $\mathbb{Q}$ is clearly not a cyclic $\mathbb{Z}$-module itself. Thus $\mathbb{Q}$ cannot be expressed as a direct sum of cyclic $\mathbb{Z}$-modules. In particular, $\mathbb{Q}$ is not a free $\mathbb{Z}$-module.

Second, the classification theorem should not be expected to hold for integral domains that are not PIDs.

Example 13.8. Let K be a field and let $R = K[x,y] = (K[x])[y]$ be polynomials in two variables over K. We have seen that R is a UFD but not a PID. Consider the ideal $I = Rx + Ry$ of R as a module over R by left multiplication. As a module, I cannot be written as an internal direct sum of two nonzero modules; if $I = J \oplus L$ for nonzero ideals J and L, then in particular $J \cap L = 0$, but if $0 \neq a \in J$ and $0 \neq b \in L$ then $0 \neq ab \in J \cap L$, a contradiction. Moreover, I is not a cyclic module itself, as it is not a principal ideal. (If $I = (z)$ then $z|x$ and $z|y$, but x and y are non-associate irreducibles in R so this forces z to be a unit, and $I = (z) = R$, but this is absurd.)

We see that I cannot be expressed as a direct sum of cyclic modules, even though I is finitely generated as a module. In particular, I is not a free module.

13.3. **The torsionfree case.** Now we start to work towards the proof of the theorem, which will be accomplished through a series of subsidiary results.

The first step is to handle the torsionfree case.

Proposition 13.9. *Let R be a PID. Let M be a torsionfree R-module which is finitely generated by n elements. Then M is free of finite rank $\leq n$.*

Proof. Suppose that M is generated by n elements as an R-module, say $M = Rm_1 + \dots Rm_n$ for $m_i \in M$. The prove the result we induct on n. The base case is where M has 0 generators, in which case $M = 0$ and the result is trivial. Now assume that $n \geq 1$ and that the result is true for fewer than n generators.

Consider the factor module M/Rm_1. Let $\text{Tors}(M/Rm_1)$ be its torsion submodule. By the correspondence theorem for modules, $\text{Tors}(M/Rm_1) = K/Rm_1$ where K is a submodule of M containing Rm_1. Explicitly, $K = \{m \in M | rm \in Rm_1 \text{ for some nonzero } r \in R\}$. Now by Lemma 13.2, $(M/Rm_1)/(K/Rm_1)$ is torsionfree, but also by the 4th isomorphism theorem for modules, this module is isomorphic to M/K.

Now M/K is torsionfree and since $m_1 \in K$, M/K is generated by the $n-1$ elements $m_2 + K, \dots, m_n + K$. By the induction hypothesis, M/K is free of rank at most $n-1$. This implies that the natural surjection $\pi : M \to M/K$ is a split surjection, by Corollary 12.50. Then by Lemma 12.49, $M \cong M/K \oplus \ker(\pi) \cong M/K \bigoplus K$. To complete the proof, it now suffices to show that K is free of rank at most 1.

Since K is isomorphic to a summand of M, K is a surjective image of M and so K is also finitely generated. Then K/Rm_1 is a finitely generated torsion module, so by Lemma 13.3, it has nonzero annhilator. Say $0 \neq x \in \text{ann}_R(K/Rm_1)$. Then $xK \subseteq Rm_1$. Now $f : K \to xK$ given by $f(k) = xk$ is an isomorphism of modules, since M and hence K is torsionfree. Similarly, there is an isomorphism of modules $R \to Rm_1$ given by $r \mapsto rm_1$. Now we just need to show that every submodule of R is free of rank ≤ 1. But this is obvious, since a submodule is a principal ideal Ry, which is either 0 or free of rank 1. □

The hypothesis of finite generation is essential in Proposition 13.9. As we saw earlier in Example 13.7, $\mathbb{Q}$ is not a free $\mathbb{Z}$-module, but it is clearly a torsionfree $\mathbb{Z}$-module. The preceding result also certainly need not be true for integral domains that are not PIDs; Example 13.8 already gave the example of the finitely generated submodule $xR + yR$ of $R = K[x, y]$ which is not free.

Corollary 13.10. *Let R be a PID. If F is a free R-module of finite rank n, then every submodule of F is free of rank at most n.*

Proof. In particular, F is torsionfree and n-generated, so the result is imemdiate from Proposition 13.9. □

Unlike Proposition 13.9, it is possible to remove the finite rank assumption from Corollary 13.10; it is true that submodules of arbitrary free modules are free, for modules over a PID. We omit the proof of this result, which is not relevant for the classification of finitely generated modules over PIDs. On the other hand, the example $I = xR + yR \subseteq R = K[x, y]$ for a field K is a non-free submodule of the rank one free module R itself, so again for non-PIDs we don't have a result like Corollary 13.10.

13.4. **The torsion case.** The remaining work is to analyze the torsion part in more detail. Recall that a PID is a UFD, and the prime and irreducible elements are the same. We will use the term prime below.

Definition 13.11. Let R be a PID and let $p \in R$ be prime. An R-module M is called *p-primary* if for all $m \in M$, there exists $n \geq 1$ such that $p^n m = 0$.

If M is a p-primary module, then for every $m \in M$, $(p^n) \subseteq \operatorname{ann}_R(m)$ for some n, and so $\operatorname{ann}_R(M) = (p^i)$ for some i since every ideal containing (p^n) is generated by a divisor of p^n and hence a power of p. Similarly, if M is a finitely generated p-primary module, then $\operatorname{ann}_R(M) = (p^i)$ for some i.

The first step in understanding finitely generated torsion modules over a PID is to show that such a module decomposes as a direct sum of p-primary submodules.

Definition 13.12. Let R be a PID and let M be an R-module. If p is a prime element of R, the *p-primary component* of M is $M_p = \{m \in M \mid p^n m = 0 \text{ for some } n \geq 1\}$.

It is easy to see that M_p is the unique largest p-primary R-submodule of M.

Proposition 13.13. *let M be a finitely generated torsion module over a PID R. Then there are pairwise non-associate primes $p_1, \ldots, p_k$ of R such that $M \cong M_{p_1} \oplus \cdots \oplus M_{p_k}$.*

Proof. Since M is finitely generated torsion, $\operatorname{ann}_R(M) \neq 0$, say $\operatorname{ann}_R(M) = (a)$ with $a \neq 0$. By unique factorization we can write $a = p_1^{e_1} p_2^{e_2} \ldots p_k^{e_k}$ for some pairwise non-associate primes p_i and integers $e_i \geq 1$. We claim that M is the internal direct sum of the submodules $M_{p_1}, \ldots M_{p_k}$, where M_{p_i} is the p_i-th primary component of M.

First, define $q_i = p_1^{e_1} \ldots p_{i-1}^{e_{i-1}} p_{i+1}^{e_{i+1}} \ldots p_k^{e_k}$, i.e. q_i is the prime factorization of a with the $p_i^{e_i}$ term removed. It is clear that $\gcd(q_1, \ldots, q_k) = 1$, since the only primes (up to associates) that divide any q_j are the primes p_i, but p_i does not divide q_i. Since R is a PID, this means that $1 = b_1 q_1 + \cdots + b_k q_k$ for some $b_i \in R$. Now if $m \in M$, then $m = 1m = b_1 q_1 m + \cdots + b_k q_k m$. By definition $p_i^{e_i} b_i q_i m = b_i a m = 0$. Thus $b_i q_i m \in M_{p_i}$ for all i. It follows that $M = M_{p_1} + \cdots + M_{p_k}$.

Next, suppose that $m \in M_{p_1} \cap (M_{p_2} + \cdots + M_{p_k})$, then p_1^s kills m for some s since $m \in M_{p_1}$, and $p_2^{n_1} \ldots p_k^{n_k}$ kills m for some $n_i \geq 1$, since $m \in M_{p_2} + \cdots + M_{p_k}$. But $\gcd(p_1^s, p_2^{n_1} \ldots p_k^{n_k}) = 1$. Since every R-linear combination of these elements will also kill m, we have $1m = 0$ and so $m = 0$. By relabeling the primes, the same argument shows that $M_{p_i} \cap (M_{p_1} + \cdots + M_{p_{i-1}} + M_{p_{i+1}} + \cdots + M_{p_k}) = 0$ for all i. We have checked both conditions for an internal direct sum, so we see that M is an internal direct sum $M = M_{p_1} \oplus \cdots \oplus M_{p_k}$ as claimed. □

The last and perhaps most sensitive step is to show that a p-primary module is a direct sum of cyclic modules. We first make an observation about modules that are killed by an actual prime (not just a prime power).

Example 13.14. Let R be a PID and let $p \in R$ be a prime element. Then we claim that an R-module M such that $(p) \subseteq \operatorname{ann}_R(M)$ is the same thing as a vector space over the field $K = R/(p)$.

In fact this is just a special case of a general phenomenon. If I is an ideal of a commutative ring R, and M is an R-module such that $IM = 0$, i.e. $I \subseteq \operatorname{ann}_R(M)$, then M is naturally an R/I-module defined by $(r + I) \cdot m = rm$; the fact that $IM = 0$ is used to show that this action is well-defined. Conversely, any R/I-module N is also an R-module, by pulling back along the ring homomorphism $\phi : R \to R/I$, in other words defining $r \cdot x = (r + I)x$, and the resulting R-module is certainly killed by I. It is easy to see that in this way R/I-modules are in bijective correspondence with R-modules that are annihilated by I.

Apply this to R and $R/(p)$, noting that $R/(p) = K$ is a field since in a PID a prime element generates a maximal ideal, and that a K-module is the same as a vector space over K.

Lemma 13.15. *Let M be a finitely generated p-primary module. Suppose that we have elements $0 \neq g_i \in M$ such that the sum $\sum_{i=1}^{n} Rg_i$ is the internal direct sum of its cyclic submodules $Rg_1, \dots, Rg_n$. Assume that there are $h_1, \dots, h_n \in M$ such that $g_i = ph_i$ for all i. Then $\sum_{i=1}^{n} Rh_i$ is also the internal direct sum of its cyclic submodules $Rh_1, \dots, Rh_n$.*

Proof. The reader may easily check that condition (ii) of Theorem 12.47 for the submodules $Rh_1, \dots, Rh_n$ is equivalent to the following statement: if $x_1 + \cdots + x_n = 0$ for $x_i \in Rh_i$, then $x_i = 0$ for all i.

Suppose that $r_1h_1 + \cdots + r_nh_n = 0$. Acting by p, we have $r_1g_1 + \cdots + r_ng_n = 0$. Since $\sum_{i=1}^{n} Rg_i$ is the internal direct sum of its submodules $Rg_1, \dots, Rg_n$, we must have $r_ig_i = 0$ for all i. Since M is p-primary, $\operatorname{ann}_R(g_i) = p^{m_i}$ for some $m_i \geq 1$ (since $g_1 \neq 0$). So each r_i is a multiple of p. But then $r_ih_i \in Rg_i$ for all i. Again since $\sum_{i=1}^{n} Rg_i$ is the internal direct sum of its submodules $Rg_1, \dots, Rg_n$, this forces $r_ih_i = 0$ for all i. □

Proposition 13.16. *Let R be a PID with prime p, and let M be a finitely generated p-primary R-module. Then $M \cong R/(p^{s_1}) \oplus R/(p^{s_2}) \oplus \cdots \oplus R/(p^{s_k})$ as R-modules, for some list of positive integers $s_1, s_2, \dots, s_k$.*

Proof. We have $\operatorname{ann}_R(M) = p^n$ for some $n \geq 0$. The proof is by induction on n. We take $n = 0$ as the base case; this is when $M = 0$ and the result holds trivially with the empty list of integers.

Now assume that $n \geq 1$ and that the result holds for all smaller n. Let $N = pM = \{pm | m \in M\}$. Then N is a submodule of M for which $p^{n-1}N = p^nM = 0$. By the induction hypothesis, we have an internal direct sum $N \cong \bigoplus_{i=1}^{l} C_i$ for some cyclic submodules C_i, where $C_i \cong R/(p^{j_1})$ for some $j_1, \dots, j_l$ (it could be that $N = 0$ and the list is empty). Let g_i be a generator of C_i, so $\operatorname{ann}_R(g_i) = \operatorname{ann}_R(C_i) = (p^{j_i})$. Then $g_i = ph_i$ for some $h_i \in M$, since $N = pM$. It follows that $\operatorname{ann}_R(h_i) = (p^{j_i+1})$. By Lemma 13.15, the submodule $H = \sum_{i=1}^{l} Rh_i$ is also the direct sum of its cyclic submodules $Rh_1, \dots, Rh_l$.

Now consider the submodule $M[p] = \{m \in M | pm = 0\}$ of M. As discussed in Example 13.14, this can be thought of as a vector space over $K = R/(p)$. In particular, $(M[p] \cap H)$ is a K-subpace of $M[p]$ and so we can choose a complement in $M[p]$, that is, a K-subspace V of $M[p]$ such that $M[p] = (M[p] \cap H) \oplus V$ as K-modules (it is a standard result for vector spaces that any subspace has a complement; or use that all K-modules are free and apply Corollary 12.50).

We now claim that we have an internal direct sum $M = H \oplus V$. First, by the choice of $V \subseteq M[p]$ we have $V \cap H \subseteq V \cap (M[p] \cap H) = 0$. Next, if $m \in M$ then $pm \in N$. By the definition of

H, $pH = N = pM$, so there is $h \in H$ such that $ph = pm$ and hence $p(h - m) = 0$. Then $h - m \in M[p] = (M[p] \cap H) + V \subseteq H + V$. So $m \in H + V$. Hence $M = H \oplus V$ as claimed.

Finally, now V is a summand of M and so is also finitely generated as an R-module (and so also as a K-space). If $v_1, \ldots, v_s$ is a K-basis of V, then $V = Rv_1 \oplus \ldots Rv_s$ where $Rv_i \cong R/(p)$ as an R-module, for all i. Since $H = Rh_1 \oplus \cdots \oplus Rh_l$ where $Rh_i \cong R/(p^{j_i+1})$ for all i, we have $M \cong R/(p^{s_1}) \oplus \cdots \oplus R/(p^{s_k})$ for some positive integers $s_1, \ldots, s_k$. □

13.5. **Proof of the classification theorem.** We now put together all of the results we have proved to give the proof of the classification of finitely generated modules over PIDs.

Proof of Theorem 13.6. (1) Let M be a finitely generated module over the PID R. Let $T = \mathrm{Tors}(M)$. Of course T is torsion, and we know that M/T is torsionfree, by Lemma 13.2. Now M/T is finitely generated, since M is. It follows that M/T is free of finite rank, by Proposition 13.9. But then the quotient homomorphism $\pi : M \to M/T$ is a split surjection, and hence there is an internal direct sum decomposition $M \cong T \bigoplus F$, where $F \cong M/T$ is free of finite rank, say r. This implies also that $M/F \cong T$, so T is also isomorphic to a factor module of M and thus T is also finitely generated. Now by Proposition 13.13, $T \cong T_{p_1} \oplus \cdots \oplus T_{p_k}$ for some pairwise nonassociate primes p_i, where T_{p_i} is p_i-primary and is finitely generated since it is a summand of T. Finally, each T_{p_i} is isomorphic to a direct sum $T_{p_i} \cong R/(p_i^{s_{i,1}}) \oplus \cdots \oplus R/(p_i^{s_{i,k_i}})$, by Proposition 13.16.

This proves that M is a direct sum of a rank r free module and a finite number of cyclic modules with annihilators generated by prime powers.

(2). Note that it is obvious that if two modules have the same rank and the same elementary divisors, then the modules are isomorphic. To prove the converse, we take two modules

$$M = \overbrace{R \oplus R \oplus \cdots \oplus R}^{r} \oplus R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \cdots \oplus R/(p_m^{e_m})$$

and

$$M' = \overbrace{R \oplus R \oplus \cdots \oplus R}^{s} \oplus R/(q_1^{f_1}) \oplus R/(q_2^{f_2}) \oplus \cdots \oplus R/(q_n^{f_n})$$

where p_i, q_i are primes and $e_i, f_i \geq 1$. It suffices to prove that if $M \cong M'$, then $r = s$, $m = n$, and after renumbering one of the sequences of prime powers we have $e_i = f_i$ and p_i and q_i are associates for all i.

Let $f : M \to M'$ be an isomorphism. It is clear that f restricts to an isomorphism of the torsion submodules $f : T = \mathrm{Tors}(M) \to T' = \mathrm{Tors}(M')$. Then f induces an isomorphism of the factor modules $F = M/T \to F' = M'/T'$. Thus M and M' have isomorphic free and torsion parts. It is clear that $T = R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \cdots \oplus R/(p_m^{e_m})$ and $T' = R/(q_1^{f_1}) \oplus R/(q_2^{f_2}) \oplus \cdots \oplus R/(q_n^{f_n})$. Thus

$F \cong R^r$ and $F' \cong R^s$. A free module over a commutative ring has a uniquely determined rank by a homework exercise. Thus $r = s$ and the ranks are the same.

We have the isomorphism $f : T \to T'$ which we still call f. For any isomorphism of torsion modules and for any prime p, it restricts to an isomorphism $f : T_p \to T'_p$ between the p-primary components, by the definition of these components. Now notice that T_p is the direct sum of all of the summands $R/(p_i^{e_i})$ (if any) such that $(p_i) = (p)$, i.e. such that p_i is an associate of p. A similar comment holds for T'.

It now suffices to work one primary component at a time and show that if f is an isomorphism from $T_p = R/(p^{s_1}) \oplus \dots \oplus R/(p^{s_k})$ to $T'_p = R/(p^{t_1}) \oplus \dots \oplus R/(p^{t_l})$, then $k = l$ and after renumbering the t_i we have $s_i = t_i$ for all i. Equivalently, we just need that for each positive integer b, the number of s_i which is equal to b is the same as the number of t_i which is equal to b.

For each $b \geq 1$, if N is a p-primary module we can define $N[b] = \{x \in N | p^b x = 0\}$. By convention we put $N[0] = 0$. These are submodules of N with $0 = N[1] \subseteq N[1] \subseteq N[2] \subseteq \dots$. Also, each factor module $N[b]/N[b-1]$ is killed by p and so is a vector space over $K = R/(p)$. In particular, a short calculation shows that $T_p[b]/T_p[b-1]$ is a K-vector space of dimension equal to the number of s_j which are greater than or equal to b.

Now the isomorphism $f : T_p \to T'_p$ restricts to an isomorphism $T_p[b] \to T'_p[b]$ for all b, and hence induces an isomorphism $T_p[b]/T_p[b-1] \to T'_p[b]/T'_p[b-1]$ for all b as R-modules, hence also as K-vector spaces. It follows that the number of s_i which are greater than or equal to b is the same as the number of t_i which are greater than or equal to b, for all b. But this implies that the number of s_i which are equal to b is the same as the number of t_i which are equal to b. $\square$

13.6. **The invariant factor form.** There is another form of the classification theorem in which the torsion part is written as a direct sum of cyclic modules in a different way. For completeness we restate the theorem in its entirety in this version.

Theorem 13.17. *Let R be a PID. Let M be a finitely generated R-module. Then*

(1)

$$M \cong \overbrace{R \oplus R \oplus \dots \oplus R}^{r} \oplus R/(a_1) \oplus R/(a_2) \oplus \dots \oplus R/(a_n)$$

as R-modules, where the $a_i \in R$ are nonzero, nonunit elements such that $a_i | a_{i+1}$ in R for all i. The number r is called the rank *of M and the elements $a_1, \dots, a_n$ are called the* invariant factors *of M.*

(2) *The rank and invariant factors are uniquely determined by M (up to replacing the a_i by associates). Two modules M and N are isomorphic if and only if they have the same rank and invariant factors (up to associates).*

There are a few reasons why sometimes one might prefer the version of the classification in terms of invariant factors. In this version the torsion part is typically given as a direct sum of fewer cyclic modules. Also, the invariant factors occur in a specific order, unlike the ambiguity of the order in which the elementary divisors appear. We will see later how this leads to the uniqueness of the rational canonical form of a matrix, which has important applications.

In practice, if one is given the torsion part of a finitely generated module over a PID in elementary divisor form or in invariant factor form, it is routine to change to the other form. The reason is the following application of the Chinese remainder theorem.

Lemma 13.18. *Let R be a PID and let $a = p_1^{e_1} p_2^{e_2} \dots p_n^{e_n}$, where the p_i are pairwise non-associate primes. Then*

$$R/(a) \cong R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \dots R/(p_n^{e_n})$$

as both rings and as R-modules.

Proof. The fact that the primes are pairwise non-associate implies that $\gcd(p_i^{e_i}, p_j^{e_j}) = 1$ for $i \neq j$, and so the ideals $(p_i^{e_i})$ and $(p_j^{e_j})$ are comaximal. The Chinese remainder theorem now gives that the natural map $\phi : R/(a) \to R/(p_1^{e_1}) \oplus R/(p_2^{e_2}) \oplus \dots R/(p_n^{e_n})$ defined by the formula $\phi(r + (a)) = (r + (p_1^{e_1}), \dots, r + (p_n^{e_n}))$ is an isomorphism of rings. But it is clear that ϕ is also a homomorphism of R-modules. □

We give a proof of the invariant factor version of the fundamental theorem, but really seeing some examples in action may make this as clear as the rather technical proof. The reader might just skip this proof and move on to the examples we present afterwards.

Proof of Theorem 13.17. We only need to show that a finitely generated torsion module M over a PID can be expressed in invariant factor form and that the invariant factors are uniquely determined up to associates. The other parts of the theorem are the same as for Theorem 13.6.

We apply Theorem 13.6 to express M in terms of cyclic modules whose annihilators are the elementary divisors. Group together those elementary divisors which are associates of the same prime and change them if necessary so they are powers of exactly the same prime (which doesn't change the ideals they generate). We see that there are pairwise non-associate primes $p_1, \dots, p_m$

and exponents $e_{i,1}, \dots, e_{i,s_i}$ for each i such that M is a direct sum $M \cong \bigoplus_{i=1}^{m} \bigoplus_{j=1}^{s_i} R/(p_i^{e_{i,j}})$. We can order the exponents so that $e_{i,1} \geq e_{i,2} \geq \dots \geq e_{i,s_i}$. We also define $e_{i,j} = 0$ for $j > s_i$.

Now define $b_j = p_1^{e_{1,j}} \dots p_m^{e_{1,j}}$ for each $j \geq 1$, where as usual $p_i^0 = 1$ by convention. Then $b_i | b_{i-1}$ for all $i \geq 1$, with $b_j = 1$ for $j > n = \max\{s_i | 1 \leq i \leq m\}$. Define $a_i = b_{n+1-i}$ for $1 \leq i \leq n$. Then it is clear that $a_i | a_{i+1}$ for all $1 \leq i \leq n$. By Lemma 13.18,

$$\bigoplus_{j=1}^{n} R/(a_j) = \bigoplus_{j=1}^{n} R/(b_j) = \bigoplus_{j=1}^{n} \bigoplus_{i=1}^{m} R/(p_i^{e_{i,j}})$$

which is equal to the elementary divisor decomposition of M we started with if we ignore any zero summands. This proves that an invariant factor decomposition $M \cong R/(a_1) \oplus \dots \oplus R/(a_n)$ with $a_i | a_{i+1}$ for all i exists.

Conversely, suppose that $M \cong R/(c_1) \oplus \dots \oplus R/(c_t)$ is an invariant factor decomposition, with $c_i | c_{i+1}$ for all i. By using Lemma 13.18, we can break up each $R/(c_i)$ as a direct sum of cyclic modules with prime power annihilators. In this way we get an elementary factor decomposition. By the uniqueness of the elementary factor decomposition, the list of all of the prime powers occurring in the prime power decompositions of the c_i is the same as the list of all of the prime powers occuring in the prime power decompositions of the a_i. Now using the divisibility conditions, we see that the power of p_j occurring in c_i (possibly 0) is less than or equal to the power of p_j occurring in c_{i+1} for all i. The same is true of the a_i. There is only one way to arrange a sequence of prime powers in nondecreasing order of exponents. It is straightforward to see now that $(c_i) = (a_i)$ for all i and $t = n$. □

13.7. **Examples.** Applying the fundamental classification theorem in the case $R = \mathbb{Z}$ immediately gives us a classification theorem for finitely generated abelian groups. We did not discuss this when we studied group theory, since it is more convenient to obtain it as a consequence of module theory. There is no proof for Abelian groups that doesn't have to do more or less the same steps as a proof for modules over general PID's.

We restate the fundamental theorem for the case $R = \mathbb{Z}$ for convenience.

Theorem 13.19. *Let G be a finitely generated abelian group. Then $G \cong \mathbb{Z}^r \oplus H$ for some uniquely determined free abelian group $\mathbb{Z}^r$ of rank r and finite abelian group H. The group H is isomorphic to $\bigoplus_{i=1}^{m} \mathbb{Z}/(p_i^{e_i})$ for some prime powers $p_i^{e_i}$ (elementary divisors) uniquely determined up to their order. H is also isomorphic to $\bigoplus_{i=1}^{n} \mathbb{Z}/(a_i)$ for some uniquely determined integers $a_i \geq 2$ (invariant factors) satisfying $a_i | a_{i+1}$ for all i.*

Example 13.20. Here are some explicit examples of going between invariant factor form and elementary divisor form for a finite abelian group.

Suppose that $G = \mathbb{Z}/(3) \oplus \mathbb{Z}/(12) \oplus \mathbb{Z}/(60) \oplus \mathbb{Z}/(360)$ is a group given in invariant factor form. To find the elementary divisor form, we simply factor each invariant factor into prime powers, and take the list of all of those prime powers. We have $3 = 3^1$, $12 = 2^2 3^1$, $60 = 2^2 3^1 5^1$, and $360 = 2^3 3^2 5^1$. Thus the elementary divisors are $2^2, 2^2, 2^3, 3, 3, 3, 3^2, 5, 5$ and

$$G \cong \mathbb{Z}/(2^2) \oplus \mathbb{Z}/(2^2) \oplus \mathbb{Z}/(2^3) \oplus \mathbb{Z}/(3) \oplus \mathbb{Z}/(3) \oplus \mathbb{Z}/(3) \oplus \mathbb{Z}/(3^2) \oplus \mathbb{Z}/(5) \oplus \mathbb{Z}/(5)$$

as $\mathbb{Z}$-modules and hence abelian groups. This is justified by Lemma 13.18.

For an example of the reverse process, consider the abelian group given in elementary divisor form by

$$G \cong \mathbb{Z}/(5) \oplus \mathbb{Z}/(5^2) \oplus \mathbb{Z}/(5^2) \oplus \mathbb{Z}/(7^2) \oplus \mathbb{Z}(7^3) \oplus \mathbb{Z}/(11).$$

The elementary divisors are $5, 5^2, 5^2, 7^2, 7^3, 11$.

To find the invariant factors, it is easiest to find them in reverse order, as in the proof of Theorem 13.17. Take the product of the largest powers of each prime among the elementary divisors, then the product of the largest powers of the primes among the remaining elementary divisors, etc. In this case we have $b_1 = (5^2)(7^3)(11)$, $b_2 = (5^2)(7^2)$, $b_3 = 5$. The invariant factors are these integers in the reverse order: $a_1 = 5$, $a_2 = (5^2)(7^2) = 1715$, $a_3 = (5^2)(7^3)(11) = 94325$. So $G \cong \mathbb{Z}/(5) \oplus \mathbb{Z}/(1715) \oplus \mathbb{Z}/(94325)$.

In the next section we will explore the consequences of the fundamental theorem when applied to modules over a polynomial ring $K[x]$ for a field K. Here is an example of moving between invariant factors and elementary divisors in that context.

Example 13.21. Let $R = \mathbb{Q}[x]$. Consider the module $M = \mathbb{Q}[x]/(x^3 - 1) \oplus \mathbb{Q}[x]/(x^6 - 1)$, which is in invariant factor form, since $x^3 - 1 | x^6 - 1$ (as $x^6 - 1 = (x^3 - 1)(x^3 + 1)$).

To put this in elementary divisor form requires factorizing $x^3 - 1$ and $x^6 - 1$ as products of powers of prime (i.e. irreducible) polynomials in $\mathbb{Q}[x]$. We will discuss irreducibility for polynomials in more detail later (we ran out of time in the ring theory part last quarter, so we will do it in the field theory part this quarter). The only thing we use here is that a degree 2 polynomial over $\mathbb{Q}$ is irreducible if and only if it does not have a root in $\mathbb{Q}$.

By the standard formula for a difference of cubes, $x^3 - 1 = (x - 1)(x^2 + x + 1)$. It is easy to see that $x^2 + x + 1$ has no root in $\mathbb{Q}$, so it is irreducible over $\mathbb{Q}$. Similarly, $x^6 - 1 = (x^3 - 1)(x^3 + 1) = (x - 1)(x^2 + x + 1)(x + 1)(x^2 - x + 1)$, and $(x^2 - x + 1)$ is irreducible over $\mathbb{Q}$. We conclude that the elementary divisor form of M is

$$M \cong \mathbb{Q}[x]/(x-1) \oplus \mathbb{Q}[x]/(x-1) \oplus \mathbb{Q}[x]/(x+1) \oplus \mathbb{Q}[x]/(x^2+x+1) \oplus \mathbb{Q}[x]/(x^2+x+1) \oplus \mathbb{Q}[x]/(x^2-x+1).$$

It happens that all primes occur to the first power in this case.

14. Canonical Forms

In this section we will use the classification theorem of finitely generated modules over PIDs to develop the theory of canonical forms for linear transformations. These forms have many theoretical as well as practical uses in linear algebra.

14.1. **Linear algebra review.** Let V be a f.d. vector space over a field F. Suppose that $\phi : V \to V$ is an F-linear transformation. Fix an F-basis $\{v_1, \dots, v_n\} = \mathcal{B}$ for V. Although we just use set bracket notation for the basis, we always assume that the order of the basis vectors is fixed as well. We can define the matrix of ϕ relative to $\mathcal{B}$ to be $M_{\mathcal{B}}^{\mathcal{B}}(\phi) = (a_{ij}) \in M_n(F)$ where $\phi(v_j) = \sum_i a_{ij} v_i$.

Then if we identify $v \in V$ with the column vector $v_{\mathcal{B}} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \in F^n$, where $v = \sum_j b_j v_j$, then

$$\phi(v) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

that is, ϕ is given by left multiplication by the matrix $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$. For each fixed basis $\mathcal{B}$, this association of matrices to linear transformations gives a ring isomorphism $\Psi : \mathrm{End}_F(V) \to M_n(F)$ defined by $\Psi(\phi) = M_{\mathcal{B}}^{\mathcal{B}}(\phi)$, such that $\phi(v)_{\mathcal{B}} = M_{\mathcal{B}}^{\mathcal{B}}(\phi) v_{\mathcal{B}}$ for all v.

Let us recall what happens to the matrix when we change the basis. If $\mathcal{B}' = \{w_1, \dots, w_n\}$ is also a basis, then we let $P = (p_{ij})$ be the change of basis matrix whose coordinates are defined by $w_j = \sum_i p_{ij} v_i$. Similarly, we can define $Q = (q_{ij})$ to be the change of basis matrix defined by $v_j = \sum_i q_{ij} w_i$. Then $v_j = \sum_i q_{ij} \sum_k p_{ki} v_k$ and so $\sum_k (\sum_i p_{ki} q_{ij}) v_k = v_j$ forces $\sum_i p_{ki} q_{ij} = \delta_{kj}$, where this is the Kronecker δ symbol defined by $\delta_{kj} = \begin{cases} 0 & k \neq j \\ 1 & k = j \end{cases}$. This implies that $PQ = I$ where I is the $n \times n$ identity matrix, so P is invertible with $Q = P^{-1}$.

Now we calculate

$$\phi(w_j) = \sum_i p_{ij} \phi(v_i) = \sum_k \sum_i p_{ij} a_{ki} v_k = \sum_k \sum_i \sum_l q_{lk} p_{ij} a_{ki} w_l$$

which implies that

$$M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)_{lj} = \sum_k \sum_i q_{lk} a_{ki} p_{ij} = [P^{-1} M_{\mathcal{B}}^{\mathcal{B}}(\phi) P]_{lj}.$$

Thus the matrix with respect to the new basis, $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi) = P^{-1} M_{\mathcal{B}}^{\mathcal{B}}(\phi) P$, is a conjugate of the matrix associated to the old basis.

Definition 14.1. Matrices $A, B \in M_n(F)$ are *similar* if there is an invertible matrix $P \in \mathrm{GL}_n(F)$ such that $P^{-1}AP = B$.

Similarity is obviously an equivalence relation on the set of all $n \times n$ matrices. The above calculations showed that if two matrices represent the same linear transformation with respect to two different bases, then the matrices are similar. It is easy to see that the converse also holds.

Given that similarity is an equivalence relation, we can consider equivalence classes of matrices in $M_n(F)$ with respect to this relation, which we call *similarity classes.* The idea of canonical forms is to choose a representative of each similarity class with a particularly nice form. Then in proofs or calculations involving properties which are independent of similarity, one can reduce to the case of these canonical forms.

Definition 14.2. A matrix $A \in M_n(F)$ is *diagonal* if $A = (a_{ij})$ with $a_{ij} = 0$ for all $i \neq j$. Then $B \in M_n(F)$ is *diagonalizable* if B is similar to a diagonal matrix. A linear transformation $\phi \in \mathrm{End}_F(V)$ is called diagonalizable if there is a basis $\mathcal{B}$ of V such that $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is diagonal.

Diagonalizable matrices are usually the simplest ones to deal with when it comes to calculations. The first canonical form we study, the Jordan canonical form, will give a matrix in each similarity class which is as close to diagonal as possible in some sense. The Jordan form is also closely related to the theory of eignenvectors.

Definition 14.3. If V is a vector space over F and $\phi \in \mathrm{End}_F(V)$, then a nonzero vector $v \in V$ is an *eigenvector* of ϕ with *eigenvalue* λ if $\phi(v) = \lambda v$. Similarly, if $A \in M_n(F)$ and $0 \neq w \in F^n$ where the elements of F^n are written as column vectors, then w is an eigenvector of A with eigenvector λ if $Aw = \lambda w$.

It should be clear that $0 \neq v$ is an eigenvector of $\phi \in \mathrm{End}_F(V)$ if and only if $v_{\mathcal{B}}$ is an eigenvector of the matrix $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ for all choices of basis $\mathcal{B}$.

The following familiar result follows immediately from the definitions reviewed above.

Lemma 14.4. *A linear transformation $\phi \in \mathrm{End}_F(V)$ is diagonalizable if and only if V has a basis $\mathcal{B}$ consisting of eigenvectors of ϕ (in which case $M^{\mathcal{B}}_{\mathcal{B}}(\phi)$ is diagonal). Similarly, a matrix $A \in M_n(F)$ is diagonalizable if and only if F^n has a basis of eigenvectors for A.*

Definition 14.5. Let $A \in M_n(F)$. The *characteristic polynomial* of A is $\mathrm{charpoly}(A) = \det(xI - A) \in F[x]$. If V is a vector space over F of dimension n and $\phi \in \mathrm{End}_F(V)$, then the *characteristic polynomial* of ϕ is $\mathrm{charpoly}(\phi) = \det(xI - A)$ where $A = M^{\mathcal{B}}_{\mathcal{B}}(\phi)$ for any basis $\mathcal{B}$ of V.

Note that the choice of basis $\mathcal{B}$ in the definition above doesn't matter, because if matrices A and B are similar, they have the same characteristic polynomial, as

$$\begin{aligned}\det(xI - P^{-1}AP) &= \det(P^{-1}(xI)P - P^{-1}AP) = \det(P^{-1}(xI - A)P) \\ &= \det(P^{-1})\det(xI - A)\det(P) = \det(xI - A).\end{aligned}$$

If v is an eigenvector of A, then $Av = \lambda v$ implies $(A - \lambda I)v = 0$ and so v is a nonzero vector in the nullspace of $(A - \lambda I)$. Thus $(A - \lambda I)$ is singular. Conversely if $(A - \lambda I)$ is singular, then a nonzero element in its nullspace will be an eigenvector of A with eigenvalue λ. It follows that the eigenvalues of A are precisely the scalars λ such that $A - \lambda I$ is singular, or equivalently $\det(A - \lambda I) = 0$. These are the roots of the characteristic polynomial $\det(A - xI)$. We have proved that the eigenvalues of A are the roots of the characteristic polynomial of A. Similarly, for an endomorphism ϕ of an n-dimensional vector space V, the eigenvalues of ϕ are the roots of the characteristic polynomial of ϕ.

In the next section we will want to focus on the case of matrices whose elementary divisors are powers of degree one primes. The next definitions are useful for this purpose.

Definition 14.6. Let $0 \neq f(x) \in F[x]$ for a field F with $d = \deg f$. We say that f *splits* over F if f factors as $f = c(x - r_1)(x - r_2)\dots(x - r_d)$ in $F[x]$ (i.e., with $c, r_1, \dots, r_d \in F$).

If a polynomial f of degree d splits over F, then it has d roots $r_1, \dots, r_d$ in F (counted with multiplicity).

Definition 14.7. A field F is *algebraically closed* if every nonzero polynomial $f \in F[x]$ splits over F.

The *Fundamental Theorem of Algebra* is the statement that the field $\mathbb{C}$ of complex numbers is algebraically closed. We will give a proof of this at the end of our study of field theory, but we simply assume it now for convenience. We will also prove later that any field is contained in an algebraically closed one.

Note that if a field F is algebraically closed, then every nonzero polynomial in $F[x]$ factors as a product of degree 1 factors in $F[x]$ and a unit. This implies that the irreducible polynomials in $F[x]$ are precisely the polynomials of degree 1. Recall that a polynomial is *monic* if its leading coefficient is 1. The monic irreducibles in $F[x]$ are just the polynomals $(x - r)$ with $r \in F$.

14.2. **Jordan canonical form.** Let F be a field. Consider a finite dimensional vector space V over F, and an F-linear transformation $\phi \in \mathrm{End}_F(V)$. Recall that given any vector space with a choice of linear endomorphism, we can encode this information by making V into a module over the ring $F[x]$, where the constant polynomials in F act by the existing scalar multiplication and x acts by $x \cdot v = \phi(v)$. The action by a general element of $F[x]$ then follows the rule $(\sum_{i=0}^{n} a_i x^i) \cdot v = \sum_{i=0}^{n} a_i \phi^n(v)$, where we take $\phi^0 = 1_V$.

Now since V is finite-dimensional over F, it is finitely generated over F (by a basis) and so it is certainly finitely generated as a module over the larger ring $F[x]$. Since $F[x]$ is a PID, we can apply the classification of finitely generated modules over a PID to the module V. Here we apply the elementary divisor version; in the next section we show how to get somewhat different information by applying the invariant factor version.

Note that a nonzero free $F[x]$-module has infinite dimension as an F-vector space. Thus applying the classification to our module V, we see that the free part of V is zero and V is a torsion $F[x]$-module. The theorem tells us that there is an $F[x]$-module isomorphism

$$V \cong F[x]/(f_1^{e_1}) \oplus F[x]/(f_2^{e_2}) \oplus \cdots \oplus F[x]/(f_s^{e_s})$$

where the $f_1, \dots, f_s$ are prime, i.e. irreducible, polynomials in $F[x]$ and the $e_i \geq 1$. The prime powers $f_i^{e_i}$ are unique up to the order in which they appear, and possibly replacing f_i with associates. In this case by multiplying each by a nonzero scalar we can insist that the f_i be monic and then there is no ambiguity up to associates.

The Jordan canonical form we want to develop exists only under an additional condition: we assume now that the all of the irreducible polynomials f_i appearing in the elementary divisors have degree 1. By the comments in the previous section, this is always the case if we assume that F is algebraically closed. When we study fields we will see that every field is contained in an algebraically closed one, so this is not a huge restriction.

With our new assumption, we have that we can write $f_i = (x - \lambda_i)$ for some $\lambda_i \in F$. So as $F[x]$-modules we have an isomorphism

$$V \cong F[x]/((x - \lambda_1)^{e_1}) \oplus F[x]/((x - \lambda_2)^{e_2}) \oplus \cdots \oplus F[x]/((x - \lambda_s)^{e_s}).$$

The λ_i are not necessarily distinct.

For the moment, consider the case where $s = 1$, that is where V has only one elementary divisor. For convenience, drop the indexing and write $V \cong F[x]/(x-\lambda)^e$ as $F[x]$-modules. Now we choose a F-basis of $F[x]/(x-\lambda)^e$ for which the multiplication by x map will have a simple form. Let $I = ((x-\lambda)^e)$ be the ideal of $R = F[x]$ generated by $(x-\lambda)^e$, so $V \cong R/I$ as R-modules.

Now notice that $\{w_1 = (x-\lambda)^{e-1} + I, w_2 = (x-\lambda)^{e-2} + I, \dots, w_{e-1} = (x-\lambda) + I, w_e = 1 + I\}$ is an F-basis of R/I. This follows just since I is generated by a polynomial of degree e, and $(x-\lambda)^i$ has degree i, so we have coset representatives of degrees $1, 2, \dots, e-1$.

In the $R = F[x]$-module action on R/I, we have

$$(x-\lambda)\cdot w_i = (x-\lambda)[(x-\lambda)^{e-i} + I] = (x-\lambda)^{e-i+1} + I = \begin{cases} w_{i-1} & 2 \le i \le e \\ 0 & i = 1. \end{cases}$$

We can rewrite $(x-\lambda)\cdot w_i = w_{i-1}$ as $x \cdot w_i = \lambda w_i + w_{i-1}$ for $2 \le i \le e$, while $(x-\lambda)\cdot w_1 = 0$ becomes $x \cdot w_1 = \lambda w_1$. In other words, w_1 is an eigenvector for the linear transformation of $F[x]/I$ given by action by x.

Now let $\theta : V \to F[x]/I$ be a $F[x]$-module isomorphism. Then define $v_i = \theta^{-1}(w_i)$ for all i. Since θ is an $F[x]$-module isomorphism, it is also an F-vector space isomorphism. Thus $\mathcal{B} = \{v_1, \dots, v_e\}$ is an F-basis of V. Moreover, we have $\theta(x \cdot v_i) = x \cdot \theta(v_i) = x \cdot w_i$ for all i. But by definition $x \cdot v_i = \phi(v_i)$ for all i. Given the rules above for how x acts on the basis $\{w_i\}$ of R/I, we have

$$\phi(v_i) = \begin{cases} \lambda v_i + v_{i-1} & 2 \le i \le e \\ \lambda v_1 & i = 1. \end{cases}$$

This shows that the matrix of ϕ with respect to the basis $\mathcal{B}$ has an especially simple form:

$$M_{\mathcal{B}}^{\mathcal{B}}(\phi) = \begin{pmatrix} \lambda & 1 & & & & \mathbf{0} \\ & \lambda & 1 & & & \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ & & & & \lambda & 1 \\ \mathbf{0} & & & & & \lambda \end{pmatrix}$$

More precisely,

$$[M^{\mathcal{B}}_{\mathcal{B}}(\phi)]_{ij} = \begin{cases} \lambda & i = j \\ 1 & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

This matrix is called the *Jordan block* of size e associated to the eigenvalue λ. Its only nonzero entries are along the main diagonal (all λs) and the diagonal just above it (all 1's). We also write this $e \times e$ matrix as $J_{\lambda,e}$.

Now we pass to the general case where there is more than one elementary divisor and an isomorphism

$$\theta : V \cong F[x]/((x-\lambda_1)^{e_1}) \oplus F[x]/((x-\lambda_2)^{e_2}) \oplus \cdots \oplus F[x]/((x-\lambda_s)^{e_s}).$$

We can choose a special basis $\{w_{i,1}, \dots, w_{i,e_i}\}$ of each summand $F[x]/((x-\lambda_i)^{e_i})$ of the right hand side, as above, where $w_{i,j} = (x-\lambda_i)^{e_i - j} + ((x-\lambda_i)^{e_i})$. Then stringing these together gives as a basis $\{w_{1,1}, \dots, w_{1,e_1}, w_{2,1}, \dots, w_{2,e_2}, \dots, w_{s,1}, \dots w_{s,e_s}\}$ of the right hand side. Applying θ^{-1} gives us a basis $\mathcal{B}$ of V. The matrix of ϕ with respect to $\mathcal{B}$ is then a block matrix

$$M^{\mathcal{B}}_{\mathcal{B}}(\phi) = \left(\begin{array}{c|c|c|c} J_{\lambda_1,e_1} & & & \mathbf{0} \\ \hline & J_{\lambda_2,e_2} & & \\ \hline & & \ddots & \\ & & \qquad \ddots & \\ \hline \mathbf{0} & & & J_{\lambda_s,e_s} \end{array}\right)$$

Here, this is a matrix of size $n \times n$ where $n = e_1 + \cdots + e_s$. The diagonal blocks are Jordan blocks, and all other blocks are 0. A matrix of this type is said to be in *Jordan canonical form.*

Theorem 14.8. *Let $\phi : V \to V$ be a linear transformation of the n-dimensional vector space V over F. Make V into an $F[x]$-module via ϕ, and assume the that elementary divisors of the $F[x]$-module V are of the form $(x-\lambda_1)^{e_1}, \dots, (x-\lambda_s)^{e_s}$. Then there is a basis $\mathcal{B}$ of V s.t. $M^{\mathcal{B}}_{\mathcal{B}}(\phi)$ is in Jordan canonical form, with Jordan blocks $J_{\lambda_1,e_1}, \dots, J_{\lambda_s,e_s}$.*

If there is another basis $\mathcal{B}'$ such that $M^{\mathcal{B}'}_{\mathcal{B}'}(\phi)$ is also in Jordan canonical form, then this matrix is the same as $M^{\mathcal{B}}_{\mathcal{B}}(\phi)$ after possibly reordering the Jordan blocks.

Proof. The existence of the basis $\mathcal{B}$ was proved in the preceding discussion.

Conversely, suppose that $\mathcal{B}'$ is another basis for which $M^{\mathcal{B}'}_{\mathcal{B}'}(\phi)$ is in Jordan form. The fact that this is a block matrix with only the blocks along the main diagonal nonzero, say with blocks of size $f_1, f_2, \dots, f_t$, means that $V \cong V_1 \oplus V_2 \oplus \cdots \oplus V_t$ with $\phi(V_i) \subseteq V_i$, where V_i is spanned by f_i elements of the basis, and with $f_1 + \cdots + f_t = n$. Since the matrix of $\phi|_{V_i}$ with respect to the corresponding f_i basis elements is a Jordan block J_{μ_i, f_i}, as an $F[x]$-module x acts on the basis $v_1, \dots, v_{f_i}$ of V_i via the rules

$$\phi(v_i) = \begin{cases} \mu_i v_i + v_{i-1} & 2 \leq i \leq f_i \\ \mu_i v_1 & i = 1. \end{cases}$$

This easily implies that $V_i \cong F[x]/(x - \mu_i)^{f_i}$ as an $F[x]$-module, by reversing the steps in the earlier argument. We conclude that

$$V \cong F[x]/((x - \mu_1)^{f_1}) \oplus F[x]/((x - \mu_2)^{f_2}) \oplus \cdots \oplus F[x]/((x - \mu_t)^{f_t}).$$

as $F[x]$-modules. But now $(x - \mu_1)^{f_1}, \dots, (x - \mu_t)^{f_t}$ are elementary divisors for the module V. By the uniqueness of elementary divisors, $s = t$, and possibly after renumbering, we have $\mu_i = \lambda_i$ and $e_i = f_i$ for all i. In other words, the Jordan form $M_{\mathcal{B}'}^{\mathcal{B}'}(\phi)$ is the same as $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ after reordering the Jordan blocks. □

Corollary 14.9. *Let F be an algebraically closed field. Then every matrix $A \in M_n(F)$ is similar to a matrix in Jordan canonical form. The Jordan form is uniquely determined up to rearrangement of the Jordan blocks.*

One reason that the Jordan form is useful is that calculation of the powers of a matrix in this form is especially simple.

Example 14.10. Let $J = J_{\lambda,3}$ be a Jordan block of size 3. Then for all $n \geq 1$ we have

$$J^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1} & \binom{n}{2}\lambda^{n-2} \\ 0 & \lambda^n & n\lambda^{n-1} \\ 0 & 0 & \lambda^n \end{pmatrix},$$

by an easy inductive proof.

Similarly, a Jordan block of any size has an explicit formula for its powers involving binomial coefficients. Then the powers of any Jordan form may also be explicitly determined, simply by taking powers of the blocks. Finally, if A is an arbitrary matrix which is similar to a Jordan form J, if we calculate explicitly the matrix P such that $A = P^{-1}JP$, then the powers of A may be explicitly determined as $A^n = P^{-1}J^nP$.

If we are trying to understand all elements of $M_n(F)$ with a certain property that is invariant under similarity, then if F is algebraically closed we can reduce to the case of a Jordan form, where the calculation is usually much easier.

Example 14.11. Suppose we would like to find all elements of $\mathrm{GL}_2(\mathbb{C})$ which have order dividing 3 in this group. In other words we want all matrices A such that $A^3 = I$.

Since we are working over the algebraically closed field $\mathbb{C}$, all matrices have a Jordan form. Note that $A^3 = I$ if and only if $B^3 = I$, for any matrix B similar to A. Thus if we find all nonsingular matrices A in Jordan form which have $A^3 = I$, then the answer will simply be the union of all similarity classes containing those Jordan forms.

Jordan forms of 2×2 matrices are either diagonal, say $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ for $\lambda_1, \lambda_2 \in \mathbb{C}$, or else a Jordan block of size 2, $\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$. The formula for powers of a 2×2 Jordan block, which is even easier than Example 14.10, is

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1} \\ 0 & \lambda^n \end{pmatrix}.$$

In particular no positive power of such a Jordan block can be the identity matrix I. We conclude that the Jordan form of an invertible matrix of order dividing 3 is a diagonal matrix. Since $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^3 = \begin{pmatrix} \lambda_1^3 & 0 \\ 0 & \lambda_2^3 \end{pmatrix}$, The Jordan forms with multiplicative order dividing 3 are

$$S = \left\{ \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \middle| \lambda_1^3 = \lambda_2^3 = 1 \right\}.$$

$\mathbb{C}$ has three cube roots of 1, $\{1, e^{2\pi i/3}, e^{4\pi i/3}\}$. There are thus 9 distinct diagonal matrices in S, and the set of matrices with multiplicative order dividing 3 is equal to the set of all conjugates of S, i.e. the union of the similarity classes containing elements of S.

Suppose we want to know how many distinct similarity classes there are of such matrices. Then we need to determine whether any of the elements in S are similar. We know the Jordan form is determined only up to rearranging the Jordan blocks; thus two diagonal matrices with the same diagonal elements in some order are similar. It is now easy to see that there are only 6 distinct similarity classes containing the invertible matrices A such that $A^3 = I$.

14.3. **Rational canonical form.** The rational canonical form is developed in a very similar way to the Jordan canonical form, except using the invariant factor form instead of the elementary divisor form of the fundamental theorem.

Again let F be any field, let V be a finite dimensional F-space, and choose an F-linear transformation $\phi \in \mathrm{End}_F(V)$. Make V into an $F[x]$-module where x acts by ϕ. Then V is a torsion

$F[x]$-module, as we have already argued earlier. Theorem 13.17 tells us that there is an $F[x]$-module isomorphism

$$V \cong F[x]/(f_1) \oplus F[x]/(f_2) \oplus \cdots \oplus F[x]/(f_m)$$

where the invariant factors $f_1, \ldots, f_m \in F[x]$ satisfy $f_i | f_{i+1}$ for all $1 \leq i \leq m-1$. The invariant factors are uniquely determined up to associates. We will insist that the f_i are monic polynomials, and then they are completely unique.

Consider first the case where $n = 1$, so there is one invariant factor. Then $V \cong F[x]/(f)$ for some monic polyomial $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 \in F[x]$, some $n \geq 1$. In other words, V is essentially an arbitrary nonzero torsion cyclic $F[x]$-module.

Let $I = (f)$ for convenience of notation. Now $\{w_1 = 1 + I, w_2 = x + I, \ldots, w_n = x^{n-1} + I\}$ is an F-basis of $F[x]/I$. In the $F[x]$-module action on R/I, we have

$$x \cdot w_i = x(x^{i-1} + I) = x^i + I = \begin{cases} w_{i+1} & 1 \leq i \leq n-1 \\ -b_{n-1}w_n - b_{n-2}w_{n-1} - \cdots - b_1 w_2 - b_0 w_1 & i = n; \end{cases}$$

the second case follows since $f \in I$ and so

$$x^n + I = -(b_{n-1}x^{n-1} + \cdots + b_1 x + b_0) + I = -b_{n-1}(x^{n-1} + I) - \cdots - b_1(x + I) - b_0(1 + I).$$

Now if we fix an $F[x]$-module isomorphism $\theta : V \to F[x]/I$ and define $v_i = \theta^{-1}(w_i)$, then $\mathcal{B} = \{v_1, \ldots, v_n\}$ is an F-basis of V for which

$$\phi(v_i) = \begin{cases} v_{i+1} & 1 \leq i \leq n-1 \\ -b_{n-1}v_n - b_{n-2}v_{n-1} - \cdots - b_1 v_2 - b_0 v_1 & i = n. \end{cases}$$

We conclude that

$$M_{\mathcal{B}}^{\mathcal{B}}(\phi) = \begin{pmatrix} 0 & 0 & & & & & -b_0 \\ 1 & 0 & & & & & -b_1 \\ 0 & 1 & & & & & -b_2 \\ & & \ddots & & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & 1 & 0 & -b_{n-2} \\ & & & & & 1 & -b_{n-1} \end{pmatrix}.$$

More precisely,

$$[M^{\mathcal{B}}_{\mathcal{B}}(\phi)]_{ij} = \begin{cases} 1 & i = j+1 \\ -b_{i-1} & j = n \\ 0 & \text{otherwise.} \end{cases}$$

This matrix is called the *companion matrix* of the monic polynomial $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$. We also write it as C_f.

In the general case where we have more than one invariant factor, we proceed exactly as we did with the Jordan form. Fixing an isomorphism

$$\theta : V \cong F[x]/(f_1) \oplus F[x]/(f_2) \oplus \cdots \oplus F[x]/(f_m)$$

we choose a basis $\mathcal{B}$ of V which is the preimage under θ of the basis of the right hand side obtained by stringing together the special bases of the modules $F[x]/(f_i)$ we picked above. The matrix of ϕ with respect to $\mathcal{B}$ is then a block matrix

$$M^{\mathcal{B}}_{\mathcal{B}}(\phi) = \left(\begin{array}{c|c|cc|c} C_{f_1} & & \cdots & \cdots & \mathbf{0} \\ \hline & C_{f_2} & & & \\ \hline \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \hline \mathbf{0} & & & & C_{f_m} \end{array}\right)$$

A matrix of this type is said to be in *rational canonical form.* Note that there is no ambiguity in the order of the blocks of a matrix in rational canonical form. The blocks must be the companion matrices of polynomials each of which divides the next as we go down the diagonal from the top left to the bottom right.

Theorem 14.12. *Let $\phi : V \to V$ be a linear transformation of the n-dimensional vector space V over F. Make V into an $F[x]$-module via ϕ, and assume that the invariant factors of the $F[x]$-module V are $f_1, \ldots, f_m$. Then there is a basis $\mathcal{B}$ of V such that $M^{\mathcal{B}}_{\mathcal{B}}(\phi)$ is in rational canonical form, with blocks $C_{f_1}, \ldots, C_{f_m}$.*

If $\mathcal{B}'$ is a basis such that $M^{\mathcal{B}'}_{\mathcal{B}'}(\phi)$ is also in rational canonical form, then $M^{\mathcal{B}'}_{\mathcal{B}'}(\phi) = M^{\mathcal{B}}_{\mathcal{B}}(\phi)$.

Proof. The proof is completely analogous to the proof of Theorem 14.8, but using the uniqueness of the invariant factor decomposition of a torsion module instead. We leave the details to the reader. □

Corollary 14.13. *Let F be an arbitrary field. Then every matrix $A \in M_n(F)$ is similar to a unique matrix in $M_n(F)$ which is in rational canonical form.*

We have emphasized the Jordan canonical form since this is the form which is most useful in calculations and in applications. There is no nice formula for the powers of a companion matrix, in contrast. The rational canonical form is useful for theoretical reasons, however, because it is defined over an arbitrary field, and it is absolutely unique. The Jordan canonical form, by contrast, only exists over certain fields, and is unique only up to permutation of the blocks.

Here is a useful theorem whose proof becomes nearly trivial with the use of the rational canonical form. Recall that $F \subseteq K$ is called a *field extension* if F and K are fields, and F is a subring of K.

Theorem 14.14. *Let $F \subseteq K$ be a field extension. Let $A, B \in M_n(F)$. Then we can also consider $A, B \in M_n(K)$. If A and B are similar in the ring $M_n(K)$ (i.e. $A = P^{-1}BP$ for some $P \in$ $\mathrm{GL}_n(K)$) then A and B are similar in $M_n(F)$.*

Proof. Let C be the rational canonical form of A over F, and let C' be the rational canonical form of B over F. Thus $C, C' \in M_n(F)$.

Now A and C are similar in $M_n(F)$, so A and C are certainly similar in $M_n(K)$ as well. But the matrix C is in rational canonical form, that is it is block diagonal with companion matrices C_{f_i} along the diagonal, for $f_i \in F[x]$ with $f_i | f_{i+1}$ for all i. Clearly then C is also in rational canonical form when considered as a matrix in $M_n(K)$. By the uniqueness of the rational canonical form, C is the rational canonical form of A in $M_n(K)$. We have in the same way that C' is the rational canonical form of B in $M_n(K)$. But by assumption A and B are similar in $M_n(K)$. Thus $C = C'$ by the uniqueness of the rational canonical form in $M_n(K)$. But then A is similar to B in $M_n(F)$ by the uniqueness of the rational canonical form in $M_n(F)$. □

The preceding result is highly non-obvious without introducing forms. If $A = P^{-1}BP$ for some $P \in \mathrm{GL}_n(K)$, there is no obvious way to adjust P to obtain a $Q \in \mathrm{GL}_n(F)$ such that $A = Q^{-1}BQ$ also.

14.4. **Characteristic and minimal polynomials.** We now relate canonical forms of a matrix to its characteristic and minimal polynomials (we will define the minimal polynomial shortly). Because the rational canonical form is defined over any field we use this form in our initial approach.

Lemma 14.15. *Let $C_f \in M_n(F)$ be a companion matrix for a monic polynomial $f \in F[x]$. Then* $\mathrm{charpoly}(C_f) = f[x]$.

Proof. Let $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$. We have

$$\operatorname{charpoly}(C_f) = \det \begin{pmatrix} x & 0 & & & & & b_0 \\ -1 & x & & & & & b_1 \\ 0 & -1 & x & & & & b_2 \\ & & \ddots & & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & -1 & x & b_{n-2} \\ & & & & & -1 & x + b_{n-1} \end{pmatrix}.$$

Expanding by minors along the first row gives

$$x \det \begin{pmatrix} x & 0 & & & & & b_1 \\ -1 & x & & & & & b_2 \\ 0 & -1 & x & & & & b_3 \\ & & \ddots & & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & -1 & x & b_{n-2} \\ & & & & & -1 & x + b_{n-1} \end{pmatrix} + (-1)^{n-1} b_0 \det \begin{pmatrix} -1 & x & & & & \\ 0 & -1 & x & & & \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & -1 & x \\ & & & & & -1 \end{pmatrix}$$

$$= x \det(C_g) + (-1)^{n-1}(-1)^{n-1} b_0 = x \det(C_g) + b_0,$$

where $g(x) = x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1$.

By induction on the size of the companion matrix, we have $\det(C_g) = g$, and so we get $\det(C_f) = xg + b_0 = f$. □

Corollary 14.16. *Let $\phi : V \to V$ be a F-linear transformation of a finite dimensional F-space V. Suppose the corresponding $F[x]$-module structure on V has invariant factors $f_1, \ldots, f_n$. Then* $\operatorname{charpoly}(\phi) = f_1 f_2 \ldots f_n$.

Proof. $\operatorname{charpoly}(\phi)$ is the same as $\operatorname{charpoly}(C)$ for a rational canonical form C of ϕ. The result follows from Lemma 14.15 and the fact that the determinant of a block diagonal matrix is the product of the determinants of the blocks. □

Let us also now discuss the minimal polynomial of a matrix or a linear transformation.

Definition 14.17. Let $f = \sum_{i=0}^{n} a_i x^i \in F[x]$. Given a matrix $A \in M_n(F)$ we define the evaluation of f at A to be $f(A) = \sum_{i=0}^{n} a_i A^i \in M_n(F)$ (where $A^0 = I$).

Note that for a fixed matrix A, the "evaulation at A" map $\epsilon_A : F[x] \to M_n(F)$ given by $\epsilon_A(f) = f(A)$ is a ring homomorphism. Thus $\ker \epsilon_A$ is an ideal of $F[x]$ and $\ker \epsilon_A = (g)$ for some $g \in F[x]$ since $F[x]$ is a PID. The map ϵ_A is also a linear transformation over F; since $\dim_F F[x] = \infty$ while $\dim_F M_n(F) = n^2$, we have $\ker \epsilon_A \neq 0$ and so $g \neq 0$.

Definition 14.18. The unique monic polynomial $g \in F[x]$ such that $\ker \epsilon_A = (g)$ is called the *minimal polynomial* of A and denoted $\operatorname{minpoly}(A)$.

Recall that a monic generator of a nonzero ideal I of $F[x]$ is the monic polynomial of uniquely smallest degree among nonzero elements of I. Thus $\operatorname{minpoly}(A)$ is the monic polynomial of smallest degree which when evaluated at A gives 0. This justifies the terminology.

Of course we can make this definition for linear transformations as well. If $\dim_F V < \infty$ and $\phi \in \operatorname{End}_F(V)$, then we can define an evaluation map $\epsilon_\phi : F[x] \to \operatorname{End}_F(V)$ by $\epsilon_\phi(\sum_{i=0}^n a_i x^i) = \sum_{i=0}^n a_i \phi^i$ (where $\phi^0 = 1_V$) and define $\operatorname{minpoly}(\phi)$ to be the unique monic generator of $\ker \epsilon_\phi$. It is easy to see that $\operatorname{minpoly}(\phi) = \operatorname{minpoly}(M_{\mathcal{B}}^{\mathcal{B}}(\phi))$ for any basis $\mathcal{B}$ of V.

Proposition 14.19. *Let $\phi : V \to V$ be a F-linear transformation of a finite dimensional F-vector space V. Let $f_1, f_2, \ldots, f_n$ be the invariant factors of V considered as an $F[x]$-module where x acts as ϕ. Then $f_n = \operatorname{minpoly}(\phi)$.*

Proof. For any commutative ring R and ideal I, $\operatorname{ann}_R(R/I) = I$. Thus $\operatorname{ann}_{F[x]} F[x]/(f_i) = (f_i)$. Also, it is clear that $\operatorname{ann}_R(M_1 \oplus \cdots \oplus M_n) = \bigcap_{i=1}^n \operatorname{ann}_R(M_i)$ for any direct sum of R-modules M_i. Since we have $V \cong F[x]/(f_1) \oplus \cdots \oplus F[x]/(f_n)$, we conclude that $\operatorname{ann}_{F[x]} V = \bigcap_{i=1}^n (f_i) = (f_n)$ since $f_i | f_n$ for $1 \leq i \leq n$.

We also claim that for $h \in F[x]$, $h \in \operatorname{ann}_{F[x]} V$ if and only if $h(\phi) = 0$. Writing $h = \sum_{i=0}^n a_i x^n$, we have

$$\begin{aligned}
&h \cdot v = 0 \text{ for all } v \in V \\
\iff &(\sum_{i=0}^n a_i x^n) \cdot v = 0 \text{ for all } v \in V \\
\iff &\sum_{i=0}^n a_i \phi^n(v) = 0 \text{ for all } v \in V \\
\iff &[h(\phi)](v) = 0 \text{ for all } v \in V \\
\iff &h(\phi) = 0
\end{aligned}$$

as claimed.

Thus $\operatorname{ann}_{F[x]} V = \ker \epsilon_\phi$ for the evaluation map $\epsilon_\phi : F[x] \to \operatorname{End}_F(V)$ and we conclude that $\ker \epsilon_\phi = (f_n)$. By definition, $f_n = \operatorname{minpoly}(\phi)$. □

Proposition 14.20. *Let $\phi \in \operatorname{End}_F(V)$, where V is a finite dimensional F-vector space.*

(1) $\operatorname{minpoly}(\phi) | \operatorname{charpoly}(\phi)$.

(2) *If $p(x)$ is irreducible in $F[x]$ and $p | \operatorname{charpoly}(\phi)$, then $p | \operatorname{minpoly}(\phi)$.*

Proof. (1) Let V be an $F[x]$-module where x acts as ϕ, and let $f_1, \dots, f_n$ be the invariant factors of this module. We have seen that $\operatorname{minpoly}(\phi) = f_n$ is the largest invariant factor by Proposition 14.19, and $\operatorname{charpoly}(\phi) = f_1 f_2 \dots f_n$ is the product of the invariant factors by Corollary 14.16. The result follows.

(2) If p is irreducible and $p | f_1 f_2 \dots f_n$, then $p | f_i$ for some i since p is prime. Since $f_i | f_n$ for all i we get that $p | f_n$. □

An immediate consequence of the results above is the following pretty result known as the Cayley-Hamilton Theorem. There are various tricky proofs of this result that may be considered more elementary since they do not rely on the theory of forms, but it is striking how the result just falls out from the simple properties of the rational canonical form we have developed.

Theorem 14.21. *Let $A \in M_n(F)$ for a field F. If $f = \operatorname{charpoly}(A)$, then $f(A) = 0$. In other words, any matrix satisfies its own characteristic polynomial.*

Proof. We have seen in Proposition 14.20 that $g = \operatorname{minpoly}(A)$ divides $f = \operatorname{charpoly}(A)$ in $F[x]$; say $f = gh$. But $g(A) = 0$ by definition. So $f(A) = g(A)h(A) = 0$ as well. □

Since the elementary divisors are related to the invariant factors in a simple way, we can also relate the characteristic polynomial and minimal polynomial to these.

Lemma 14.22. *Let $\phi : V \to V$ be a F-linear transformation of a finite dimensional F-vector space V. Consider V as an $F[x]$-module where x acts as ϕ. We can write the elementary divisors of V in the form $\{p_i^{e_{i,j}} | 1 \leq i \leq m, 1 \leq j \leq s_i\}$ where the p_i are monic and pairwise distinct primes in $F[x]$, and where $e_{i,1} \geq e_{i,2} \geq \dots \geq e_{i,s_i}$ for each i. Then $\operatorname{charpoly}(\phi)$ is the product of all of the elementary divisors, and $\operatorname{minpoly}(\phi) = p_1^{e_{1,1}} p_2^{e_{2,1}} \dots p_m^{e_{m,1}}$ is the product of the largest powers of the distinct primes occurring among the elementary divisors.*

Proof. This follows from now the invariant factors are related to elementary divisors, as in the proof of Theorem 13.17. In particular, since the product of the elementary divisors is the product of the invariant factors, $\operatorname{charpoly}(\phi)$ is the product of all elementary divisors. Since the largest invariant factor is the product of the largest powers of the distinct primes occurring among the elementary divisors, we get $\operatorname{minpoly}(\phi) = p_1^{e_{1,1}} p_2^{e_{2,1}} \dots p_m^{e_{m,1}}$. □

We can use this to give an easy to understand condition for when a particular linear transformation has a Jordan canonical form over a field F.

Corollary 14.23. *Let $\phi : V \to V$ be an F-linear transformation of a finite dimensional F-vector space V. Then ϕ has a Jordan canonical form in $M_n(F)$ if and only if* charpoly(ϕ) *splits over F.*

Proof. Consider the elementary divisors of V as an $F[x]$-module via ϕ. Then ϕ has a Jordan form if and only if those elementary divisors are all powers of degree 1 irreducibles $(x - \lambda)$ in $F[x]$. Since charpoly(ϕ) is the product of the elementary divisors, this is if and only if charpoly(ϕ) is a product of degree 1 irreducibles, i.e. it splits over F. □

For a small matrix, often knowledge of the characteristic and minimal polynomials is enough to determine what the Jordan and rational forms must be. We illustrate this in the following simple example.

Example 14.24. Consider $A = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

Since A is upper triangular, clearly charpoly$(A) = (x-2)^2(x-1)$. Recall from Proposition 14.20(b) that every prime dividing the characteristic polynomial divides the minimal polynomial, so minpoly(A) is either $(x-2)(x-1)$ or $(x-2)^2(x-1)$.

We can calculate

$$(A - 2I)(A - I) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \neq 0$$

and thus we must have minpoly$(A) = (x-2)^2(x-1) =$ charpoly(A). This implies that A has a single invariant factor $f_1 = (x-2)^2(x-1) = x^3 - 5x^2 + 8x - 4$. The rational canonical form of A is

$$\begin{pmatrix} 0 & 0 & 4 \\ 1 & 0 & -8 \\ 0 & 1 & 5 \end{pmatrix}.$$

Because charpoly(A) is a product of linear factors in $F[x]$, A also has a Jordan form over F, no matter what the field F is. The elementary divisors are $(x-2)^2$ and $(x-1)$ (note that in any field F, $1 \neq 2$).

So the Jordan form is

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

14.5. **Generalized eigenspaces and the Jordan form.** For convenience, let us assume now that the field F is algebraically closed, so that every matrix in $M_n(F)$ has a Jordan canonical form. In applications of the Jordan form it is useful to relate it to the theory of generalized eigenvalues and eigenvectors.

Let V be a vector space with $\dim_F V = n$ and let $\phi \in \operatorname{End}_F(V)$.

Definition 14.25. Given $\lambda \in F$, we say that $v \in V$ is a *generalized eigenvector* with eigenvalue λ if $(\phi - \lambda 1_V)^n(v) = 0$ for some $n \geq 1$.

Thinking of V as an $F[x]$-module where x acts as ϕ, it is equivalent to define a generalized eigenvector to be $v \in V$ such that $(x-\lambda)^n \cdot v = 0$. It is easy to see that

$$V_\lambda = \{v \in V | v \text{ is a generalized eigenvector with eigenvalue } \lambda\}$$

is an $F[x]$-submodule of V; in other words it is an F-subspace such that $\phi(V_\lambda) = V_\lambda$. V_λ is called a *generalized eigenspace*. If $(x-\lambda)^n \cdot v = 0$ and n is the minimal exponent for which this holds, then $0 \neq w = (x-\lambda)^{n-1} \cdot v$ and $(x-\lambda) \cdot w = 0$, so w is a genuine eigenvector for ϕ with eigenvalue λ. Thus the λ such that the generalized eigenspace V_λ is nonzero are exactly the eigenvalues of ϕ.

Since F is algebraically closed, the monic irreducible polynomials in $F[x]$ are just the polynomials $(x-\lambda)$ for $\lambda \in F$. We see that by definition V_λ is precisely the $(x-\lambda)$-primary component of V as defined earlier. By Proposition 13.13, $V \cong V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_m}$ for some distinct λ_i, as $F[x]$-modules. In other words, V is the direct sum of finitely many generalized eigenspaces for ϕ.

Moreover, by Proposition 13.16, each primary component V_λ is a direct sum of cyclic modules, say $V_\lambda \cong F[x]/(x-\lambda)^{e_1} \cdots \oplus \ldots F[x]/(x-\lambda)^{e_s}$. In other words, $(x-\lambda)^{e_1}, \ldots, (x-\lambda)^{e_s}$ are those elementary divisors of ϕ which are powers of the prime $(x-\lambda)$. By the proof of Theorem 13.6, we can determine the list of positive integers e_i as follows: the number of $e_i \geq b$ is the dimension over F of $V_\lambda[b]/V_\lambda[b-1]$, where $V_\lambda[b] = \{v \in V_\lambda | (x-\lambda)^b \cdot v = 0\}$. Thus it suffices to find $\dim_F V_\lambda[b]$ for each b.

Note that $V_\lambda[1]$ is precisely the space of eigenvectors for ϕ with eigenvalue λ, and $\dim_F V_\lambda[1]$ is equal to the number of the e_i, which is all of them. $\dim_F V_\lambda[1]$ is the same as the number of elementary divisors which are powers of $(x-\lambda)$. Thus the number of independent eigenvectors with eigenvalue λ is the same as the number of Jordan blocks of the Jordan form which are associated to the eigenvalue λ.

This also makes sense because a Jordan block only has a 1-dimensional space of eigenvectors; it is useful for intuition to see what is going on directly in that case:

Example 14.26. Suppose that $\phi : V \to V$ has a basis $\mathcal{B} = \{v_1, \dots, v_n\}$ with respect to which $M_{\mathcal{B}}^{\mathcal{B}}(\phi)$ is a single Jordan block $J_{\lambda,e}$. Then $\phi(v_i) = \lambda v_i + v_{i-1}$ for $i \geq 1$, and $\phi(v_1) = \lambda v_1$, as we have seen. By definition $(\phi - \lambda 1_V)(v_i) = v_{i-1}$ for $i \geq 1$, and $(\phi - \lambda 1_V)(v_1) = 0$. From this one sees that $(\phi - \lambda 1_V)^i(v_i) = 0$ while $(\phi - \lambda 1_V)^{i-1}(v_i) = v_1 \neq 0$, for $1 \leq i \leq n$. It easily follows that $v_i + V_\lambda[i-1]$ is a basis of $V_\lambda[i]/V_\lambda[i-1]$ and so all of these factor spaces are 1-dimensional, for $1 \leq i \leq n$. This is consistent with the fact that the corresponding $F[x]$-module structure on V is isomorphic to $F[x]/(x-\lambda)^n$ and there is only one elementary divisor $(x-\lambda)^n$. The eigenspace of ϕ is 1-dimensional, so ϕ is highly defective in the sense that V is far from being spanned by eigenvectors (but every element of V is a generalized eignevector for the eigenvalue λ).

We can use the generalized eigenvector point of view to help determine the Jordan form of a matrix. We give a simple example next.

Example 14.27. Let $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & -2 & 0 & 1 \\ -2 & 0 & -1 & 2 \end{pmatrix}$.

We would like to find the Jordan form of A over $\mathbb{C}$. To put this explicitly in the context of the previous discussion, we can let $V = \mathbb{C}^4$ be a space column vectors and make it an $F[x]$-module where x acts by the linear transformation $\phi : V \to V$ determined by left multiplication by A.

The first step is to calculate the characteristic polynomial of A:

$$\det \begin{pmatrix} x-1 & 0 & 0 & 0 \\ 0 & x-1 & 0 & 0 \\ 2 & 2 & x & -1 \\ 2 & 0 & 1 & x-2 \end{pmatrix} = (x-1)^2 \det \begin{pmatrix} x & -1 \\ 1 & x-2 \end{pmatrix} = (x-1)^2(x^2-2x-1) = (x-1)^4.$$

We see that A has only one eigenvalue, $\lambda = 1$. Thus the elementary divisors of A must all be of the form $(x-1)^i$. Moreover, by Lemma 14.22, the product of the elementary divisors is $(x-1)^4$. Let the elementary divisors be $(x-1)^{e_1}, (x-1)^{e_2}, (x-1)^{e_3}, \dots$ (we don't know how many there are yet, though there are at most 4).

We first calculate the dimension of the 1-eigenspace of A. Since

$$(A-I)=\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -2 & -2 & -1 & 1 \\ -2 & 0 & -1 & 1 \end{pmatrix}$$

is clearly a matrix of rank 2, its nullspace has dimension 2. This means we have two linearly independent eigenvectors and the Jordan form has two Jordan blocks.

We still don't know if those blocks have size 1 and 3 or size 2 and 2. For this we can easily calculate that

$$(A-I)^2=\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix}.$$

This matrix as rank 1, so nullspace of dimension 3.

We have $V = V_1$, that is, V is the 1-generalized eigenspace. In the notation above, $\dim_F V_1[1] = 2$, and $\dim_F V_1[2] = 3$. This says that there are 2 e_i's with $e_i \geq 1$, and $\dim_F V_1[2]/V_1[1] = 3-2 = 1$ of the e_i's with $e_i \geq 2$. The only possibility is that $e_1 = 3$, $e_2 = 1$ and the elementary divisors are $(x-1)^3, (x-1)^1$. Thus the Jordan form of A is

$$J=\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is also easy to determine the rational canonical form. Clearly the invariant factors are also $f_1 = (x-1)$, $f_2 = (x-1)^3$. The minimal polynomial is thus $f_2 = (x-1)^3 = x^3 - 3x^2 + 3x - 1$. The rational form is

$$C=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & 3 \end{pmatrix}.$$

We did an example above with only one eigenvalue to demonstrate the method; in general, one would need to calculate the dimensions of the nullspaces of $(A-\lambda_i I)^j$ for each of the eigenvalues

λ_i and each $j \geq 1$ until one had enough information to determine the sizes of the Jordan blocks associated to λ_i.

Throughout this section, we have concentrated on the theoretical aspects of canonical forms rather than methods of calculation. The reader interested in computations can find more details in Chapter 12 of Dummit and Foote's book.

15. Tensor products

15.1. **Balanced maps.** Tensor products are very important gadgets, but they take a while to get comfortable with and see why they are so useful. It is natural the first time you learn about them to understand the basics of how to manipulate them, but feel like you still don't quite ”get” them. I think once they appear in your later classes and you see them in action they tend to sink in better.

For a little motivation we consider the case of vector spaces (i.e. modules over a field F) where tensor products are already interesting. Let V and W be F-vector spaces. As we know, we have the direct sum of vector spaces $V \oplus W$ which as a set is just the cartesian product $V \times W = \{(v, w) | v \in V, w \in W\}$. There may be natural functions from $V \times W$ to another vector space U which are not F-linear maps, but rather *F-linear in each coordinate separately.* Here is a very common example.

Example 15.1. Let V be a vector space over F and define $V^* = \mathrm{Hom}_F(V, F)$. Since F is a commutative ring, V^* is again an F-module, i.e. vector space, as we have seen. It is called the *dual* vector space, or the space of linear functionals on V. There is a very natural function $\theta : V^* \times V \to F$ defined by $\theta(\psi, v) = \psi(v)$. (Note that for notational convenience we write $\theta(\psi, v)$ instead of $\theta((\psi, v))$, which would be more technically correct.) This function satisfies $\theta(\psi+\phi, v) = [\psi+\phi](v) = \psi(v) + \phi(v) = \theta(\psi, v) + \theta(\phi, v)$ as well as $\theta(\psi, v+w) = \psi(v+w) = \psi(v) + \psi(w) = \theta(\psi, v) + \theta(\psi, w)$. Thus θ is linear in each coordinate and we say that θ is *bilinear.* However, θ is not an F-linear transformation from $V^* \oplus V$ to F. This would require $\theta((\psi, v) + (\phi, w)) = \theta(\psi, v) + \theta(\phi, w)$. But the left hand side of this equation is $\theta(\psi + \phi, v + w) = [\psi + \phi](v + w) = \psi(v) + \phi(v) + \psi(w) + \phi(w)$, while the right hand side is $\psi(v) + \phi(w)$. These are certainly not equal in general.

On the other hand, when doing linear algebra one would really like to work with linear maps. We are going to define a vector space $V^* \otimes_F V$, the tensor product of V^* and V over F, and an F-linear map $\widetilde{\theta} : V^* \otimes_F V \to F$ which contains the same information as the map θ. In some sense $V^* \otimes_F V$ is the vector space where bilinear functions like θ naturally live. We will see that to make this work $V^* \otimes_F V$ will have to be a bigger vector space than $V^* \oplus V$.

Let us begin the technical work to define tensor products. We are going to make the main definitions for modules over an arbitrary, possibly noncommutative ring R.

Definition 15.2. Let R be a ring. Let M be a right R-module and N a left R-module. Let P be any abelian group. A function $\phi : M \times N \to P$ is called *R-balanced* if

(1) $\phi(m_1 + m_2, n) = \phi(m_1, n) + \phi(m_2, n)$ for all $m_1, m_2 \in M$ and $n \in N$.

(2) $\phi(m, n_1 + n_2) = \phi(m, n_1) + \phi(m, n_2)$ for all $m \in M$ and $n_1, n_2 \in N$.

(3) $\phi(mr, n) = \phi(m, rn)$ for all $m \in M$, $r \in R$, and $n \in N$.

Note that the first two conditions say that an R-balanced map respects addition separately in each coordinate. It may seem awkward that this definition is made for one right module and one left module; this will also appear in the definition of tensor product to come shortly and it is essential to make the theory work properly. In the applications to modules over commutative rings R, of course, we can identify left and right modules and one can stick to modules over one side. We will make more comments about this later.

Example 15.3. Consider again the map $\theta : V^* \times V \to F$ defined by $\theta(\phi, v) = \phi(v)$ that was defined in Example 15.1. Since F is commutative, we can think of V^* as a right F-module just as well, where $\phi \cdot a = a\phi$ for $a \in F$, $\phi \in V^*$. With this convention we claim that the map θ is F-balanced. We have already seen that it respects sums separately in each coordinate. Moreover

$$\theta(\phi \cdot a, v) = [\phi \cdot a](v) = [a\phi](v) = a\phi(v) = \phi(av) = \theta(\phi, av),$$

for all $\phi \in V^*$, $a \in F$, and $v \in V$.

Example 15.4. Let M be a left R-module. Note that R is naturally a right R-module by right multiplication. Then the map $\psi : R \times M \to M$ defined by $\psi(r, m) = rm$ is R-balanced.

We now define a tensor product as a universal object for R-balanced maps from $M \times N$ to any abelian group.

Definition 15.5. Let M be a right R-module, and N a left R-module. A *tensor product* for M and N (over R) is an abelian group T together with an R-balanced map $\theta : M \times N \to T$ with the following universal property: for any abelian group P and R-balanced map $\phi : M \times N \to P$, there exists a unique homomorphism of abelian groups $\psi : T \to P$ such that $\theta\psi = \phi$.

Here is the commutative diagram which represents the universal propery of the tensor product:

$$\begin{tikzcd}
M \times N \arrow[r, "\theta"] \arrow[dr, "\phi"'] & T \arrow[d, dotted, "\exists ! \psi"] \\
 & P
\end{tikzcd}$$

The important thing is that ϕ is not a homomorphism of abelian groups (it does not respect addition), while ψ *is*. Moreover, since θ is a given part of the structure of the tensor product, the homomorphism ψ contains all of the information in the map ϕ (as we can recover ϕ by $\phi = \theta\psi$). So the tensor product allows one to replace balanced maps by actual homomorphisms, without losing information.

As a warmup, let us give an example of a tensor product in a special case where we can check directly that the definition holds.

Example 15.6. Let R be a right R-module by right multiplication, and let M be a left R-module. Then the R-balanced map $\theta : R \times M \to M$ from Example 15.4, where $\theta(r, m) = rm$, is a tensor product of R and M over R.

Proof. Suppose we have an abelian group P and an R-balanced map $\phi : R \times M \to P$. We need to find a group homomorphism $\psi : M \to P$ such that $\psi\theta = \phi$, and show that ψ is unique with this property.

Define $\psi : M \to P$ by $\psi(m) = \phi(1, m)$. Then since ϕ is R-balanced, for all $r \in R$ and $m \in M$ we have

$$\phi(r, m) = \phi(1r, m) = \phi(1, rm) = \psi(rm) = \psi\theta(r, m).$$

Thus $\psi\theta = \phi$. ψ is certainly uniquely determined, since if $\psi'\theta = \phi$ then $\psi'(m) = \psi'\theta(1, m) = \phi(1, m)$ and so $\psi' = \psi$.

Finally, ψ is a group homomorphism since $\psi(m_1 + m_2) = \phi(1, m_1 + m_2) = \phi(1, m_1) + \phi(1, m_2) = \psi(m_1) + \psi(m_2)$, using that ϕ is R-balanced. □

15.2. **Existence and uniqueness.** Let us first state the uniqueness of tensor products up to isomorphism. This follows by exactly the same argument as we have already seen for other universal properties, and so we leave the proof to the reader.

Proposition 15.7. *Let M be a right R-module and N a left R-module. Suppose that $\theta_1 : M \times N \to T_1$ and $\theta_2 : M \times N \to T_2$ are both tensor products of M and N over R. Then there is a unique isomorphism of abelian groups $\theta : \psi : T_1 \to T_2$ such that $\psi\theta_1 = \theta_2$.*

The proposition shows that there is essentially only one tensor product of M and N over R (if there is any). Now let us show that the tensor product always exists.

Theorem 15.8. *Let M be a right R-module and N a left R-module. Then there is an abelian group T and an R-balanced map $\theta : M \times N \to T$ which is a tensor product of M and N over R.*

Proof. Consider $S = M \times N = \{(m, n) | m \in M, n \in N\}$ as a set. Construct a free $\mathbb{Z}$-module (i.e. abelian group) indexed by S, in other words $F = \bigoplus_{s \in S} \mathbb{Z}$. We introduce a new formal symbol $m \otimes n$ to represent the standard basis element which is 1 in the (m, n)-spot and 0 elsewhere. Then a general element of F looks like $a_1(m_1 \otimes n_1) + \cdots + a_k(m_k \otimes n_k)$ for $a_i \in \mathbb{Z}$, $m_i \in M$, $n_i \in N$.

Now let $T = F/I$ where I is the subgroup of F generated by all elements of the form

$$\begin{aligned} (m_1 + m_2) \otimes n - m_1 \otimes n - m_2 \otimes n, &\qquad m_1, m_2 \in M,\ n \in N; \\ m \otimes (n_1 + n_2) - m \otimes n_1 - m \otimes n_2 &\qquad m \in M,\ n_1, n_2 \in N; \\ mr \otimes n - m \otimes rn &\qquad m \in M,\ r \in R,\ n \in N. \end{aligned}$$

We claim that $\theta : M \times N \to T = F/I$ defined by $\theta(m, n) = (m \otimes n) + I$ is a tensor product of M and N over R.

First we need that θ is R-balanced. This is immediate from the "relations" we threw into the subgroup I. For example, let us check the third condition of the R-balanced property:

$$\theta(mr, n) = (mr \otimes n) + I = (m \otimes rn) + I = \theta(m, rn) \text{ for all } m \in M, r \in R, n \in N,$$

where we have used that the cosets $(mr \otimes n) + I$ and $(m \otimes rn) + I$ are equal because $mr \otimes n - m \otimes rn \in I$ by definition. The other two conditions of the balanced property, that sums are respected in each coordinate, follow immediately in the same way.

Next, we need that if $\phi : M \times N \to P$ is R-balanced, for some abelian group P, there there is a unique linear map $\psi : T \to P$ such that $\phi = \psi\theta$. This also follows very formally from the fact that F is free, and that to produce T we have modded out the subgroup generated exactly those relations that represent being R-balanced.

More precisely, we first get that there is a unique homomorphism of $\mathbb{Z}$-modules (i.e. abelian groups) $\widehat{\psi} : F \to P$ such that $\widehat{\psi}(m \otimes n) = \phi(m, n)$ for all $m \in M$, $n \in N$. This is just because F is a free $\mathbb{Z}$-module on the basis $\{m \otimes n | m \in M, n \in N\}$. Second, because ϕ is R-balanced, it is easy to check that every element in the generating set of I must be in the kernel of $\widehat{\psi}$. Since $\ker \widehat{\psi}$ is a subgroup of F, it must contain all of I. This implies that $\widehat{\psi}$ factors through I to give a group homomorphism $\psi : T = F/I \to P$ such that $\psi((m \otimes n) + I) = \phi(m, n)$. The fact that $\phi = \psi\theta$ is immediate.

Finally, if ψ' were another homomorphism such that $\phi = \psi'\theta$, we would have $\psi((m \otimes n) + I) = \phi(m, n) = \psi'\theta(m, n) = \psi'((m \otimes n) + I)$ for all $m \in M$, $n \in N$. Since the basis elements $\{(m \otimes n) | m \in M, n \in N\}$ generate F as an abelian group, their images $\{(m \otimes n) + I | m \in M, n \in N\}$ generate $T = F/I$ as an abelian group. Thus ψ' and ψ agree on a generating set of T. Since they are group homomorphisms, $\psi' = \psi$. □

Remark 15.9. From now on we use the following standard notation. Given a right R-module M and a left R-module N, we now know from Theorem 15.8 that there exists a tensor product of M and N over R, given by a group homomorphism $\theta : M \times N \to T$ for some abelian group T. By Proposition 15.7, this tensor product is unique up to isomorphism. The standard notation for the abelian group T is $M \otimes_R N$, and we will adopt this notation from now on.

Technically the tensor product of M and N over R is the abelian group $M \otimes_R N$ together with a R-balanced map $\theta : M \times N \to M \otimes_R N$. In practice, we refer to the abelian group $M \otimes_R N$ as the tensor product and suppress the map θ. Instead θ is remembered by writing the element $\theta(m, n)$ as $m \otimes n$ for each $m \in M, n \in N$ (this notation was already suggested by the notation used in the proof of Theorem 15.8). These elements $m \otimes n$ in $M \otimes_R N$ are referred to as *pure tensors*. The fact that θ is R-balanced means that we have the following rules for manipulating pure tensors:

$$(m_1 + m_2) \otimes n = m_1 \otimes n + m_2 \otimes n \text{ for all } m_1, m_2 \in M, n \in N;$$

$$m \otimes (n_1 + n_2) = m \otimes n_1 + m \otimes n_2 \text{ for all } m \in M, n_1, n_2 \in N; \text{and}$$

$$mr \otimes n = m \otimes rn \text{ for all } m \in M, r \in R, n \in N.$$

As we will see shortly, θ is not surjective in general, and so it is important to realize that not all elements of $M \otimes_R N$ are equal to pure tensors. Rather, by the construction in Theorem 15.8, an arbitrary element of $M \otimes_R N$ has the form $\sum_{i=1}^{d} a_i(m_i \otimes n_i)$ for $a_i \in \mathbb{Z}$, $m_i \in M$, and $n_i \in N$. Using that the map θ is additive in the first coordinate, this is the same as $\sum_{i=1}^{d}(a_i m_i \otimes n_i) = \sum_{i=1}^{d}(m'_i \otimes n_i)$ for some $m'_i \in M$. We conclude that *every element of $M \otimes_R N$ is a finite sum of pure tensors.* Of course in general an element can be written as a sum of pure tensors in many different ways.

We have seen that the tensor product $M \otimes_R N$ always exists and is unique up to isomorphism. The proof of existence in Theorem 15.8 is very formal, and unfortunately it does not really give any intuition for what a particular tensor product looks like. For example, the tensor product of R and M over R is given by the natural multiplication map $R \times M \to M$, as we saw in Example 15.6. The proof of Theorem 15.8 also constructs a tensor product of R and M over R as a factor group of a massive free abelian group. This must be isomorphic to M as an abelian group in this case; but this is certainly not obvious. In practice, when working with tensor products, it is usually best to try to understand them using their defining universal property and to forget the formal construction as a factor group of a free abelian group which appeared in the proof of Theorem 15.8.

The tensor product can behave in ways that are quite unintuitive at first. For example, it can easily happen that the tensor product of two nonzero modules is 0.

Lemma 15.10. *In any tensor product $M \otimes_R N$, we have $0 \otimes n = 0 = m \otimes 0$ for all $m \in M, n \in N$.*

Proof. We have $(0 \otimes n) = (0 + 0) \otimes n = 0 \otimes n + 0 \otimes n$. Subtracting, we get $0 \otimes n = 0$. Similarly, $0 = m \otimes 0$. □

Example 15.11. Let G be a torsion abelian group, thought of as a right $\mathbb{Z}$-module. Consider $\mathbb{Q}$ as a left $\mathbb{Z}$-module as usual. Then we claim that $G \otimes_{\mathbb{Z}} \mathbb{Q} = 0$.

Consider a pure tensor in $G \otimes_{\mathbb{Z}} \mathbb{Q}$. It has the form $g \otimes a/b$ for $a, b \in \mathbb{Z}$ with $b \neq 0$. Since G is torsion, $g \cdot n = ng = 0$ for some $n \geq 1$. Then

$$g \otimes a/b = g \otimes an/bn = gn \otimes a/bn = 0 \otimes a/bn = 0,$$

using Lemma 15.10. Thus all pure tensors are equal to 0. Since every element of $G \otimes_{\mathbb{Z}} \mathbb{Q}$ is a finite sum of pure tensors, $G \otimes_{\mathbb{Z}} \mathbb{Q} = 0$ as claimed.

15.3. **Functoriality; module structure on the tensor product.** It is important that the formation of tensor products is *functorial*: this means that the operation $- \otimes_R N$ of tensoring modules with N respects homomorphisms (and similarly in the other coordinate). Here is what we mean precisely.

Lemma 15.12. *Let R be a ring, let M, M' be right R-modules, and let N, N' be left R-modules.*

(1) *Let $f : M \to M'$ be a homomorphism of right R-modules. Then there is a homomorphism of abelian groups $f \otimes 1 : M \otimes_R N \to M' \otimes_R N$ given by $[f \otimes 1](m \otimes n) = f(m) \otimes n$.*

(2) *Let $g : N \to N'$ be a homomorphism of left R-modules. Then there is a homomorphism of abelian groups $1 \otimes g : M \otimes_R N \to M \otimes_R N'$ given by $[1 \otimes g](m \otimes n) = m \otimes g(n)$.*

Before beginning the proof we make some comments about statements like this. It is not at all clear that a formula such as $[f \otimes 1](m \otimes n) = f(m) \otimes n$ defines a function. The problem is that it is not clear it is well-defined, as there are many relations among the pure tensors; for example, two pure tensors might well be equal. (Remember that $m \otimes n$ means $\theta(m, n)$ for the underlying map

$\theta : M \times N \to M \otimes_R N$ of the tensor product, and there is no reason why θ should be injective.) So we need to make sure those relations are respected. The best way to do this is to use the universal property of the tensor product, as we will see in the proof.

Note also that the formula $[f \otimes 1](m \otimes n) = f(m) \otimes n$, even once we show it is well-defined, only gives the action of the function on pure tensors. But since we require $f \otimes 1$ to be a homomorphism of groups, the action on an arbitrary element, i.e. a sum of pure tensors, is uniquely determined. For this reason it is standard to use only pure tensors in the formulas for functions and actions, and often one only verifies these formulas for pure tensors in proofs. You should be careful not to let the appearance of those formulas seduce you into forgetting that not every element of the tensor product is a pure tensor.

Proof. (1) We define a function $\phi : M \times N \to M' \otimes N$ by $\phi(m, n) = f(m) \otimes n$. Using the rules for manipulating the tensor product symbol and the fact that f is a homomorphism of right modules, it is easy to check that ϕ is R-balanced. It follows from the universal property of the tensor product that there is a unique group homomorphism $f \otimes 1 : M \otimes_R N \to M' \otimes_R N$ such that $[f \otimes 1](m \otimes n) = f(m) \otimes n$ for $m \in M, n \in N$, as required.

(2) This is proved in a symmetric manner. □

Before studying some more examples we should discuss when $M \otimes_R N$ is actually a module and not just an abelian group.

Definition 15.13. Let M be an abelian group. Then M is called an (R, S)-bimodule if M is both a left R-module and a right S-module, and the two module structures are compatible in the sense that $(rm)s = r(ms)$ for all $r \in R, m \in M, s \in S$.

Proposition 15.14. *Let M be a right R-module and N a left R-module.*

1. *If M is an (S, R)-bimodule, then $M \otimes_R N$ is a left S-module, where $s \cdot (m \otimes n) = sm \otimes n$.*
2. *If N is an (R, T)-bimodule, then $M \otimes_R N$ is a right T-module, where $(m \otimes n) \cdot t = m \otimes nt$.*
3. *If both (1) and (2) occur then $M \otimes_R N$ is an (S, T)-bimodule.*

Proof. (1) For any $s \in S$, define $\ell_s : M \to M$ by $\ell_s(m) = sm$. This "left multiplication by s" map is not a homomorphism of M as a left S-module in general (unless S is commutative) but it *is* always a right R-module map, since $\ell_s(mr) = s(mr) = (sm)r = \ell_s(m)r$ for $r \in R$.

By the functoriality of the tensor product given in Lemma 15.12, we get a homomorphism of abelian groups $\ell_s \otimes 1 : M \otimes_R N \to M \otimes_R N$ such that $[\ell_s \otimes 1](m \otimes n) = (sm \otimes n)$. Thus there is in fact a well-defined left action of S on $M \otimes_R N$ for which the action on pure tensors is given by $s \cdot (m \otimes n) = [\ell_s \otimes 1](m \otimes n) = sm \otimes n$, and for which left action by s is an abelian group homomorphism. This also implies one of the module axioms $((s \cdot (x + y) = s \cdot x + s \cdot y)$ for $s \in S$, $x, y \in M \otimes_R N)$, and the others are easy to check.

(2) This is proved by a completely symmetric proof to the proof of part (1).

(3) On pure tensors we have

$$[s \cdot (m \otimes n)] \cdot t = (sm \otimes n) \cdot t = (sm \otimes nt) = s \cdot (m \otimes nt) = s \cdot [(m \otimes n) \cdot t]$$

and this extends immediately to the action on a finite sum of pure tensors, i.e. a general element of $M \otimes_R N$. □

Proposition 15.14 is analogous to earlier observations we made about $\mathrm{Hom}_R(M, N)$ for two left R-modules M and N. In general this Hom-space is just an abelian group, but in an exercise on the homework you verified that if either M or N is a bimodule then $\mathrm{Hom}_R(M, N)$ obtains a module structure as well, and it is a bimodule if both M and N are bimodules.

Example 15.15. Let R be a ring with ideal I. Thus R/I is naturally an (R, R)-bimodule. Let M be a left R-module. We claim that $(R/I) \otimes_R M \cong M/IM$ as left R-modules. (This generalizes Example 15.6.)

Since R/I is an (R, R)-bimodule, $R/I \otimes_R M$ is a left R-module by Proposition 15.14. Define a map $\phi : R/I \times M \to M/IM$ by $\phi(r + I, m) = rm + IM$. If $r + I = r' + I$, then $r - r' \in I$, so $(r - r')m \in IM$ and hence $rm + IM = r'm + IM$. Hence ϕ is well-defined. It is now easy to check that ϕ is R-balanced. Thus the universal property of the tensor product gives us a unique group homomorphism $\psi : R/I \otimes M \to M/IM$ such that $\psi((r+I) \otimes m) = rm + IM$ for all $r \in R, m \in M$. But from this formula, in fact we see that ψ is an R-module homomorphism, since for $x \in R$ we have $\psi(x(r+I \otimes m)) = \psi(xr+I \otimes m) = (xr)m+IM = x(rm)+IM = x(rm+IM) = x\psi(r+I \otimes m)$.

To show that ψ is an isomorphism, one can define an inverse map $\rho : M/IM \to (R/I) \otimes_R M$ by $\rho(m+IM) = (1+I) \otimes m$. To see that this is well defined, if $m+IM = m'+IM$, then $m-m' \in IM$, say $m - m' = \sum x_i m_i$ with $x_i \in I$, $m_i \in M$. Then

$$\begin{aligned}(1 + I) \otimes m - (1 + I) \otimes m' = (1 + I) \otimes (m - m') = (1 + I) \otimes \sum x_i m_i = \sum (1 + I) x_i \otimes m_i \\ = \sum (x_i + I) \otimes m_i = \sum (0 + I) \otimes m_i = 0.\end{aligned}$$

Thus ρ is well defined. It is obvious that $\psi\rho = 1_{M/IM}$. On the other hand, $\rho\psi((r + I) \otimes m) = \rho(rm + I) = (1 + I) \otimes rm = (1 + I) r \otimes m = (r + I) \otimes m$, so $\rho\psi$ is also the identity and we are done.

15.4. **The commutative case.** The tensor product over a commutative ring R is often described in a slightly different way. Recall that any right R-module M is naturally also a left R-module with $r \cdot m = mr$; symmetrically, every left module is also a right module. So there is no need to use one left and one right module in the definition of a tensor product, and usually the definition is made in terms of left modules only.

Definition 15.16. Let R be a commutative ring and let M, N, and P be left R-modules. A function $\phi : M \times N \to P$ is *R-bilinear* if

(1) $\phi(r_1m_1 + r_2m_2, n) = r_1\phi(m_1, n) + r_2\phi(m_2, n)$ for all $r_1, r_2 \in R$, $m_1, m_2 \in M$, $n \in N$; and
(2) $\phi(m, r_1n_1 + r_2n_2) = r_1\phi(m, n_1) + r_2\phi(m, n_2)$ forall $r_1, r_2 \in R$, $m \in M$, $n_1, n_2 \in N$.

In the commutative case, the universal property for the tensor product can be (and usually is) phrased in terms of R-bilinear maps, as in part (3) of the next result.

Proposition 15.17. *Let R be a commutative ring and let M and N be left R-modules. Considering M as a right R-module, we can define $T = M \otimes_R N$. Then*

(1) *T is naturally a left R-module with $r \cdot (m \otimes n) = rm \otimes n = m \otimes rn$.*
(2) *The map $\theta : M \times N \to M \otimes_R N$ given by $\theta(m, n) = m \otimes n$ is R-bilinear.*
(3) *Given any R-module P and an R-bilinear map $\phi : M \times N \to P$, there is a unique R-module homomorphism $\psi : M \otimes_R N \to P$ such that $\psi\theta = \phi$.*

Proof. (1) Since R is commutative, we can actually think of M as an (R, R)-bimodule with the same left and right actions: $mr = rm$ for $r \in R$, $m \in M$. (This is a bimodule since $s(mr) = s(rm) = (sr)m = (rs)m = r(sm) = (sm)r$.) Then $T = M \otimes_R N$ obtains a left R-module structure by Proposition 15.14, where $r \cdot (m \otimes n) = rm \otimes n$. We also have $rm \otimes n = mr \otimes n = m \otimes rn$.

(2) By the R-module structure on T given in (1), we have

$$(r_1m_1 + r_2m_2) \otimes n = (r_1m_1 \otimes n) + (r_2m_2 + n) = r_1(m_1 \otimes n) + r_2(m_2 \otimes n)$$

and

$$m \otimes (r_1n_1 + r_2n_2) = m \otimes r_1n_1 + m \otimes r_2n_2 = r_1(m \otimes n_1) + r_2(m \otimes n_2).$$

(3) The map ϕ is also R-balanced, since $mr \otimes n = rm \otimes n = m \otimes rn$ by (1). Thus there is a unique group homomorphism $\psi : T \to P$ such that $\psi(m \otimes n) = \phi(m, n)$. Then ψ is an R-module homomorphism because $\psi(r(m \otimes n)) = \psi(rm \otimes n) = \phi(rm, n) = r\phi(m, n) = r\psi(m \otimes n)$. □

An important special case of the tensor product over a commutative ring R is the case where R is a field F. In this case one can get a very explicit description of the tensor product of two vector spaces over F.

Theorem 15.18. *Let V and W be vector spaces over the field F. Suppose that $\{v_i | i \in I\}$ is an F-basis for V and $\{w_j | j \in J\}$ is an F-basis for W.*

Then $V \otimes_F W$ is an F-vector space with basis $\{v_i \otimes w_j | i \in I, j \in J\}$. In particular, if $\dim_F V = m$ and $\dim_F W = n$ then $\dim_F V \otimes W = mn$.

Proof. Given a pure tensor $v \otimes w \in V \otimes_F W$, write $v = \sum a_i v_i$ and $w = \sum b_j w_j$, where all but finitely many of the a_i and b_j are nonzero. Then

$$v \otimes w = (\sum_i a_i v_i) \otimes (\sum_j b_j w_j) = \sum_i \sum_j a_i b_j (v_i \otimes w_j).$$

Thus any pure tensor is in the F-span of $\{v_i \otimes w_j | i \in I, j \in J\}$. We have also seen that an arbitrary element of $V \otimes_F W$ is a finite sum of pure tensors. Hence $\{v_i \otimes w_j | i \in I, j \in J\}$ spans $V \otimes_F W$.

Now suppose that we have an independence relation $\sum_{i,j} a_{ij}(v_i \otimes w_j) = 0$, where $a_{ij} = 0$ for all but finitely many $(i, j) \in I \times J$. Then we have

$$0 = \sum_i \sum_j (a_{ij} v_i \otimes w_j) = \sum_j (\sum_i a_{ij} v_i) \otimes w_j = \sum_j v'_j \otimes w_j$$

for some elements $v'_j \in V$. For each $j \in J$ let $w^*_j \in W^* = \mathrm{Hom}_F(W, F)$ be the linear functional with $w^*_j(w_i) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

For any $k \in J$, the map $\phi : V \times W \to V$ defined by $\phi(v, w) = w^*_k(w)v$ is F-bilinear, as is easy to check. Thus there is an F-linear map $\psi : V \otimes_F W \to V$ with $\psi(v \otimes w) = w^*_k(w)v$. Applying this to our relation above gives $0 = \psi(0) = \sum_j \psi(v'_j \otimes w_j) = \sum_j w^*_k(w_j)v'_j = v'_k$. We conclude that $v'_k = 0$ for all k. Then $v'_k = \sum_i a_{ik} v_i = 0$, and by the linear independence of the v_i, we get $a_{ik} = 0$ for all i. Since k was arbitrary, $a_{ik} = 0$ for all $i \in I, k \in J$, and thus $\{v_i \otimes w_j | i \in I, j \in J\}$ is also F-independent. Hence it is an F-basis as claimed. □

15.5. **Extension of scalars.** Suppose that we have a ring homomorphism $\phi : R \to S$. Recall that if M is a left S-module, there is an easy way to make M into a left R-module: just define $r \cdot m = \phi(r)m$ for $r \in R$, $m \in m$. This is called *restriction of scalars*, as we have already mentioned, since in the case where R is a subring of S and $\phi : R \to S$ is just the inclusion map, then we areliterally just restricting the elements that act on M to a smaller set.

Now that we have developed the tensor product, we can easily define a process that goes the other way.

Definition 15.19. Let $\phi : R \to S$ be a ring homomorphism. Suppose that M is a left R-module. Then $S \otimes_R M$ is naturally a left S-module, where $s \cdot (t \otimes m) = st \otimes m$. This process is called *extension of scalars.*

Again, when R is a subring of S we are extending the ring acting to a larger ring; hence the name.

The fact that $S \otimes_R M$ is a left S-module with action on pure tensors by the given formula is immediate from the fact that S is an (S, R)-bimodule and Proposition 15.14. Here the bimodule structure on S is given by the natural S-action by multiplication on the left, and the right R-action is $s \cdot r = s\phi(r)$, for $s \in S$, $r \in R$. In other words, S is a right R-module by restricting the scalars from the action of S on the right by multiplication.

Let us give several applications of extension by scalars.

Example 15.20. Suppose that $F \subseteq K$ is an inclusion of fields. If V is an F-vector space, then $K \otimes_F V$ is a K-vector space by extension of scalars.

Suppose that $\mathcal{B} = \{v_i | i \in I\}$ is an F-basis of V. Then $\mathcal{B}' = \{1 \otimes v_i | i \in I\}$ is a K-basis of $K \otimes_F V$ (we leave this as an exercise). In other words, $\dim_K(K \otimes_F V) = \dim_F V$. Moreover, suppose that $\phi : V \to V$ is an F-linear transformation. Then $1 \otimes \phi : K \otimes_F V \to K \otimes_F V$ is a K-linear transformation. If $\dim_F V$ is finite and we consider the matrix $M^{\mathcal{B}}_{\mathcal{B}}(\phi)$ of ϕ with respect to $\mathcal{B}$, then one may check that $M^{\mathcal{B}'}_{\mathcal{B}'}(1 \otimes \phi)$ is the same matrix $M^{\mathcal{B}}_{\mathcal{B}}(\phi)$. So the linear transformation is given by the same matrix, just working over a larger field.

This is a very useful operation, for example, if we are initially working over a field F and would like to work over an algebraically closed one K. (We will prove later in the notes that F is always a subfield of an algebraically closed field K.) By extending V to $K \otimes_F V$ and the map ϕ to $1 \otimes \phi$ we put ourselves in a setting where the Jordan canonical form is defined.

Example 15.21. Suppose that R is an integral domain. Let K be the field of fractions of R. If M is a left R-module we define the *rank* of M to be $\dim_K(K \otimes_R M)$.

The point is that by extending scalars to a field K, we have access to the notion of dimension of a vector space, which was not available over the original ring R. One may check that if R is a PID and M is a finitely generated R-module, then writing $M \cong R^r \oplus T$ where T is torsion, the rank of M is r. So this notion of rank just recovers the rank of the free part of a finitely generated module in this case.

The rank is defined over any integral domain however, which makes it more generally applicable. One may show that in general the rank of a finitely generated module M is equal to the maximum r such that there is an R-submodule of M isomorphic to R^r. Though we could use this as a definition of rank instead, the properties of rank are easiest to prove by extending scalars to K.

Example 15.22. Suppose R is any commutative ring. We know that R has some maximal ideal $\mathfrak{m}$. We have the homomorphism $\phi : R \to R/\mathfrak{m}$ where $F = R/\mathfrak{m}$ is a field.

This gives another way one can sometimes reduce problems about R-modules to problems about fields. Given an R-module M, when we "extend" scalars using the map ϕ we get the F-module $F \otimes_R M \cong M/\mathfrak{m}M$, using Example 15.15.

If M is a free R-module, say $M \cong \bigoplus_{i \in I} R$, then by picking a basis one may show that $F \otimes_R M \cong \bigoplus_{i \in I} F$ as F-modules; that is, $F \otimes_R M$ is a vector space with dimension $|I|$ over F. This can be used to show that two isomorphic free modules over R must have the same rank, by reducing to the case of vector spaces.

As another example, Let R be a local commutative ring with maximal ideal $\mathfrak{m}$ (that is, $\mathfrak{m}$ is the unique maximal ideal of R). In this case $F = R/\mathfrak{m}$ is called the *residue field*. If M is a finitely generated R-module, then $\dim_F(F \otimes_R M) = \dim_F M/\mathfrak{m}M$ is equal to the minimum number n such that M can be generated by n elements as an R-module. This is a consequence of the result known as Nakayama's Lemma.

15.6. **Tensor products of algebras.**

Definition 15.23. Let R be a commutative ring. A ring A is an R-*algebra* if A is also a left R-module, and for all $r \in A$, $a, b \in A$ we have $r \cdot ab = (r \cdot a)b = a(r \cdot b)$.

A homomorphism of R-algebras is a function $f : A \to B$ which is both a ring homomorphism and a R-module homomorphism. It is an isomorphism of algebras if f is bijective. A *subalgebra* of an algebra A is a subset which is both a submodule over R and a subring; so it is naturally an R-algebra again.

Thus an algebra is both a ring and a module, with those structures being compatible in a certain way. The compatibility can be framed in the following alternative way which is perhaps more natural.

Remark 15.24. If A is an R-algebra, then there is a function $\phi : R \to A$ given by $\phi(r) = r \cdot 1$. This is clearly a homomorphism of abelian groups with $\phi(1) = 1$; moreover $\phi(rs) = (rs) \cdot 1 = r \cdot (s \cdot 1)$ (by module axioms) and $r \cdot (s \cdot 1) = r \cdot ((1)(s \cdot 1)) = (r \cdot 1)(s \cdot 1) = \phi(r)\phi(s)$. Thus ϕ is a homomorphism of rings.

We also check using the axioms of an algebra that $(r \cdot 1)a = r \cdot (1a) = r \cdot a = r \cdot (a1) = a(r \cdot 1)$ for all $r \in R$, $a \in A$. This shows that $\phi(R)$ is contained in the center $Z(A)$ of A.

Conversely, if R and A are arbitrary rings with R commutative and $\phi : R \to A$ is a ring homomorphism such that $\phi(R) \subseteq Z(A)$, then defining a left R-module structure on A by $r{\cdot}a = \phi(r)a$ it is not hard to check that A is an R-algebra.

In this way, one can see that to make a ring A into an R-algebra is equivalent to giving a homomorphism $\phi : R \to A$ such that $\phi(R) \subseteq Z(A)$.

Example 15.25. Let R be a commutative ring. The polynomial ring $R[x]$ and the power series ring $R[[x]]$ are both R-algebras in an obvious way. Similarly, the polynomial ring $R[x_1, \dots, x_n]$ in finitely many variables is an R-algebra. A noncommutative example of an R-algebra is the ring $M_n(R)$ of $n \times n$ matrices over R, where R acts by scalar multiplication.

We are especially interested in algebras over fields. If A is an F-algebra for a field F, then by the remark above we can think of this via a homomorphism $\phi : F \to A$ with $\phi(F) \subseteq Z(A)$. Since F is a field, ϕ is injective and we usually identify F with its image $\phi(F)$. Thus an algebra over a field F is just a ring A together with a subring of the center $Z(A)$ which is isomorphic to F, which gives A an F-module structure (i.e. vector space structure) by restriction.

Example 15.26. Let F be a field. The polynomial ring $F[x]$ and the power series ring $F[[x]]$ are both F-algebras in an obvious way. Similarly, the polynomial ring $F[x_1, \dots, x_n]$ in finitely many variables is an F-algebra. A noncommutative example of an F-algebra is the ring $M_n(F)$ of $n \times n$ matrices over F, where the copy of F in the center is the subring of scalar matrices $\{\lambda I | \lambda \in F\}$.

We do not define the notion of R-algebra when R is not commutative. Remark 15.24 shows why this wouldn't be useful: if we just apply the given definition to an arbitrary ring R, then $\phi(R)$ must lie in the center of A and thus be a commutative ring. If $I = \ker \phi$ then A is an R/I-module where R/I is commutative, and we might as well just think of A as an algebra over the commutative ring R/I.

One of the reasons why algebras are useful is that we can take a tensor product of two algebras and obtain another algebra.

Theorem 15.27. *Let A and B be algebras over a (commutative) ring R. Then $A \otimes_R B$ is again an R-algebra, where it is an R-module via Proposition 15.17 and the product is given by $(a \otimes b)(c \otimes d) = ac \otimes bd$.*

Proof. Since R is commutative, we know that $A \otimes_R B$ is again an R-module, where $r \cdot (a \otimes b) = (ra \otimes b) = (a \otimes rb)$ for $r \in R$, $a \in A$, $b \in B$.

Now for $a \in A, b \in B$, we define a map $\psi_{a,b} : A \otimes_R B \to A \otimes_R B$ by the formula $c \otimes d \mapsto ac \otimes bd$. This exists from the universal property since the function $A \times B \to A \otimes_R B$ given by $(c, d) \mapsto ac \otimes bd$ is R-bilinear. So $\psi_{a,b} \in \operatorname{End}_R(A \otimes_R B)$.

Then define $\widehat{\Psi} : A \times B \to \operatorname{End}_R(A \otimes_R B)$ by $(a, b) \mapsto \psi_{a,b}$. Again this is an R-bilinear map. So we get an R-module homomorphism $\Psi : A \otimes_R B \to \operatorname{End}_R(A \otimes_R B)$ such that $\Psi(a \otimes b) = \psi_{a,b}$.

In other words, we can think of $\Psi(a \otimes b)$ as "left multiplication by $a \otimes b$". This allows us to define a product $*$ on $A \otimes_R B$, where for $x, y \in A \otimes_R B$ we define $x * y = [\Psi(x)](y)$. On pure tensors this product has the formula $(a \otimes b)(c \otimes d) = ac \otimes bd$ as claimed. It is now routine to check that with this product that $A \otimes_R B$ is a ring, and in fact an R-algebra. □

While the product operation in $A \otimes_R B$ is done just my multiplying "coordinate-wise", the tensor product of algebras is a very different operation from the direct product of rings, because the underlying space $A \otimes_R B$ is much different from $A \times B$. The following examples should help illustrate how the tensor product of algebras behaves.

Example 15.28. Let R be a commutative ring. Then the ring $M_n(R)$ consisting of matrices with entries in R is an R-algebra. Let A be another R-algebra; for simplicity assume that the corresponding map $\phi : R \to A$ defined by $\phi(r) = r \cdot 1$ is injective. We claim that $A \otimes_R M_n(R) \cong M_n(A)$ as R-algebras.

We can identify R with its image $\phi(R)$ and thus think of R as a subring of $Z(A)$. There is an R-bilinear map $A \times M_n(R) \to M_n(A)$ given by $(a, B) \mapsto aB$ for $a \in A, B \in M_n(R)$. (a scalar a times a matrix B means multiply every entry of the matrix B by a on the left, or in other words take the matrix product $(aI)B$.) So we get an R-module homomorphism $\psi : A \otimes M_n(R) \to M_n(A)$. Note that for $a \in A$, and any matrix $B \in M_n(R)$, the matrices aI and B commute (this is because $R \subseteq Z(A)$). Thus we have

$$\psi((a \otimes B)(c \otimes D)) = \psi(ac \otimes BD) = (acI)BD = (aI)B(cI)D = \psi(a \otimes B)\psi(c \otimes D).$$

Thus ψ is a homomorphism of R-algebras.

Let e_{ij} be the matrix with 1 in the (i, j)-entry and 0 in all other entries (in $M_n(R)$ or in $M_n(A)$). These n^2 elements are often traditionally called *matrix units* (but they are not units in the ring). An arbitrary matrix $B \in M_n(A)$, where $B_{ij} = a_{ij} \in A$, can be written as $B = \sum_{i,j} a_{ij} e_{ij}$. Now define $\rho : M_n(A) \to A \otimes M_n(R)$ by $\rho(\sum_{i,j} a_{ij} e_{ij}) = \sum_{i,j} a_{ij} \otimes e_{ij}$. It is obvious that $\psi\rho = 1_{M_n(A)}$. Conversely, on pure tensors we have $\rho\psi(a \otimes \sum_{i,j} r_{i,j} e_{i,j}) = \rho(\sum_{i,j} ar_{i,j} e_{i,j}) = \sum_{i,j} ar_{i,j} \otimes e_{i,j} = \sum_{i,j} a \otimes r_{i,j} e_{i,j}$ since $r_{i,j} \in R$. It follows that $\rho\psi = 1_{A \otimes M_n(R)}$. Thus ψ is an isomorphism.

Example 15.29. Let R be a commutative ring. Then $M_n(R) \otimes_R M_m(R) \cong M_{mn}(R)$ as R-algebras.

The previous example actually gives $M_n(R) \otimes_R M_m(R) \cong M_m(M_n(R))$ as R-algebras. We leave the proof that $M_m(M_n(R)) \cong M_{mn}(R)$ as an exercise for the reader (think about multiplication of matrix blocks).

Example 15.30. For any R-algebra A, $A \otimes_R R[x] \cong A[x]$ as R-algebras. This is proved similarly as in Example 15.28 and left as an exercise.

For example, this gives $R[y] \otimes_R R[x] \cong (R[y])[x]$. By definition this is the polynomial ring in two variables $R[y, x]$.

Example 15.31. Tensor product of fields over a common subfield can behave in unexpected ways.

For example, consider $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ as an $\mathbb{R}$-algebra. We know that $\mathbb{C}$ is a vector space over $\mathbb{R}$ with basis $\{1, i\}$ and so by our description of the tensor product over a field, $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is a vector space over $\mathbb{R}$ with basis $\{1 \otimes 1, 1 \otimes i, i \otimes 1, i \otimes i\}$. A pure tensor has the form $(a + bi) \otimes (c + di) = ac(1 \otimes 1) + bc(i \otimes 1) + ad(1 \otimes i) + bd(i \otimes i)$. This is 0 only if $ac = bc = ad = bd = 0$, which happens only if $a = b = 0$ or $c = d = 0$. Thus a pure tensor $(w \otimes z)$ is zero only if $w = 0$ or $z = 0$. In particular, since $(w \otimes z)(x \otimes y) = wx \otimes zy$, a product of nonzero pure tensors is nonzero.

This is a case where thinking about pure tensors only for intuition can lead one astray, however, as $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is not a domain. One may easily check that

$$[(1 \otimes i) - (i \otimes 1)][(1 \otimes i) + (i \otimes 1)] = 0.$$

In fact $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C} \cong \mathbb{C} \times \mathbb{C}$ as $\mathbb{R}$-algebras, as you will be asked to show on the homework.

15.7. **Exact sequences and flatness.**

Definition 15.32. Fix a ring R and let M, N, P be left R-modules. Consider a sequence of R-module homomorphisms

$$M \xrightarrow{f} N \xrightarrow{g} P.$$

The sequence is called *exact* at N if $\ker g = f(M)$. More generally, it is called a *complex* if $f \circ g = 0$, i.e. $f(M) \subseteq \ker g$. In this case, the *homology group* at N is the R-module $H = (\ker g)/f(M)$. So a complex is exact at N if and only if $H = 0$.

Note that $0 \longrightarrow N \xrightarrow{g} P$ is automatically a complex, and it is exact at N if and only if $0 = \ker(g)$, that is, g is an injective homomorphism. Dually, $M \xrightarrow{f} N \longrightarrow 0$ is a complex which is exact at N if and only if $f(M) = N$, that is, f is surjective.

A longer sequence of modules and homomorphisms is called an exact sequence (or a complex) if it is exact (or a complex, respectively) at every spot that has an arrow both entering and leaving. In particular we have

Definition 15.33. A sequence of R-modules and maps

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

which is exact (that is, exact at M, N, and P) is called a *short exact sequence.*

From the comments above, we see that explicitly the definition of short exact sequence is equivalent to f being injective, g being surjective, and $f(M) = \ker(g)$. Given such a short exact sequence, we also say that N is an *extension* of P by M. This is because by the first isomorphism theorem applied to g, $N/f(M) = N/(\ker g) \cong P$. Also since f is injective, $f(M) \cong M$. So N is built out of the submodule $f(M) \cong M$ and the factor module $N/f(M) \cong P$; the short exact sequence tells us precisely how M and P are put together to form N. In this point of view, a short exact sequence is just a convenient way to represent the information of an extension of two modules.

If you have taken a course in algebraic topology, you will recognize the definitions above. The development of Algebraic Topology had a lot of influence on algebra. A whole field of homological algebra developed from this which abstracts definitions and techniques derived from topology. These homological techniques have had and continue to have great importance in the study of algebra. In this short section we will just be able to scratch the surface. We recommend the book of Weibel, ("An Introduction to Homological Algebra"), or of Rotman (also called "An Introduction to Homological Algebra") if you would like to learn more about this interesting subject.

We would like to consider the following question. To what extent does the operation of tensoring with a module preserve exact sequences? More precisely, suppose that

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$$

is a short exact sequence where M, N, and P are left R-modules and f and g are R-module homomorphisms. If Q is a right R-module, then we can apply the operation $Q \otimes_R -$ to the entire sequence using the functoriality result of Lemma 15.12, to obtain a sequence

$$0 \longrightarrow Q \otimes_R M \xrightarrow{1\otimes f} Q \otimes_R N \xrightarrow{1\otimes g} Q \otimes_R P \longrightarrow 0$$

in which the maps are homomorphisms of abelian groups. If Q is an (S, R)-bimodule the maps are even left S-module homomorphisms. This sequence is a complex but it turns out to not always be a short exact sequence, as we will see next; the problem is at the $Q \otimes_R M$ spot.

Theorem 15.34. *Let $0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0$ be a short exact sequence of left R-modules. Let Q be a right R-module. Then*

$$Q \otimes_R M \xrightarrow{1\otimes f} Q \otimes_R N \xrightarrow{1\otimes g} Q \otimes_R P \longrightarrow 0$$

is exact, i.e. exact at the $Q \otimes_R N$ and $Q \otimes_R P$ spots.

This result can be described by saying that the operation of tensoring with a module is *right exact*. It preserves the exactness at the right two terms of the sequence only.

Proof. Let $q \otimes p$ be a pure tensor in $Q \otimes_R P$. Since g is surjective, there is $n \in N$ such that $g(n) = p$. Then $[1 \otimes g](q \otimes n) = q \otimes p$. Thus all pure tensors are in the image of $1 \otimes g$. Since all elements in $Q \otimes_R P$ are sums of pure tensors and $1 \otimes g$ is a homomorphism of abelian groups, $1 \otimes g$ is surjective. Thus we have exactness at the $Q \otimes_R N$ spot.

Now since $g \circ f = 0$, note that $(1 \otimes g) \circ (1 \otimes f) = 0$. Thus our supposed right exact sequence is at least a complex, in other words we have $\operatorname{im}(1 \otimes f) \subseteq \ker(1 \otimes g)$. Let $L = (Q \otimes_R N)/\operatorname{im}(1 \otimes f)$, which is again a left R-module. Note that because we have $\operatorname{im}(1 \otimes f) \subseteq \ker(1 \otimes g)$, we can define a map $\overline{1 \otimes g} : L \to Q \otimes_R P$ by $\overline{1 \otimes g}(x + \operatorname{im}(1 \otimes f)) = [1 \otimes g](x)$ for $x \in Q \otimes_R N$. Since $1 \otimes g$ is surjective, so is $\overline{1 \otimes g}$.

Now we claim that there is a homomorphism of abelian groups

$$\psi : Q \otimes P \to L = (Q \otimes_R N)/\operatorname{im}(1 \otimes f)$$

given by the formula $\psi(q \otimes p) = q \otimes n + \operatorname{im}(1 \otimes f)$, where $n \in N$ is any element such that $g(n) = p$. To see that this formula does not depend on the choice of n, note that if $g(n) = g(n') = p$ then $n - n' \in \ker(g) = \operatorname{im}(f)$ and so $n' - n = f(m)$ for some $m \in M$. Then $q \otimes n - q \otimes n' = q \otimes (n - n') = q \otimes f(m) \in \operatorname{im}(1 \otimes f)$. thus $q \otimes n + \operatorname{im}(1 \otimes f) = q \otimes n' + \operatorname{im}(1 \otimes f)$. The existence of the map ψ is then proved in the usual way by first defining an R-balanced map and applying the universal property.

Now for $(q \otimes n) + \operatorname{im}(1 \otimes f) \in L$ applying $\overline{1 \otimes g}$ gives $q \otimes g(n)$ and then applying ψ gives $q \otimes n' + \operatorname{im}(1 \otimes f)$ for some n' such that $g(n) = g(n')$. As we just saw, $q \otimes n + \operatorname{im}(1 \otimes f) = q \otimes n' + \operatorname{im}(1 \otimes f)$. Thus $\psi \circ \overline{1 \otimes g} = 1_L$.

In particular $\overline{1 \otimes g}$ must be injective. So $0 = \ker(\overline{1 \otimes g}) = \ker(1 \otimes g)/\operatorname{im}(1 \otimes f)$ and thus $\ker(1 \otimes g) = \operatorname{im}(1 \otimes f)$. Thus we have exactness at the $Q \otimes_R N$ spot and we are done. □

It is easy to give an example showing that the result above is the best we can do; the operation of tensoring with a module does not preserve exactness at the left in general.

Example 15.35. Let R be an integral domain which is not a field. Let x be a nonunit in R. Consider the module homomorphism $f : R \to R$ given by $f(a) = ax$. Since R is a domain, f is injective. Thus f is the left map in a short exact sequence $0 \to R \xrightarrow{f} R \xrightarrow{\pi} R/xR \to 0$ where π is the natural surjection, and $R/xR \neq 0$ since x is not a unit.

Now let us tensor this short exact sequence with R/xR, obtaining

$$0 \to (R/xR) \otimes_R R \xrightarrow{1 \otimes f} (R/xR) \otimes_R R \xrightarrow{1 \otimes \pi} (R/xR) \otimes_R (R/xR) \to 0.$$

We know the resulting sequence will be exact at the right by Theorem 15.34. Let us see that it is not exact at the left; in other words $1 \otimes f$ is not injective.

We have seen that there is an isomorphism $(R/xR) \otimes_R R \to R/xR$ given by $(a+xR) \otimes b \mapsto ab+xR$; see Example 15.15. In particular, $(R/xR) \otimes_R R \neq 0$. On the other hand,

$$[1 \otimes f]((a + xR) \otimes b) = (a + xR) \otimes xb = x(a + xR) \otimes b = (xa + xR) \otimes b = (0 + xR) \otimes b = 0$$

which implies that $1 \otimes f = 0$. In particular, $1 \otimes f$ is not injective.

The following definition is made to focus on those modules that do not have the problem of failure of left exactness as in the previous example.

Definition 15.36. Let R be any right. A right R-module Q is called *flat* (over R) if for all short exact sequences of left R-modules $0 \to M \xrightarrow{f} N \xrightarrow{g} P \to 0$, the sequence obtained by applying $Q \otimes_R -$ to this short exact sequence,

$$0 \to Q \otimes_R M \xrightarrow{1 \otimes f} Q \otimes_R N \xrightarrow{1 \otimes g} Q \otimes_R N \to 0,$$

is again short exact.

Because of Theorem 15.34, it is easy to see that Q is a flat right R-module if and only if for all injective homomorphisms of left R-modules $f : M \to N$, then $1 \otimes f : Q \otimes_R M \to Q \otimes_R N$ is still injective.

In the terminology of category theory, $Q \otimes_R -$ is what is known as a *functor*: it is an operation that sends every left R-module M to an abelian group $Q \otimes_R M$, and comes along with an action on homomorphisms which sends a module homomorphism $f : M \to N$ to an abelian group homomorphism $1 \otimes f : Q \otimes_R M \to Q \otimes_R N$. A functor which when applied to a short exact sequence returns another short exact sequence is called an *exact* functor. So Q is flat if $Q \otimes_R -$ is an exact functor, by definition. We don't have time here to go more in the details of category theory, but may occasionally adopt this terminology.

Of course there is nothing special about the side on which we have chosen to make this definition: a left R-module L is called flat if $- \otimes_R L$ is an exact functor when applied to short exact sequences of right R-modules.

Example 15.37. Let R be an integral domain which is not a field. If x is not a unit in R, then R/xR is an R-module which is not flat, by Example 15.35. In fact, it is possible to show that a flat module over an integral domain must be torsionfree.

Now let us give examples of modules that are flat by showing that free modules are always flat. We leave the following result to the reader; it is a good exercise in applying the universal property of the tensor product.

Lemma 15.38. *Let $\{M_\alpha | \alpha \in I\}$ be an indexed family of right R-modules. For any left R-module N, we have an isomorphism of abelian groups*

$$\Phi : \big(\bigoplus_{\alpha \in I} M_\alpha\big) \otimes_R N \to \bigoplus_{\alpha \in I} \big(M_\alpha \otimes_R N\big),$$

where $\Phi((m_\alpha)_{\alpha \in I} \otimes n) = (m_\alpha \otimes n)_{\alpha \in I}$.

Remark 15.39. The corresponding statement for products is not true in general: there is a natural homomorphism $\Psi : (\prod_{\alpha \in I} M_\alpha) \otimes_R N \to \prod_{\alpha \in I}(M_\alpha \otimes_R N)$ given by the same formula, but it need not be an isomorphism when the index set I is infinite.

Proposition 15.40. *Let F be a free right R-module. Then F is flat.*

Of course, the same result holds for left R-modules.

Proof. We know that $F \cong \bigoplus_{\alpha \in I} R$ for some index set I. It is easy to see that flatness is an invariant of a module up to isomorphism, so we just need to prove that $\bigoplus_{\alpha \in I} R$ is flat. Now for any right module M we have

$$(\bigoplus_{\alpha \in I} R) \otimes_R M \cong \bigoplus_{\alpha \in I} (R \otimes_R M) \cong \bigoplus_{\alpha \in I} M$$

using Lemma 15.38 and the fact that $R \otimes_R M \cong M$, as in Example 15.15.

If $f : M \to N$ is an injective homomorphism of right R-modules, using the isomorphisms above it is straightforward to check that $1 \otimes f : (\bigoplus_{\alpha \in I} R) \otimes_R M \to (\bigoplus_{\alpha \in I} R) \otimes_R N$ can be identified with the homomorphism $\bigoplus f : \bigoplus_{\alpha \in I} M \to \bigoplus_{\alpha \in I} N$ which simply applies f in every coordinate. But since f is injective, this homomorphism is clearly also injective. □

The fact that direct sums pull out of tensor products can also be used to prove that a more general class of modules is flat.

Corollary 15.41. *Suppose that F is a free right R-module and that $F \cong P \oplus Q$ for right R-submodules P and Q. Then P is a flat module.*

Proof. Let $f : M \to N$ be an injective homomorphism of left R-modules. We have seen that free modules are flat and hence

$$(P \oplus Q) \otimes_R M \xrightarrow{1 \otimes f} (P \oplus Q) \otimes_R N$$

is also injective.

Now we have an injection map $i : P \to P \otimes Q = F$ given by $i(p) = (p, 0)$, and a projection map $\pi : F = P \oplus Q \to P$ given by $\pi(p, q) = p$, where $\pi \circ i = 1_P$. Moreover, there is a commutative diagram

$$\begin{tikzcd}
P \otimes_R M \arrow[r, "1\otimes f"] \arrow[d, "i\otimes 1"] & P \otimes_R N \arrow[d, "i\otimes 1"] \\
(P \oplus Q) \otimes_R M \arrow[r, "1\otimes f"] & (P \oplus Q) \otimes_R N
\end{tikzcd}$$

Now if $x \in P \otimes_R M$ satisfies $[1 \otimes f](x) = 0$, then $[1 \otimes f] \circ [i \otimes 1](x) = 0$; by the flatness of $F = P \oplus Q$, we have $[i \otimes 1](x) = 0$. But since $\pi \circ i = 1_P$, $[\pi \otimes 1] \circ [i \otimes 1] = 1_{P \otimes_R M}$. It follows that $x = 0$. Hence $1 \otimes f : P \otimes_R M \to P \otimes_R N$ is injective and P is flat. □

We have proved that direct summands of free modules are flat. These modules have a name and another interesting description.

Definition 15.42. Let P be a right R-module. P is a *projective* module if given any surjective homomorphism of right modules $g : M \to N$ and a homomorphism of right modules $f : P \to N$, there is a homorphism $h : P \to M$ such that $g \circ h = f$.

This property can be respresented by the following commutative diagram:

$$\begin{tikzcd}
 & P \arrow[ld, dotted, "\exists h"'] \arrow[d, "f"] & \\
M \arrow[r, "g"] & N \arrow[r] & 0
\end{tikzcd}$$

The additional arrow to the right of N pointing to 0 is to remind one that in this property g is assumed to be surjective, that is, that the bottom row is exact at N. The property satisfied by P is not a universal property the way those are usually understood, because the map h is only assumed to exist, and need not be (in fact almost never is) unique.

One of the important properties of projective modules is that any surjection onto a projective module is split.

Lemma 15.43. *Let P be a projective right R-module. Suppose $g : N \to P$ is a surjective homomorphism of right R-modules. Then g is a split surjection.*

Recall that for g to be a split surjection means that there is a homomorphism $h : P \to N$ such that $g \circ h = 1_P$, and that by Lemma 12.49 this has as a consequence that N is an internal direct sum $N = \ker(g) \oplus \operatorname{im}(h) \cong \ker(g) \oplus P$.

Proof. Taking $f = 1_P : P \to P$ and $g : N \to P$ as given, the existence of $h : N \to P$ such that $g \circ h = 1_P$ is immediate from the definition of projective module. □

Let us now relate this definition of projective module to the modules that appeared in Corollary 15.41.

Theorem 15.44. *A right R-module P is projective if and only if there exists a right R-module Q such that $P \oplus Q$ is a free right R-module. In particular, free modules are projective.*

As usual, in the definition of projective and this theorem, there exist left-sided versions which are stated and proved in the analogous way. We have focused on right modules here only because our primary definition of flatness was for right modules, so we have stayed on that side for consistency.

Proof. First we claim that a free right R-module F is projective. Suppose that $f : F \to N$ and $g : M \to N$ are given right module homomorphisms with g surjective. Let $\{x_\alpha | \alpha \in I\}$ be a basis for F as a free right R-module. For each $\alpha \in I$ we can choose $m_\alpha \in M$ such that $g(m_\alpha) = f(x_\alpha)$, since g is surjective. We need to find h completing the diagram

$$\begin{tikzcd}
 & F \arrow[ld, dotted, "h"'] \arrow[d, "f"] \\
M \arrow[r, "g"] & N
\end{tikzcd}$$

By the universal property of a free module, there exists a unique homomorphism $h : F \to M$ such that $h(x_\alpha) = m_\alpha$ for all α. Then $f(x_\alpha) = [h \circ g](x_\alpha)$ for all α. This shows that the diagram commutes for the basis elements x_α. Since the basis generates F as an R-module and all maps are R-module homomorphisms, $h \circ g = f$ and the diagram commutes. Thus F is projective as claimed.

Now we need to extend this result to show that a direct summand of a free module F is projective. Suppose that we have an internal direct sum $F = P \oplus Q$. Let $i : P \to P \oplus Q = F$ and $\pi : F = P \oplus Q \to P$ be the injection and projection maps associated to the first coordinate of the direct sum, as in the proof of Corollary 15.41. Given f and g as above we have a diagram

$$\begin{tikzcd}
 & F = P \oplus Q \arrow[d, "\pi"] \arrow[ldd, dotted, "\widehat{h}"'] \\
 & P \arrow[d, "f"] \arrow[ld, dotted, "h"'] \\
M \arrow[r, "g"] & N
\end{tikzcd}$$

Now there exists a homomorphism $\widehat{h}$ making the outer triangle commute, i.e. $g \circ \widehat{h} = f \circ \pi$, since F is free. Define $h = \widehat{h} \circ i$. Then

$$g \circ h = g \circ \widehat{h} \circ i = f \circ \pi \circ i = f$$

because $\pi \circ i = 1_P$. This proves that P is projective.

For the converse, we need to prove that a projective module is necessarily a direct summand of a free module. Let P be projective and pick any generating set $\{p_\alpha | \alpha \in I\}$ whatsoever for P as a right R-module. By the universal property of a free module, there is a unique homomorphism $g : F = \bigoplus_{\alpha \in I} R \to P$ such that $g(e_\alpha) = p_\alpha$ for all α, where $\{e_\alpha\}$ is the standard basis of F. Since the $\{p_\alpha\}$ are a generating set for P, g is surjective. Now because P is projective, Lemma 15.43 shows that g is split. Thus $F \cong (\ker g) \oplus P$ by lemma 12.49. Taking $Q = \ker g$ we see that $P \oplus Q$ is free as required. □

Corollary 15.45. *Projective modules are flat.*

Proof. This follows from Theorem 15.44 and Corollary 15.41. □

Note that Corollary 12.50, which showed that a surjective homomorphism onto a free module splits, can now just be seen as a special case of Lemma 15.43 since free modules are projective. We needed the fact that surjections onto free modules are split for the theory of modules over a PID, but did not wish to introduce projective modules at that point.

Let us note that not all projective modules are free.

Example 15.46. Let $R = M_2(F)$ where F is a field. Then $R = I \oplus J$ as right R-modules, for right ideals $I = \left\{ \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} \middle| a, b \in F \right\}$ and $J = \left\{ \begin{pmatrix} 0 & 0 \\ c & d \end{pmatrix} \middle| c, d \in F \right\}$. Thus each of I and J is a summand of a free module of rank 1 and so I and J are projective right R-modules. Note however that every right R-module is also an F-vector space, since R is an F-algebra. Since $\dim_F R = 4$, any free R-module has F-dimension which a multiple of 4 (or infinite). Since $\dim_F I = 2$, I cannot be free. Of course J is also not free for the same reason, but in fact it is easy to see that $I \cong J$ as right R-modules.

Example 15.47. Let R be any commutative ring and let X be a multiplicative system in R. Then one can define the localization RX^{-1}. We concentrated on the case where R is an integral domain earlier, but you will study the general case in Math 200C. In that course you will also see that RX^{-1} is a flat R-module.

For example, if R is an integral domain with field of fractions K, then K is a flat R-module.

Example 15.48. $\mathbb{Q}$ is a flat $\mathbb{Z}$-module which is not projective. It is flat as remarked in the previous example. A $\mathbb{Z}$-module M is *divisible* if for all $x \in M$ and positive integer $n > 0$, there exists $y \in M$ such that $ny = x$. It is obvious that $\mathbb{Q}$ is a divisible $\mathbb{Z}$-module.

On the other hand, for a free $\mathbb{Z}$-module $F \cong \bigoplus \mathbb{Z}$, it is easy to see that no nonzero element of F can satisfy the divisibility property; thus F has no nonzero divisible submodules. In particular, $\mathbb{Q}$ cannot be isomorphic to a summand of a free module and hence it is not projective.

Example 15.49. If R is a PID, then it is true that all projective R-modules are free. This is easy to prove for finitely generated projective modules using the classification theorem. In general it follows from the fact that submodules of free modules over R are free (which we only proved for finitely generated modules).

16. Field basics

16.1. **Irreducible Polynomials.** Before we begin our study of fields in earnest we need to discuss some results about irreducible polynomials. These results could have been presented when we studied rings, but they also fit in here.

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 16.1. *Let $f \in F[x]$ where F is a field. Given $a \in F$, we have $f = q(x-a) + r$ where $q \in F[x]$ and $r = f(a) \in F$. In other words $f(a)$ is the remainder when f is divided by $(x-a)$. In particular, $f(a) = 0$ if and only if $(x-a)|f$ in $F[x]$.*

Proof. We know that $F[x]$ is a Euclidean domain with respect to the function $d : F[x] \to \mathbb{N}$ given by $d(0) = 0$, $d(f) = \deg(f)$ for $f \neq 0$. Since $g = (x-a)$ has degree 1, we have $f = qg + r$ with $d(r) < d(g) = 1$ or $r = 0$. Thus $d(r) = 0$ and hence r is a constant. Now since evaluation at a is a homomorphism, we must have $f(a) = r(a) = r$. The last statement follows since q and r are unique. □

The fact that the remainder when we divide f by $(x-a)$ is equal to $f(a)$ is often called the "remainder theorem", and the fact that $(x-a)|f$ if and only if $f(a) = 0$ is often called the "factor theorem". We say that $a \in F$ is a *root* of $f \in F[x]$ if $f(a) = 0$.

Corollary 16.2. *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. □

Corollary 16.3. *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 $f = (x - a)f'$ for some $f' \in F[x]$. Since $\deg f \geq 2$, $\deg f' \geq 1$. 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 $s, t \in R$ 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 16.4. *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 $s, 0 \neq t \in R$ 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 $t|a_m$. □

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 16.5. 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 16.6. If F is a finite field, for example $F = \mathbb{Z}/p\mathbb{Z}$ 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 $F = \mathbb{Z}/2\mathbb{Z} = \{0, 1\}$. 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 16.7 (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 10.5), 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 16.8. $f(x) = 5x^7 + 3x^6 - 9x^3 + 6$ 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 10.6).

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 16.9. Let $f = x + x^2y^{n-1} + y^n \in F[x, y] = (F[x])[y]$, where F is a field. We claim that f is an irreducible element in $F[x, y]$. To see this we embed $R = F[x]$ in its field of fractions $K = F(x)$, and consider $f \in K[y]$. Now we can consider f as a polynomial in y over the field $K = F(x)$. The element x is irreducible in $R = F[x]$. Writing $f = (1)y^n + (x^2)y^{n-1} + (x)y^0$ we see that x does not divide the leading coefficient in R, it divides the other coefficients, and x^2 does not divide the constant term. Thus Eisenstein's criterion applies and shows that f is an irreducible polynomial in $F(x)[y]$. Then f is also irreducible in $F[x][y] = F[x, y]$ by Corollary 10.6 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 16.10. 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 $\mathbb{Q}[x]$. □

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 16.11 (Reduction mod p)**.** *Let R be a UFD with field of fractions F. Let $f = a_nx^n + \cdots + a_1x + a_0 \in R[x]$. 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 16.7), 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 these to be equalities, this forces $\deg \overline{g} = \deg g \geq 1$, $\deg \overline{h} = \deg h \geq 1$. But then $\overline{f} = \overline{g}\overline{h}$ contradicts that $\overline{f}$ is irreducible in $\overline{R}[x]$. □

Example 16.12. 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 $(\mathbb{Z}/p\mathbb{Z})[x]$. Obviously $p = 2$ won't work as the constant term will die, so we try $p = 3$. Consider $\overline{f} = x^4 + x + 2 \in (\mathbb{Z}/3\mathbb{Z})[x]$. Clearly this polynomial has no root in $\mathbb{Z}/3\mathbb{Z} = \{0, 1, 2\}$. Following the method of Example 16.6, one may find all degree 2 irreducibles and show that $\overline{f}$ is not a product of 2 degree 2 irreducibles. Thus $\overline{f}$ is irreducible in $(\mathbb{Z}/3\mathbb{Z})[x]$ and hence f is irreducible in $\mathbb{Q}[x]$ by Proposition 16.11.

Remark 16.13. 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 $\overline{f} \in (\mathbb{Z}/p\mathbb{Z})[x]$ is always reducible. A simple example is $f(x) = x^4 + 1$.

16.2. **Field extensions.** Recall that a *field* F is a commutative ring such that every nonzero element is a unit. In our study of fields we will sometimes want to refer to auxiliary commutative rings which are not themselves fields; but we will not have any use for noncommutative rings in this section. Every ring should be assumed to be commutative unless told otherwise.

There is a short list of fields that arise naturally as fields of numbers appearing throughout mathematics, and which we have already encountered: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}/p\mathbb{Z}$ for a prime p. In this section we will write the field of integers mod p as $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, to emphasize that we are considering it as a field and not just a group. Actually, how "natural" the field of real numbers $\mathbb{R}$ is may be debatable, since to define it precisely involves a nontrivial limiting process of some kind. But we will take the existence of $\mathbb{R}$ and its basic properties as a given. It is certainly natural in the sense of its many applications to the physical sciences.

Besides the basic examples above we saw in the ring theory section two ring theoretic constructions which lead to many new examples of fields. Both will be of fundamental importance in our study of fields.

First, if R is any commutative ring whatsoever, and I is a maximal ideal of R, then the factor ring R/I is a field. This gives a way of producing potentially a number of different fields from a given commutative ring. For example, taking $R = \mathbb{Z}$ then the maximal ideals are those of the form $p\mathbb{Z}$ for primes p, and we recover all of the fields $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ through this construction. Even if we have a general commutative ring R about which we know nothing, we have seen as an application of Zorn's lemma that R must have some maximal ideal I; thus R has at least one factor ring R/I that is a field.

Second, if R is an integral domain, then we defined the field of fractions K of R to be the set of formal fractions $\{r/s | r \in R, 0 \neq s \in R\}$ under a natural equivalence relation that $r/s = t/u$ if $ru = st$, with addition and multiplication defined as usual for fractions. We always identify R with the subring $\{r/1 | r \in R\}$ of K, so that $R \subseteq K$. Thus for any integral domain we can always produce at least one field by taking the field of fractions. Of course $\mathbb{Q}$ can be produced in this way by taking the field of fractions of $\mathbb{Z}$. Recall also that if F is a field, the field of fractions of the polynomial ring $F[x]$ is called the *field of rational functions* $F(x)$. Its elements consists of formal ratios of polynomials.

Usually field theory does not study a field in isolation but rather its relation to other fields. The basic object of our study of fields will be the following.

Definition 16.14. A *field extension* or *extension of fields* is an inclusion of fields $F \subseteq K$; that is, K is a field and F is a subring of K which is also a field, which we also call a *subfield*.

K is always a left K-module by multiplication, so it is a left F-module by restriction. In other words, K is a vector space over F. As such it has a dimension and we define the *degree* of the field extension $F \subseteq K$ to be the number $[K : F] = \dim_F K$.

It is also common to use the notation K/F for a field extension $F \subseteq K$; K/F is read as "K over F" and is meant to emphasize that we are considering K in relation to the subfield F it lies over. The notation K/F is one whole unit and does not indicate any kind of quotient construction. The field F is also called the *base field*.

Example 16.15. $\mathbb{R} \subseteq \mathbb{C}$ is a field extension and $[\mathbb{C} : \mathbb{R}] = 2$ because by construction an $\mathbb{R}$-basis for $\mathbb{C}$ is given by $\{1, i\}$.

Example 16.16. In our study of ring theory we introduced for any square-free integer D the ring $\mathbb{Q}(\sqrt{D}) = \{a + b\sqrt{D} | a, b \in \mathbb{Q}\}$, as a subring of $\mathbb{C}$, and proved this ring is a field. Since D is squarefree $\sqrt{D} \notin \mathbb{Q}$, so $\{1, \sqrt{D}\}$ form a basis for $\mathbb{Q}(\sqrt{D})$ over $\mathbb{Q}$ and $[\mathbb{Q}(\sqrt{D}) : \mathbb{Q}] = 2$.

Actually both field extensions $\mathbb{R} \subseteq \mathbb{C}$ and $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})$ arise from a general ring-theoretic construction which will be of crucial importance from now on. There is nothing here we haven't already seen in our study of rings, but we remind the reader of the details since they are so important.

Lemma 16.17. *Let F be a field. Let $f \in F[x]$ be an irreducible polynomial.*

(1) *$F[x]/(f) = K$ is again a field.*

(2) *The function $\theta : F \to K = F[x]/(f)$ defined by $\theta(a) = a + (f)$ is an injective ring homomorphism.*

(3) *Identifying F with $\theta(F)$ we have a field extension $F \subseteq K$. As such, $[K : F] = \deg f$.*

Proof. (1) Since $F[x]$ is a PID, we know that an irreducible element generates a maximal ideal (f) by Lemma 9.7.

(2) This function is obviously a ring homomorphism by the definition of multiplication in $F[x]/(f)$. Since f is irreducible, by definition it is not a unit and so $\deg f \geq 1$. Since f has minimal degree among nonzero elements of (f), the ideal (f) contains no nonzero constant polynomials. Thus $\ker\theta = 0$ and θ is injective.

(3) Since θ is injective it gives an isomorphism from F to $\theta(F)$, so we can take this to be an identification, under which F now consists of the elements in $F[x]/(f)$ whose coset representatives are constant polynomials. Let $\deg f = n$. If $h \in F[x]$, then $h = qf + r$ under polyomial long division, where $\deg r < \deg f$. Hence $h + (f) = r + (f)$ with $\deg r \leq n-1$. Moreover, if $r' + (f) = r + (f)$, where $\deg r \leq n-1$ as well, then $r - r' \in (f)$ with $\deg r - r' \leq n-1$; since the smallest degree of nonzero elements of (f) is n, $r - r' = 0$ and $r = r'$. We see that every element of K is of the form $r + (f)$ for a *unique* polynomial r of degree at most $n-1$. It follows that $\{1 + (f), x + (f), \ldots, x^{n-1} + (f)\}$ is a basis for $F[x]/(f)$ as an F-vector space, and so $[K : F] = \deg f = n$. □

Whenever we are in the situation of the previous lemma, it is convenient to always identify F with its image under θ and consider K as an extension of F.

Example 16.18. Let us revisit Example 16.15. Usually the construction of $\mathbb{C}$ from $\mathbb{R}$ is done by defining $\mathbb{C}$ to a be an $\mathbb{R}$ vector space with basis $\{1, i\}$ made into a ring by defining the multiplication explicitly using $i^2 = 1$; technically, one needs to check this multiplication is associative and that the resulting ring is a field.

Lemma 16.17 gives a more abstract but cleaner way to go: It is immediate that $x^2 + 1$ is an irreducible polynomial in $\mathbb{R}[x]$, since it has no real roots. Thus $F = \mathbb{R}[x]/(x^2+1)$ is a field extension of $\mathbb{R}$ such that $[F : \mathbb{R}] = 2$. It is immediate that F contains an element $x + (x^2 + 1)$ whose square is equal to $-1 + (x^2 + 1)$, so abstractly we have found a field extension of $\mathbb{R}$ in which -1 has a square root. Of course, $F \cong \mathbb{C}$ as we already saw in Example 8.3; we have just defined the usual complex numbers in a different way.

Just as for groups and modules, the notion of a subfield generated by a subset will be important.

Definition 16.19. Let $F \subseteq K$ be a field extension. For any subset X of K, the *subfield of K generated by X over F* is the intersection of all subfields of K which contain both F and X. It is written as $F(X)$. When $X = \{\alpha_1, \dots, \alpha_n\}$ is finite we write this as $F(\alpha_1, \dots, \alpha_n)$. A field generated by one element over F, $F(\alpha)$, is called a *simple extension* of E.

It is easy to see that an arbitrary intersection of subfields is again a subfield, so that $F(X)$ is the unique smallest subfield of K which contains both F and X. Note that this notation only makes sense when working inside some larger field K that contains the elements in X. The field K is understood and not part of the notation.

Example 16.20. Let D be a squarefree integer. We have already defined a subfield of $\mathbb{C}$ called $\mathbb{Q}(\sqrt{D})$ in Example 16.16, where taking $\sqrt{D}$ to be either of the square roots of D in $\mathbb{C}$, we have $\mathbb{Q}(\sqrt{D}) = \mathbb{Q} + \mathbb{Q}\sqrt{D} \subseteq \mathbb{C}$. We have seen this is a field, and it obviously contains $\mathbb{Q}$ and $\sqrt{D}$. Conversely, any subfield of $\mathbb{C}$ which contains $\mathbb{Q}$ and $\sqrt{D}$ would contain $\mathbb{Q}+\mathbb{Q}\sqrt{D}$ just because it is closed under addition and multiplication. Thus $\mathbb{Q}(\sqrt{D})$ is indeed the subfield of $\mathbb{C}$ generated by $\sqrt{D}$ over $\mathbb{Q}$, so this notation we have used for it agrees with our new notation for subfield generation.

In our study of module theory over a commutative ring R, we saw that R-modules generated by a single element (cyclic modules) are the modules of the form R/I for an ideal I. Now we will see that field extensions generated by a single element, or simple extensions as we have named them, also have a very rigid structure.

Theorem 16.21. *Let $F \subseteq K$ be a field extension. Let $\alpha \in K$. There is a canonical homomorphism of rings*

$$\begin{aligned} \phi : F[x] &\to F(\alpha) \\ f(x) &\mapsto f(\alpha) \end{aligned}$$

and moreover exactly one of the following two cases occurs:

(i) $\ker \phi = (f)$ *for a unique, monic irreducible polynomial $f \in F[x]$. Moreover, ϕ is surjective, $F(\alpha) \cong F[x]/(f)$ as fields, and $[F(\alpha) : F] = \deg f < \infty$.*

(ii) $\ker \phi = 0$. *In this case ϕ extends to an isomorphism $F(x) \to F(\alpha)$, where $F(x)$ is the field of rational functions in one variable over F. Moreover $[F(\alpha) : F] = \infty$.*

Proof. The map ϕ is simply the evaluation at $\alpha \in K$, where $f(x) = \sum_{i=0}^{n} a_i x^i \in F[x]$ maps to $f(\alpha) = \sum_{i=0}^{n} a_i \alpha^i$ (with $\alpha^0 = 1$). We saw in our study of ring theory that such a map is a homomorphism of rings $F[x] \to K$. But every element of the image is of the form $\sum_{i=0}^{n} a_i \alpha^i$ with

$a_i \in F$, which is clearly contained in any subfield of K which contains F and α. So $\mathrm{im}(\phi) \subseteq F(\alpha)$ and we can think of ϕ as a map $F[x] \to F(\alpha)$.

Assume that $\ker \phi \neq 0$, so we are in case (i). Since $F[x]$ is a PID, $\ker \phi = (f)$ for some unique monic polynomial $f \in F[x]$. Now by the 1st isomorphism theorem for rings, $F[x]/(f) \cong \mathrm{im}(\phi)$. Since $\mathrm{im}(\phi)$ is a subring of K, it is a domain, so $F[x]/(f)$ is a domain. Thus (f) is a prime ideal, but in a PID, nonzero prime ideals are maximal. So $F[x]/(f)$ is a field and hence so is $\mathrm{im}(\phi)$. Now $\mathrm{im}(\phi)$ is a subfield of K which clearly contains F (the image of the constant polynomials) and α (the image of x). Thus $F(\alpha) \subseteq \mathrm{im}(\phi)$ by the definition of subfield generation. On the other hand, we already saw that $\mathrm{im}(\phi) \subseteq F(\alpha)$. Thus $\mathrm{im}(\phi) = F(\alpha)$ and we have an isomorphism $F[x]/(f) \to F(\alpha)$. In a PID maximal ideals are generated by irreducible polynomials, so f is irreducible. By Lemma 16.17, $[F(\alpha) : F] = \dim_F(F(\alpha)) = \dim_F(F[x]/(f)) = \deg f$ since ϕ is also an F-vector space isomorphism.

Otherwise, $\ker \phi = 0$ and we are in case (ii). Now the homomorphism of rings $\phi : F[x] \to F(\alpha)$ has the property that for every nonzero element $f \in F[x]$, $0 \neq \phi(f)$ is a unit, since $F(\alpha)$ is a field. By the universal property of the localization in Theorem 6.1, ϕ extends uniquely to a homomorphism $\widetilde{\phi} : F(x) \to F(\alpha)$, where $F(x)$ is the field of fractions of $F[x]$, in other words the localization of $F[x]$ at the set $X = F[x] - \{0\}$. The formula for $\widetilde{\phi}$ is $\widetilde{\phi}(f/g) = \phi(f)(\phi(g))^{-1} = f(\alpha)g(\alpha)^{-1}$ for all $0 \neq g, f \in F[x]$. Since $F(x)$ is a field, $\widetilde{\phi}$ must be injective and so $F(x) \cong \mathrm{im}(\widetilde{\phi})$. Thus $\mathrm{im}(\widetilde{\phi})$ is a subfield of K containing F and α and since $F(\alpha)$ is the unique smallest such, $F(\alpha) \subseteq \mathrm{im}(\widetilde{\phi})$. On the other hand, the formula for $\widetilde{\phi}$ above shows that $\mathrm{im}(\widetilde{\phi}) \subseteq F(\alpha)$. Thus $\mathrm{im}(\widetilde{\phi}) = F(\alpha)$ and in this case we have $F(x) \cong F(\alpha)$ as fields. Since $\dim_F F[x] = \infty$ already, certainly $\dim_F F(x) = \infty$. Thus $[F(\alpha) : F] = \dim_F F(\alpha) = \dim_F F(x) = \infty$. □

Definition 16.22. Let $F \subseteq K$ be a field extension, and let $\alpha \in K$. If case (i) occurs in Theorem 16.21 we say that α is *algebraic over* F; the monic irreducible polynomial f with $(f) = \ker(\phi)$ is called the *minimal polynomial* of α over F and is written as $\mathrm{minpoly}_F(\alpha)$. Otherwise, case (ii) occurs and we say that α is *transcendental over* F.

Note that α is algebraic over F precisely when there is some nonzero polynomial $g \in F[x]$ such that $g(\alpha) = 0$; the set of all such polynomials is then $\ker \phi = (f)$ in the notation of Theorem 16.21. Thus the minimal polynomial of α over F is the monic polynomial f of minimal possible degree such that $f(\alpha) = 0$. We also know that f is irreducible, and so clearly f is the unique monic irreducible polynomial such that $f(\alpha) = 0$ as well.

Example 16.23. Consider the extension $\mathbb{Q} \subseteq \mathbb{R}$. $f(x) = x^3 - 2$ is irreducible in $\mathbb{Q}[x]$ since it has no root in $\mathbb{Q}$, or by the Eisenstein criterion. Since the real cube root $\alpha = \sqrt[3]{2} \in \mathbb{R}$ satisfies $f(\alpha) = 0$, $f = \operatorname{minpoly}_{\mathbb{Q}}(\alpha)$. Thus α is algebraic over $\mathbb{Q}$ and $\mathbb{Q}(\alpha) \cong \mathbb{Q}[x]/(x^3 - 2)$. In particular $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$.

On the other hand, it is known that π is a transcendental number over $\mathbb{Q}$, though this is not particularly easy to prove. Thus $\mathbb{Q}(\pi) \cong \mathbb{Q}(x)$. In general *transcendence theory* refers to the study of which particular real or complex numbers are transcendental over $\mathbb{Q}$. This subject tends to involve sensitive results in analysis. As an example of the difficulty of this theory, the number e has also been proved to be transcendental over $\mathbb{Q}$, but it is unknown whether $e + \pi$ and $e\pi$ are transcendental.

We now know how to understand simple extensions $F(\alpha)$ of a field F, in terms of the properties of α. But in fact, an extension generated by a finite set of elements can be expressed as a series of simple extensions.

Lemma 16.24. *Let $F \subseteq K$ be a field extension. For any elements $\alpha_1, \dots, \alpha_n \in K$, we have $F(\alpha_1, \dots, \alpha_n) = F(\alpha_1)(\alpha_2)\dots(\alpha_n)$.*

Proof. Let us argue the case $n = 2$; the general case follows easily by induction. Note that when we write $F(\alpha_1)(\alpha_2)$, we mean $[F(\alpha_1)](\alpha_2)$; that is we are now applying the definition of generation to the field extension $F(\alpha_1) \subseteq K$ and the element α_2. In other words, $F(\alpha_1)(\alpha_2)$ is the unique smallest subfield of K which contains the field $F(\alpha_1)$ and the element α_2. On the other hand, the field $F(\alpha_1, \alpha_2)$ is the unique smallest subfield of K which contains F, α_1, and α_2.

The field $F(\alpha_1)(\alpha_2)$ contains $F(\alpha_1)$ and α_2, so it contains F, α_1, and α_2. Thus $F(\alpha_1, \alpha_2) \subseteq F(\alpha_1)(\alpha_2)$. Conversely, the field $F(\alpha_1, \alpha_2)$ contains F and α_1, so it contains $F(\alpha_1)$, the unique smallest such subfield. Thus $F(\alpha_1, \alpha_2)$ contains $F(\alpha_1)$ and α_2 and so $F(\alpha_1)(\alpha_2) \subseteq F(\alpha_1, \alpha_2)$. □

Note that when we write $F(\alpha_1, \dots, \alpha_n)$, the order in which we write the elements α_i is immaterial; we are taking the subfield generated by them as a set. On the other hand, when we treat this as $F(\alpha_1)(\alpha_2)\dots(\alpha_n)$ we have chosen a specific order, though the end result must be the same regardless. If we are doing calculations, one order might be easier to handle than another.

Example 16.25. Let $\mathbb{Q} \subseteq \mathbb{C}$ and consider $\mathbb{Q}(\sqrt{2}, i)$. Think of this as $\mathbb{Q}(\sqrt{2})(i)$. We know that $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$, since $\operatorname{minpoly}_{\mathbb{Q}}(\sqrt{2}) = x^2 - 2$ as $x^2 - 2$ does not have a root in $\mathbb{Q}$. Now i is a root of $x^2 + 1 \in \mathbb{Q}[x] \subseteq \mathbb{Q}(\sqrt{2})[x]$. So $\operatorname{minpoly}_{\mathbb{Q}(\sqrt{2})}(i)$ has degree at most 2. If it has degree 1, that means that $i \in \mathbb{Q}(\sqrt{2})$, but this is false since $\mathbb{Q}(\sqrt{2})$ consists of real numbers. Hence $\operatorname{minpoly}_{\mathbb{Q}(\sqrt{2})}(i) =$

x^2+1. We conclude that $\mathbb{Q}(\sqrt{2}, i) \cong \mathbb{Q}(\sqrt{2})[x]/(x^2+1)$ and $[\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}(\sqrt{2})] = 2$. By results we will see in the next section this means that $[\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}(\sqrt{2})][\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 4$.

It is also possible to analyze this extension by considering it as $\mathbb{Q}(i)(\sqrt{2})$ instead, but the other order is a bit easier because we can use the trick of considering real versus complex numbers. In this order we would have to determine whether or not $\sqrt{2} \in \mathbb{Q}(i)$; this is not difficult but it is a bit more work.

16.3. **Algebraic extensions.**

Definition 16.26. Let $F \subseteq K$ be a field extension. The extension is called *algebraic* if every $\alpha \in K$ is algebraic over F.

We will explore the properties of algebraic extensions in this section. Note that in an algebraic extension, for every $\alpha \in K$ we have $[F(\alpha) : F]$ is finite, equal to the degree of $\operatorname{minpoly}_F(\alpha)$. But these degrees can vary widely as we range over elements α, and $[K : F]$ could still be infinite overall, as we will see.

The key lemma for calculating the degree of an algebraic extension is the following.

Lemma 16.27. *Let $E \subseteq F \subseteq K$ be fields.*

(1) *If either $[K : F]$ or $[F : E]$ is infinite, then so is $[K : E]$.*
(2) *If $[K : F]$ and $[F : E]$ are finite, then so is $[K : E]$ and in fact $[K : E] = [K : F][F : E]$.*

Proof. (1) If $[F : E] = \infty$, in other words $\dim_E F = \infty$, then since the E-vector space K contains the E-subspace F, $\dim_E K = \infty$ also. If instead $[K : F] = \infty$, then a basis of K as an F-space is certainly also still a linearly independent set over E. Since any linearly independent set can be extended to a basis, $\dim_E K = \infty$.

(2) The proof will show more than the statement; we will see how to construct, given a basis of K as an F-space and a basis of F as an E-space, an explicit basis of K as an E-space.

Thus let $\{\alpha_1, \dots, \alpha_m\}$ be an E-basis for F and $\{\beta_1, \dots, \beta_n\}$ an F-basis for K. We claim that $S = \{\alpha_i\beta_j | 1 \leq i \leq m, 1 \leq j \leq n\}$ is an E-basis for K.

First, if $\gamma \in K$, then $\gamma = \sum_{j=1}^n a_j\beta_j$ for $a_j \in F$ since $\{\beta_j\}$ spans K as an F-space. But then each $a_j = \sum_{i=1}^m b_{ij}\alpha_i$ for some $b_{ij} \in E$ since the α_i space F as an E-space. We conclude that $\gamma = \sum_{i,j} b_{ij}\alpha_i\beta_j$ and so the set S spans K over E.

Next, suppose that $\sum_{i,j} b_{ij}\alpha_i\beta_j = 0$ for some $b_{ij} \in E$. Then $\sum_j(\sum_i b_{ij}\alpha_i)\beta_j = 0$ with $\sum_i b_{ij}\alpha_i \in F$, since $E \subseteq F$. Since the β_j are independent over F we get $\sum_i b_{ij}\alpha_i = 0$ for all i. But then since the α_i are independent over E we get $b_{ij} = 0$ for all j, for all i. This shows that S is independent over E. We have shown that S is a basis for K over E as claimed.

Now $[K : E] = |S| = nm = [K : F][F : E]$. □

Corollary 16.28. *If $E \subseteq F \subseteq K$ are fields with $[K : E] < \infty$, then $[F : E]$ and $[K : F]$ divide $[K : E]$.*

Corollary 16.29. *If $E \subseteq K$ is a field extension with prime degree $[K : E] = p$, then for any field F with $E \subseteq F \subseteq K$, either $E = F$ or $E = K$.*

Lemma 16.27 and its corollaries are reminiscent of Lagrange's theorem in finite group theory. This is not an accident; the Fundamental Theorem of Galois Theory which we prove later gives strong connections between fields and groups.

Many important examples involve considering extensions generated over $\mathbb{Q}$ inside the field extension $\mathbb{Q} \subseteq \mathbb{C}$. From now on when we write $\mathbb{Q}(\alpha_1, \ldots, \alpha_n)$ for certain complex numbers α_i, it should be assumed that we are taking this extension inside $\mathbb{C}$ even if that is not explicitly mentioned.

Lemma 16.27 is very useful for doing calculations of degrees of extensions. Here is one example.

Example 16.30. Consider $K = \mathbb{Q}(\sqrt[3]{2}, \sqrt{3})$ where $\sqrt[3]{2}$ is the real cube root of 2. We claim that $[K : \mathbb{Q}] = 6$. We know that $x^3 - 2$ and $x^2 - 3$ are irreducible over $\mathbb{Q}$, so $\text{minpoly}_{\mathbb{Q}}(\sqrt[3]{2}) = x^3 - 2$ and $\text{minpoly}_{\mathbb{Q}}(\sqrt{3}) = x^2 - 3$. Thus $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ and $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$. Since of course $x^3 - 2 \in \mathbb{Q}(\sqrt{3})[x]$ as well, $\text{minpoly}_{\mathbb{Q}(\sqrt{3})}(\sqrt[3]{2})$ has degree at most 3 and

$$[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}(\sqrt{3})] = [\mathbb{Q}(\sqrt{3})(\sqrt[3]{2}) : \mathbb{Q}(\sqrt{3})] \leq 3.$$

By Lemma 16.27, $[K : \mathbb{Q}] \leq 6$. On the other hand, since $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{3}) \subseteq K$ we have $[K : \mathbb{Q}]$ is divisible by $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$ (using Lemma 16.27 again), and since $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2}) \subseteq K$ it also follows that $[K : \mathbb{Q}]$ is divisible by $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$. The only possibility is $[K : \mathbb{Q}] = 6$.

As a consequence, we conclude that $[K : \mathbb{Q}(\sqrt{3})] = 3$. Since $K = \mathbb{Q}(\sqrt{3})(\sqrt[3]{2})$, this means that $\sqrt[3]{2}$ must have minimal polynomial $x^3 - 2$ over $\mathbb{Q}(\sqrt{3})$. It is not obvious that $x^3 - 2$ is irreducible over $\mathbb{Q}(\sqrt{3})$, or equivalently that it has no roots in this field, and this would be more awkward to prove directly. Similarly, we conclude that the minimal polynomial of $\sqrt{3}$ over $\mathbb{Q}(\sqrt[3]{2})$ is still $x^2 - 3$.

As well as being useful for calculations, Lemma 16.27 leads to a number of interesting general results about algebraic elements and extensions.

First we see that a finite degree extension is always algebraic.

Lemma 16.31. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$. Then K/F is an algebraic extension.*

Proof. Let $\alpha \in K$. We have $F \subseteq F(\alpha) \subseteq K$. Because $[K : F] < \infty$, we have $[F(\alpha) : F] < \infty$ by Lemma 16.27. But then we know that α is algebraic over F by Theorem 16.21. Since $\alpha \in K$ was arbitrary, K/F is algebraic. □

Proposition 16.32. *Let $F \subseteq K$ be a field extension and suppose that $\alpha, \beta \in K$ are both algebraic over F. Then $\alpha - \beta$, $\alpha\beta$, and α^{-1} (if $\alpha \neq 0$) are also algebraic over F.*

Proof. By Theorem 16.21, since α is algebraic over F we have $[F(\alpha) : F] < \infty$. Since β is algebraic over F, it is certainly algebraic over $F(\alpha)$; if $f(\alpha) = 0$ with $0 \neq f \in F[x]$ then just consider $f \in F(\alpha)[x]$. Thus $[F(\alpha)(\beta) : F(\alpha)] < \infty$. By Lemma 16.27, $[F(\alpha, \beta) : F] < \infty$. Thus $F(\alpha, \beta)/F$ is an algebraic extension, by Lemma 16.31.

Now notice that since $F(\alpha, \beta)$ is a subfield of K containing α and β, it certainly also contains $\alpha - \beta$, $\alpha\beta$, and α^{-1} (if $\alpha \neq 0$). Hence these elements are all algebraic over F. □

Corollary 16.33. *Let $F \subseteq K$ be a field extension. Define $L = \{\alpha \in K | \alpha \text{ is algebraic over } F\}$. Then L is a subfield of K containing F and L/F is algebraic.*

Proof. Proposition 16.32 shows that L is closed under difference, product, and inverses, which implies that L is a subfield of K. It is obvious that L/F is algebraic by definition. □

Definition 16.34. The subset

$$\overline{\mathbb{Q}} = \{\alpha \in \mathbb{C} | \alpha \text{ is algebraic over } \mathbb{Q}\}$$

is called the *field of algebraic numbers.* It is also called the *algebraic closure* of $\mathbb{Q}$.

Note that $\overline{\mathbb{Q}}$ really is a subfield of $\mathbb{C}$ by Corollary 16.33. For a fixed prime number p, and each $n \geq 2$, note that $x^n - p$ is irreducible over $\mathbb{Q}$ by the Eisenstein criterion. Thus $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[n]{p}) \subseteq \overline{\mathbb{Q}}$ with $[\mathbb{Q}(\sqrt[n]{p}) : \mathbb{Q}] = n$. It follows that $[\overline{\mathbb{Q}} : \mathbb{Q}] = \infty$. Thus $\overline{\mathbb{Q}}/\mathbb{Q}$ is an example of an infinite degree algebraic extension.

Proposition 16.32 showed that the algebraic numbers over $\mathbb{Q}$ are closed under the field operations. The proof is abstract, however, and does not show how one can find a polynomial that has a given difference, product, or inverse of algebraic numbers as a root.

Example 16.35. Since $\sqrt{2}$ and $\sqrt{3}$ are both algebraic over $\mathbb{Q}$, with minimal polynomials $x^2 - 2$ and $x^2 - 3$ respectively, we know that $\alpha = \sqrt{2} + \sqrt{3}$ is algebraic over $\mathbb{Q}$. Here is a way to find a polynomial in $\mathbb{Q}$ with α as a root (this method is rather special to the case of a sum of two square roots, though).

Note that $\alpha^2 = (\sqrt{2} + \sqrt{3})^2 = 2 + 2\sqrt{6} + 3 = 5 + 2\sqrt{6}$. Thus $2\sqrt{6} = \alpha^2 - 5$. Squaring both sides, $24 = (\alpha^2 - 5)^2$. It follows that $f(x) = (x^2 - 5)^2 - 24 = x^4 - 10x^2 + 1$ has α as a root. In fact one may show that $f = \mathrm{minpoly}_{\mathbb{Q}}(\alpha)$.

The following definition is similar to what we have defined for generation in other contexts like groups and modules.

Definition 16.36. Let $F \subseteq K$ be a field extension. We say that the extension K/F is *finitely generated* if $K = F(\alpha_1, \alpha_2, \dots, \alpha_n)$ for some $\alpha_1, \alpha_2, \dots, \alpha_n \in K$.

Lemma 16.37. *Let $F \subseteq K$ be a field extension. Then $[K : F] < \infty$ if and only if K/F is finitely generated and algebraic.*

Proof. If $[K : F] < \infty$ then we have seen that K/F is algebraic in Lemma 16.31. It is also easy to see that K/F is finitely generated; for example, if $\alpha_1, \dots, \alpha_n$ is an F-basis of K then certainly $K = F(\alpha_1, \dots, \alpha_n)$ (though likely fewer than n elements suffice).

Conversely, suppose that $K = F(\alpha_1, \alpha_2, \dots, \alpha_n)$ for some elements $\alpha_i \in K$, and that K/F is algebraic. Thus each α_i is algebraic over F. Define $d_i = [F(\alpha_1, \dots, \alpha_i) : F(\alpha_1, \dots, \alpha_{i-1})]$ for each $1 \leq i \leq n$. Since $F(\alpha_1 \dots, \alpha_i) = F(\alpha_1, \dots, \alpha_{i-1})(\alpha_i)$, we see that $d_i = \deg \mathrm{minpoly}_{F(\alpha_1, \dots, \alpha_{i-1})}(\alpha_i)$. If $e_i = \deg \mathrm{minpoly}_F(\alpha_i)$, then $d_i \leq e_i$ for all i since any polynomial in $F[x]$ with α_i as a root is also a polynomial in $F(\alpha_1, \dots, \alpha_{i-1})[x]$. Now

$$\begin{aligned}[K : F] &= [F(\alpha_1, \dots, \alpha_n) : F(\alpha_1, \dots, \alpha_{n-1})][F(\alpha_1, \dots, \alpha_{n-1}) : F(\alpha_1, \dots, \alpha_{n-2}] \dots [F(\alpha_1) : F] \\ &= d_n d_{n-1} \dots d_1 \leq e_n e_{n-1} \dots e_1 < \infty,\end{aligned}$$

by repeated use of Lemma 16.27. In particular $[K : F] < \infty$. □

Examining the proof of the lemma, we immediately have the following consquence.

Corollary 16.38. *If $K = F(\alpha_1, \dots, \alpha_n)$ where each α_i is algebraic over F with $e_i = \deg \mathrm{minpoly}_F(\alpha_i) = [F(\alpha_i) : F]$, then $[K : F] \leq e_1 e_2 \dots e_n$.*

Example 16.30, where $K = \mathbb{Q}(\sqrt[3]{2}, \sqrt{3})$, is an example where the upper bound given by the corollary is actually acheived. In general it is just an upper bound, as can be seen from silly examples like $K = \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{6})$, where $\sqrt{6} \in \mathbb{Q}(\sqrt{2}, \sqrt{3})$ already, so in fact $[K : \mathbb{Q}] \leq 4$, while applying the corollary with the 3 generators blindly gives $[K : \mathbb{Q}] \leq 8$.

We may now show that an algebraic extension of an algebraic extension is algebraic.

Theorem 16.39. *Let $E \subseteq F \subseteq K$ where both F/E and K/F are algebraic extensions. Then K/E is algebraic.*

Proof. Let $\alpha \in K$. We know that α is algebraic over F. Let $f = \text{minpoly}_F(\alpha) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in F[x]$. Now each coefficient $a_i \in F$ is algebraic over E. By Lemma 16.37, $[E(a_0, \ldots, a_{n-1}) : E] < \infty$. Now note that $f \in E(a_0, \ldots, a_{n-1})[x]$. This means that α is algebraic over the field $E(a_0, \ldots, a_{n-1})$. Hence $[E(a_0, \ldots, a_{n-1}, \alpha) : E(a_0, \ldots, a_{n-1})] < \infty$. Now applying Lemma 16.27 gives $[E(a_0, \ldots, a_{n-1}, \alpha) : E] < \infty$. In particular, since α belongs to the field $E(a_0, \ldots, a_{n-1}, \alpha)$, α is algebraic over E by Lemma 16.31. □

Example 16.40. Suppose that $F \subseteq K$ is a field extension. We say that F is *algebraically closed in* K if given any chain of subfields $F \subseteq L \subseteq K$ with L/F algebraic, then $L = F$. In other words, no subfield of K properly containing F is algebraic over F.

Now given an arbitrary extension $E \subseteq K$, we can define $F = \{\alpha \in K | \alpha \text{ is algebraic over } E\}$. As we have seen, $E \subseteq F \subseteq K$ with F a subfield of K algebraic over E. Now we can see that F is algebraic closed in K. For if $F \subseteq L \subseteq K$ with L/F algebraic, then L/E is algebraic by Theorem 16.39. Thus every element of L is algebraic over E, which means $L \subseteq F$ and hence $L = F$.

In particular, considering $\mathbb{Q} \subseteq \overline{\mathbb{Q}} \subseteq \mathbb{C}$, we see that $\overline{\mathbb{Q}}$ is algebraically closed in $\mathbb{C}$.

16.4. **Splitting fields.** Let F be a field and let $f(x) \in F[x]$ be an irreducible polynomials. Recall that if $\deg f \geq 2$, then f has no roots in F (Corollary 16.3). Of course, if we have a field extension $F \subseteq K$, then f might well have a root in K.

It is natural to wonder if we just start with a field F and have no prior knowledge about any extension fields, does there always exist a field extension $F \subseteq K$ such that the irreducible polynomial $f \in F[x]$ has a root in K? The next lemma answers this question. While the proof of the lemma is easy, the idea behind it is rather subtle.

Lemma 16.41. *Let $f \in F[x]$ be irreducible. Then $K = F[x]/(f)$ is a field, and identifying F with a subfield of K as usual, then f has a root α in K.*

Proof. We already know that K is a field and we saw that the natural map $\theta : F \to F[x]/(f)$ is an injective homomorphism, allowing us to identify F with the cosets of constant polynomials in $K = F[x]/(f)$. This was Lemma 16.17.

But now observe that for $\alpha = x + (f) \in K$, evaluating $f = \sum_{i=0}^{n} a_i x^i$ at α gives

$$\sum_{i=0}^{n}(a_i + (f))(x + (f))^i = \sum_{i=0}^{n}(a_i x^i + (f)) = (\sum_{i=0}^{n} a_i x^i) + (f) = f + (f) = 0 + (f) = 0.$$

Thus f has a root in K as we wished. □

Recall that a polynomial $f \in F[x]$ *splits* (over F) if f factors as a product of degree 1 irreducibles. Since every degree 1 irreducible is of the form $(x - a)$ up to associates, this is the same as saying that we can write $f = c(x - \alpha) \dots (x - \alpha_n)$ for some $c, \alpha_1, \dots, \alpha_n \in F$. A field F is *algebraically closed* if every polynomial in $F[x]$ splits over F, or equivalently if every irreducible polynomial in $F[x]$ is of degree 1. We will freely use in examples that $\mathbb{C}$ is algebraically closed, though we don't prove this until much later.

Definition 16.42. Let $F \subseteq K$ be a field extension and let $F \in F[x]$. We say that K is a *splitting field for f over F* if

(i) $f = c(x - \alpha_1) \dots (x - \alpha_n)$ in $K[x]$, i.e. f splits over K.

(ii) $K = F(\alpha_1, \dots, \alpha_n)$.

Roughly, this definition is saying that a splitting field K is a larger field in which f splits, but where K is no larger than necessary for this to happen. Certainly any field over which f splits must contain the field generated over F by the roots of f.

We have seen that we can always find a larger field in which any given irreducible polynomial has a root. We can now extend this to see that any polynomial has a splitting field.

Lemma 16.43. *Let F be a field and let $f \in F[x]$. Then there exists a field extension $F \subseteq K$ such that K is a splitting field for f over F.*

Proof. First we prove that there exists a field extension $F \subseteq L$ such that f splits in $L[x]$. We induct on $\deg f$. If f already splits in $F[x]$, take $L = F$. Otherwise, a factorization of f into a product of irreducibles in $F[x]$ contains at least one irreducible factor, say g, with $\deg g \geq 2$. By Lemma 16.41, there is an extension $F \subseteq L'$ such that g has a root α_1 in L'. Then by the factor theorem, in $L'[x]$ the polynomial f factors as $f = (x - \alpha_1)f'$ for $f' \in L'[x]$. Since $\deg f' < \deg f$, by the induction hypothesis there is an extension $L' \subseteq L$ such that f' splits in $L[x]$, say $f' = c(x - \alpha_2) \dots (x - \alpha_n)$. Then $f = c(x - \alpha_1)(x - \alpha_2) \dots (x - \alpha_n)$ in $L[x]$, so f splits in $L[x]$.

Finally, define $K = F(\alpha_1, \dots, \alpha_n) \subseteq L$. It is clear that K is a splitting field for f over F. □

As we saw in the proof above, to find a splitting field K for a polynomial $f \in F[x]$, it suffices to find a field extension $F \subseteq L$ such that f splits in $L[x]$ with $f = c(x - \alpha_1) \dots (x - \alpha_n)$ and then let

$K = F(\alpha_1, \dots, \alpha_n) \subseteq L$. In particular, since we know any polynomial in $\mathbb{C}[x]$ splits, if $F \subseteq \mathbb{C}$ we can find a splitting field for $f \in F[x]$ by finding the roots of f in $\mathbb{C}$ and adjoining them to F inside $\mathbb{C}$.

Example 16.44. Let $f = x^n - 1 \in \mathbb{Q}[x]$ for some $n \geq 1$. We know that $\mathbb{C}$ has n distinct nth roots of 1, namely $\alpha_j = e^{2j\pi i/n}$ for $1 \leq j \leq n$. Thus $x^n - 1 = (x - \alpha_1)\dots(x - \alpha_n)$ since both are monic and have the same distinct n roots in $\mathbb{C}$. So $K = \mathbb{Q}(\alpha_1, \dots, \alpha_n)$ is a splitting field for $x^n - 1$ over $\mathbb{Q}$. Now note that setting $\zeta = \alpha_1 = e^{2\pi i/n}$, we have $\alpha_i = \zeta^i$ for all i. It follows that $K = \mathbb{Q}(\zeta, \zeta^2, \dots, \zeta^n) = \mathbb{Q}(\zeta)$ and so K is a simple extension of $\mathbb{Q}$. It is called the *nth cyclotomic field*.

Consider the special case where $n = p$ is prime. In this case $x^p - 1 = (x - 1)g(x)$ where $g(x) = x^{p-1} + \dots + x + 1$ was shown to be irreducible over $\mathbb{Q}$ in Example 16.10. Clearly $\zeta = e^{2\pi i/p}$ is a root of g and therefore $g = \text{minpoly}_{\mathbb{Q}}(\zeta)$. Hence $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$ in this case.

We will calculate the degree $[\mathbb{Q}(\zeta) : \mathbb{Q}]$ for arbitrary n later.

Example 16.45. Generalizing the previous example, let us consider the splitting field of $f = x^n - a \in \mathbb{Q}[x]$ over $\mathbb{Q}$ for some $n \geq 1$ and $a \neq 0$. Let α be any nth root of a in $\mathbb{C}$. As in the previous example, let $\zeta = e^{2\pi i/n}$ so that $\{1, \zeta, \dots, \zeta^{n-1}\}$ is the set of distinct complex nth roots of 1. Then the set $\{\alpha, \alpha\zeta, \alpha\zeta^2, \dots, \alpha\zeta^{n-1}\}$ consists of n distinct complex numbers and they are all roots of $x^n - a$; thus $x^n - a = (x - \alpha)(x - \alpha\zeta)\dots(x - \alpha\zeta^{n-1})$ in $\mathbb{C}[x]$. Now a splitting field for f over $\mathbb{Q}$ can be constructed inside $\mathbb{C}$ as $\mathbb{Q}(\alpha, \alpha\zeta, \dots \alpha\zeta^{n-1}) = \mathbb{Q}(\alpha, \zeta)$.

Let us consider the special case where $f = x^p - q$ for prime numbers p, q (not necessarily distinct). The number q has a unique positive pth root $\alpha = \sqrt[p]{q} \in \mathbb{R}$. By the Eisenstein criterion applied to the prime q, f is irreducible over $\mathbb{Q}$. Thus $f = \text{minpoly}_{\mathbb{Q}}(\alpha)$ and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = p$. We have seen that $\text{minpoly}_{\mathbb{Q}}(\zeta) = x^{p-1} + \dots + x + 1$ in this case, so $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$. By Corollary 16.38, $d = [\mathbb{Q}(\alpha, \zeta) : \mathbb{Q}] \leq p(p-1)$. On the other hand, $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$ and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = p$ are both divisors of d by Lemma 16.27. Since $\gcd(p, p-1) = 1$, we see that this forces $d = p(p-1)$. It also means that $[\mathbb{Q}(\alpha, \zeta) : \mathbb{Q}(\alpha)] = p - 1$ and thus $\deg \text{minpoly}_{\mathbb{Q}(\alpha)}(\zeta) = p - 1$; this forces $\text{minpoly}_{\mathbb{Q}(\alpha)}(\zeta) = x^{p-1} + \dots + x + 1$. It would be rather awkward to prove that this polynomial is irreducible over $\mathbb{Q}(\alpha)$ directly. Similarly, $[\mathbb{Q}(\alpha, \zeta) : \mathbb{Q}(\zeta)] = p$ and $\text{minpoly}_{\mathbb{Q}(\zeta)}(\alpha) = x^p - q$.

It turns out that any two splitting fields of a polynomial $f \in F[x]$ are isomorphic as fields. This is perhaps not too surprising, since by definition, in some sense the splitting field is a “smallest” field extension over which f splits, and that ought to be determined by f. However, along the way we will actually show something stronger: if f is irreducible in $F[x]$, then given two splitting fields

K and K', we can choose an isomorphism between them which sends a specified root of f in K to any root of f in K' we please. Applying this to a single splitting field will give us a way to produce automorphisms of a field which move the roots of f around. This will lay the foundation for our study of Galois theory later.

Before attacking the general case of splitting fields, we first example what happens when we adjoin to a field F two possibly different roots of the same irreducible polynomial $f \in F[x]$. For technical reasons, it is necessary to state this result in a generality that at first seems rather awkward: instead of fixing a base field, we work over two different base fields with an isomorphism between them. It doesn't really make the proof harder, and its utility will be seen in the proof of the proposition to follow.

Lemma 16.46. *Let $\phi : F \to F'$ be an isomorphism of fields. This induces an isomorphism of polynomial rings $\phi : F[x] \to F'[x]$ we give the same name, by applying ϕ to the coefficients. Let $f \in F[x]$ be an irreducible polynomial (over F) and let $f' = \phi(f) \in F'[x]$. Suppose that $F \subseteq K$ and $F' \subseteq K'$ are field extensions such that $\alpha \in K$ is a root of f and $\alpha' \in K'$ is a root of f'. Then considering $F \subseteq F(\alpha) \subseteq K$ and $F' \subseteq F'(\alpha') \subseteq K'$, there is an isomorphism $\theta : F(\alpha) \to F'(\alpha')$ such that $\theta(\alpha) = \alpha'$ and $\theta|_F = \phi$.*

Proof. Since f is irreducible over F, $f = \mathrm{minpoly}_F(\alpha)$. Since $\phi : F[x] \to F'[x]$ is an isomorphism of rings, $f' = \phi(f)$ is irreducible over F', and so $f' = \mathrm{minpoly}_{F'}(\alpha')$. Now by our theorem on the structure of a simple extension, there are isomorphisms $\sigma_1, \sigma_2, \sigma_3$ forming a chain

$$F(\alpha) \xrightarrow{\sigma_1} F[x]/(f) \xrightarrow{\sigma_2} F'[x]/(f') \xrightarrow{\sigma_3} F'(\alpha')$$

Where $\sigma_1^{-1} : F[x]/(f) \to F(\alpha)$, $\sigma_3 : F'[x]/(f') \to F'(\alpha')$ are the isomorphisms coming from Theorem 16.21(i), and σ_2 is induced by the isomorphism $\phi : F[x] \to F'[x]$ and the fact that $\phi(f) = f'$. Now take $\theta = \sigma_3\sigma_2\sigma_1$. Then

$$\theta(\alpha) = [\sigma_3\sigma_2\sigma_1](\alpha) = [\sigma_3\sigma_2](x + (f)) = \sigma_3(x + (f')) = \alpha',$$

and $\theta|_F = \phi$ since $\sigma_2|_F = \phi$ while $\sigma_1|_F = 1_F$ and $\sigma_3|_{F'} = 1_{F'}$. □

Corollary 16.47. *Let $F \subseteq K$ be a field extension, and let $f \in F[x]$ be irreducible over F. Suppose that $\alpha_1, \alpha_2 \in K$ are roots of f. Then there is an isomorphism $\sigma : F(\alpha_1) \to F(\alpha_2)$ such that $\sigma(\alpha_1) = \alpha_2$ and $\sigma|_F = 1_F$.*

Proof. This is immediate by applying the lemma to $F = F'$, $\phi = 1_F$, $K' = K$. □

We see thus that *adding two roots of the same irreducible polynomial give isomorphic simple extensions.* All roots of an irreducible polynomial are "equal" in this sense.

Example 16.48. Let $\zeta = e^{2\pi i/3}$ be a primitive 3rd root of 1. Consider the splitting field K of $f(x) = x^3 - 2$ over $\mathbb{Q}$. As we saw in Example 16.45, f is irreducible over $\mathbb{Q}$. If $\alpha = \sqrt[3]{2}$ is the positive real cube root of 2, then the roots of f in $\mathbb{C}$ are $\{\sqrt[3]{2}, \sqrt[3]{2}\zeta, \sqrt[3]{2}\zeta^2\}$. By Corollary 16.47, there is an isomorphism $\phi : \mathbb{Q}(\sqrt[3]{2}) \to \mathbb{Q}(\sqrt[3]{2}\zeta)$ fixing $\mathbb{Q}$ and sending $\sqrt[3]{2}$ to $\sqrt[3]{2}\zeta$. It doesn't matter, for example, that one of these fields is contained in $\mathbb{R}$ and the other isn't.

Now we are ready to prove the uniqueness up to isomorphism of splitting fields.

Proposition 16.49. *Let $\phi : F \to F'$ be an isomorphism of fields, inducing the isomorphism of rings $\phi : F[x] \to F'[x]$. Suppose K is a splitting field of $f \in F[x]$ over F, and K' is a splitting field of $\phi(f) \in F'[x]$ over F'.*

(1) *There is an isomorphism $\sigma : K \to K'$ such that $\sigma|_F = \phi$.*

(2) *If $g \in F[x]$ is any irreducible factor of f in $F[x]$, then for any $\alpha \in K$ which is a root of g, and any $\alpha' \in K'$ which is a root of $\phi(g)$, we can choose a σ in (1) with $\sigma(\alpha) = \alpha'$.*

Proof. Suppose that we have proved (1) and that g is an irreducible factor of degree 1 in part (2). It is no loss of generality to assume that g is monic, so $g = x - \alpha \in F[x]$. Then $\phi(g) = x - \phi(\alpha) \in F'[x]$. So the only root of g is α and the only root of $\phi(g)$ is $\phi(\alpha)$, and since $\sigma|_F = \phi$ we certainly have $\sigma(\alpha) = \phi(\alpha)$. So part (2) is automatic for irreducibles of degree 1.

The proof in general is by induction on degree f. If f splits over F already, then $K = F$. This also means that f' splits over F', so $K' = F'$; thus we take $\sigma = \phi$ in part (1). In this case all irreducible factors of f have degree 1 so (2) is also clear by the remark above. In particular, if $\deg f \leq 1$ then f splits, so the base case holds.

Now we assume that f has some irreducible factor $g \in F[x]$ with $\deg g \geq 2$, and that the result is true for all polynomials of degree smaller than f. Let $g' = \phi(g)$. Then g' is irreducible in $F'[x]$. Since f and hence g splits in K, we can pick a root $\alpha \in K$ of g. Similarly we pick a root $\alpha' \in K'$ of g'.

By Lemma 16.46, there is an isomorphism $\theta : F(\alpha) \to F'(\alpha')$ such that $\theta(\alpha) = \alpha'$ and $\theta|_F = \phi$. Now the key is to treat θ as the new isomorphism of base fields. Since α is a root of g and hence of f, we have $f = (x - \alpha)h$ for some $h \in F(\alpha)[x]$. Since f splits over K, all of the roots of f in K other than α, say $\beta_1, \dots, \beta_m$, are roots of h. It follows that $K = F(\alpha)(\beta_1, \dots, \beta_m)$ is a splitting field of $h \in F(\alpha)[x]$ over $F(\alpha)$. As usual we extend θ to an isomorphism of polynomial rings

$\theta : F(\alpha)[x] \to F'(\alpha')[x]$, and as such $\theta(f) = (x-\theta(\alpha))\theta(h) = (x-\alpha')h'$ where $h' = \theta(h) \in F'(\alpha')[x]$. Thus we similarly conclude that K' is a splitting field of h' over $F'(\alpha')$.

We thus have an isomorphism $\theta : F(\alpha) \to F'(\alpha')$ and splitting fields K of $h \in F(\alpha)[x]$ and K' of $\theta(h) = h' \in F'(\alpha')[x]$. Since $\deg h < \deg f$, by the induction hypothesis there is an isomorphism $\sigma : K \to K'$ such that $\sigma|_{F(\alpha)} = \theta$ (in particular, $\sigma(\alpha) = \theta(\alpha) = \alpha'$). We also have $\sigma|_F = \theta|_F = \phi$.

This proves part (1). We have proved part (2) along the way, since we saw that (2) is trivial if f splits over F, and otherwise to do the proof we chose an arbitrary irreducible factor g of f with $\deg g \geq 2$ and constructed a σ with $\sigma(\alpha) = \alpha'$, where α and α' were arbitrary roots of g and $\phi(g) = g'$, respectively. □

It should be clear now why Lemma 16.46 and Proposition 16.49 were stated using an isomorphism of base fields $\phi : F \to F'$ rather than a fixed base field. In the proof of the induction step in Proposition 16.49, even if we had started with $F = F'$ and $\phi = 1_F$ at the beginning, we would be forced to consider an isomorphism of base fields $\theta : F(\alpha) \to F(\alpha')$ at the induction step, where these fields are isomorphic but different in general. To make the induction work it is necessary for the statement to be in terms of an isomorphism of base fields from the start.

Nonetheless, we usually apply Proposition 16.49 in the case where $F = F'$ and $\phi = 1_F$. In this case, it tells us that if $F \subseteq K$ and $F \subseteq K'$ are both splitting fields of $f \in F[x]$, then there is an isomorphism $\sigma : K \to K'$ with $\sigma|_F = 1_F$. This tells us that splitting fields are unique up to isomorphism, as we claimed earlier. The proposition also tells us how to construct automorphisms of a splitting field which move roots around; this special case is worth singling out:

Corollary 16.50. *Let f be a polynomial in $F[x]$ and let K be a splitting field for f over F. If g is an irreducible factor of f in $F[x]$ and $\alpha, \alpha' \in K$ are both roots of g, then there exists an automorphism $\sigma : K \to K$ such that $\sigma(\alpha) = \alpha'$.*

Proof. Just take $F = F'$, $\phi = 1_F$, and $K = K'$ in Proposition 16.49. □

Example 16.51. Let us revisit Example 16.48. A splitting field of $x^3 - 2$ over $\mathbb{Q}$ is $K = \mathbb{Q}(\alpha, \zeta)$ where $\alpha = \sqrt[3]{2}$, $\zeta = e^{2\pi i/3}$, and $[K : \mathbb{Q}] = 6$. Let us construct 6 different automorphisms of K.

Since $x^3 - 2$ is irreducible, by Corollary 16.50 we can find automorphisms σ, τ of K such that $\sigma(\alpha) = \alpha\zeta$ and $\tau(\alpha) = \alpha\zeta^2$. We also saw in Example 16.45 that $g = x^2 + x + 1 = \text{minpoly}_{\mathbb{Q}}(\zeta)$ remains irreducible over $\mathbb{Q}(\alpha)$. Clearly K is the splitting field of g over $\mathbb{Q}(\alpha)$. Thus by the Corollary again, there is an automorphism ρ of K such that $\rho|_{\mathbb{Q}(\alpha)} = 1_{\mathbb{Q}(\alpha)}$ but $\rho(\zeta) = \zeta^2$.

Now it is easy to check that the 6 automorphisms $\{1_K, \sigma, \tau, \rho, \sigma\rho, \tau\rho\}$ of K are all different, as no two act the same way on both α and ζ.

16.5. **Separability.** Suppose that $f \in F[x]$ is a monic polynomial over a field F and that $F \subseteq K$ is a field extension such that f splits in $K[x]$, say $f = (x - \alpha_1) \dots (x - \alpha_n) \in K[x]$. Could it be that some of the α_i are equal? Of course there are easy ways to make this happen; for example we could have $f = (x - \alpha)^2$ with $\alpha \in F$ already; or slightly less trivially, $f = g^2$ for some irreducible polynomial $g \in F[x]$ which splits over K, so that each root appears twice when we factor f into linear factors over K. An example of the latter phenomenon would be $f = (x^2 + 1)^2 \in \mathbb{Q}[x]$.

What is less obvious is whether there could be an *irreducible* polynomial $f \in F[x]$ where some of the α_i are equal in the factorization of f over K. In fact this does happen, but only for special kinds of fields and field extensions which are quite different from the examples given so far. The goal of this section is to study this phenomenon, and also show that it is something that doesn't happen in many of the most common situations.

Definition 16.52. A polynomial $f \in F[x]$ is called *separable* if given a splitting field K for f over F, f factors as $f = c(x - \alpha_1) \dots (x - \alpha_n) \in K[x]$ with $\alpha_1, \alpha_2, \dots, \alpha_n$ distinct elements of K. Otherwise we say that the polynomial f is *inseparable*.

We have seen that if $F \subseteq K$ and $F \subseteq K'$ are both splitting fields for $f \in F[x]$, then there is an isomorphism $\sigma : K \to K'$ such that $\sigma|_F = 1_F$. Using this it is easy to see that the definition above is independent of the choice of splitting field; if f splits with distinct roots in one splitting field, the same will be true in any other. Note that f is separable if and only if f has $\deg f$ distinct roots in a splitting field K.

Example 16.53. $(x^2 + 1)^2 \in \mathbb{Q}[x]$ is an inseparable polynomial, as already mentioned; in $\mathbb{C}[x]$ it factors as $(x + i)(x + i)(x - i)(x - i)$. If $a \neq 0$, the polynomial $x^n - a \in \mathbb{Q}[x]$ is separable over $\mathbb{Q}$ for all $n \geq 1$, as we have seen that it has n distinct roots in $\mathbb{C}$ in Example 16.45.

Example 16.54. Here is an example of a polynomial which is inseparable and also irreducible. Let $F = \mathbb{F}_2(y)$ be the field of rational functions in one variable over the field $\mathbb{F}_2$ with two elements. Note that $\operatorname{char} F = 2$. We claim that $f = x^2 - y \in F[x]$ is an irreducible polynomial over F. This follows from the Eisenstein criterion, thinking of F as the field of fractions of $\mathbb{F}_2[y]$, since y is prime in $\mathbb{F}_2[y]$. Now let $K = F[x]/(x^2 - y)$ and think of $F \subseteq K$ as a field extension as usual. In K there is a root $\alpha = x + (x^2 - y)$ of the polynomial $x^2 - y$. In other words, $\alpha^2 = y$ in K. Now note that $(x - \alpha)^2 = x^2 - 2\alpha + \alpha^2 = x^2 + y = x^2 - y$ since we are in characteristic 2. Thus the irreducible polynomial $x^2 - y$ factors in $K[x]$ as the square $(x - \alpha)^2$ and thus has only the single root α in K. Hence f is inseparable.

The example above may seem complicated at first, but it is in some sense the simplest example of an inseparable irreducible polynomial. That will become clear from the next results.

A useful technical tool in studying separability is given by the formal derivative of a polynomial.

Definition 16.55. Let F be any field. If $f = a_nx^n + \cdots + a_1x + a_0 \in F[x]$ we define its *derivative* as

$$f' = na_nx^n + (n-1)a_{n-1}x^{n-1} + \cdots + 2a_2x + a_1 = \sum_{i=1}^{n} ia_ix^{i-1} \in F[x].$$

Remark 16.56. This definition requires some interpretation. In the formula $\sum_{i=0}^{n} ia_ix^{i-1}$ for f', the coefficient ia_i is the "ith multiple" of a_i, which is defined in any field F. In other words, i really means the ith multiple of 1, or the image of i under the canonical ring homomorphism $\mathbb{Z} \to F$. In particular, if $\operatorname{char} F = p > 0$ then some coefficients of f' may become 0; for example $(x^p)' = px^{p-1} = 0$.

It is also worth pointing out that there is no limiting process involved here as is used in the definition of the derivative in calculus. We are only defining the derivative of a polynomial, which is done with an explicit formula. Nonetheless, it is easy to check that this definition satisfies all of the usual differentation formulas. In particular, if $f, g \in F[x]$ then $(f+g)' = f' + g'$; $(fg)' = fg' + f'g$; and $(f^d)' = df^{d-1}f'$ for any positive integer d.

The next result gives an explicit connection between the derivative of a polynomial and separability.

Lemma 16.57. *Let $f \in F[x]$. Then f is separable if and only if* $\gcd(f, f') = 1$.

Proof. Let $F \subseteq K$ be a splitting field for f over F. In $K[x]$ we have

$$f = c(x-\alpha_1)^{e_1}(x-\alpha_2)^{e_2}\dots(x-\alpha_m)^{e_m},$$

where $\alpha_1, \alpha_2, \dots, \alpha_m$ are distinct in K; f is separable if and only if $e_i = 1$ for all i.

Note that the derivative f' of f is independent of whether we are thinking of f as a polynomial over $F[x]$ or over $K[x]$. Also, the product rule for derivatives extends to more than 2 factors as $(f_1f_2\dots f_m)' = \sum_{i=1}^{m} f_1f_2\dots f_{i-1}(f_i)'f_{i+1}\dots f_m$. Thus we have

$$f' = \sum_{i=1}^{m} ce_i(x-\alpha)^{e_1}\dots(x-\alpha_{i-1})^{e_{i-1}}(x-\alpha_i)^{e_i-1}(x-\alpha_{i+1})^{e_{i+1}}\dots(x-\alpha_m)^{e_m}.$$

From this we see that if $e_i \geq 2$, then $(x-\alpha_i)$ divides every term of the sum and so $(x-\alpha_i)|f'$ (in $K[x]$). Of course $(x-\alpha_i)|f$ also, so $(x-\alpha_i)|\gcd_{K[x]}(f, f')$. Conversely, if $e_i = 1$, then $(x-\alpha_i)$

divides every term of the sum except the ith; so $(x - \alpha_i)$ does not divide f'. Since the $(x - \alpha_i)$ are the only irreducible factors of f in $K[x]$, if $e_i = 1$ for all i then we get $\gcd_{K[x]}(f, f') = 1$.

We have proved that f is separable if and only if $\gcd_{K[x]}(f, f') = 1$. However, the gcd of two polynomials in $F[x]$ is the same whether calculated over $F[x]$ or over $K[x]$; this can be shown by noting that the steps in the Euclidean algorithm are the same in either case. Thus f is separable if and only if $\gcd_{F[x]}(f, f') = 1$. □

The lemma has immediate interesting consequences for what an irreducible inseparable polynomial could possibly look like.

Proposition 16.58. *Suppose that $f \in F[x]$ is irreducible over F. Then f is inseparable if and only if*

(i) $\operatorname{char} F = p$ *for some* $p > 0$*; and*

(ii) $f = \sum_{i=0}^{n} b_i x^{ip}$ *for some* $b_i \in F$*.*

Proof. Suppose that f is inseparable. By Lemma 16.57, $\gcd(f, f') \neq 1$. However, since f is irreducible, its only divisors (up to associates) are 1 and f. Thus $\gcd(f, f') = f$. But how can this happen? Note that $\deg f' < \deg f$ always. If $f | f'$ this forces $f' = 0$.

Now since f is irreducible, $\deg f \geq 1$. If $f = \sum_{i=0}^{n} a_i x^i$ then $f' = \sum_{i=1}^{n} i a_i x^{i-1}$ so $i a_i = 0$ for all $i \geq 1$. We can think of this as $(i \cdot 1) a_i = 0$ where $i \cdot 1$ is the ith multiple of 1. Since F is a domain, for each i either $a_i = 0$ or else $i \cdot 1 = 0$; the latter happens if and only if $\operatorname{char} F = p > 0$ and i is a multiple of p. It follows that if $\operatorname{char} F = 0$ then f can have only a constant term, so $\deg f = 0$ and f is not irreducible, a contradiction. Thus we must have as in (i) that $\operatorname{char} F = p > 0$, and we see that f is a polynomial whose only nonzero coefficents are the a_i where i is a multiple of p. By reindexing such a polynomial we get one in the form of (ii).

Conversely, if (i) and (ii) hold, a similar argument shows that $f' = 0$, and thus $\gcd(f, f') = f \neq 1$. Thus f is inseparable by Lemma 16.57. □

The proposition implies the useful fact that for a field of characteristic 0, all irreducible polynomials are separable. In particular, when working over $\mathbb{Q}$, as is the main setting for many investigations in field theory, separability becomes a non-issue. The characteristic p setting is still very important to applications, however, and so it is interesting to push our results in this case further. We will see shortly that certain special fields of characterstic p also have no irreducible inseparable polynomials.

Definition 16.59. Let R be a commutative ring with $\operatorname{char} R = p > 0$. Then the *Frobenius homomorphism* is the map $\phi : R \to R$ given by $\phi(a) = a^p$.

Note that while the pth power map preserves multiplication in any commutative ring, the preservation of addition works here only because we are in characteristic p. This follows from the binomial formula:

$$\phi(a+b) = (a+b)^p = \sum_{i=0}^{p} \binom{p}{i} a_i b^{p-i} = a^p + b^p$$

because $\binom{p}{i} = p!/(i)!(p-i)!$ is a multiple of p for all $0 < i < p$. Thus the Frobenius homomorphism really is a homomorphism of rings.

Proposition 16.60. *Let F be a field with $\operatorname{char} F = p > 0$. If the Frobenius homomorphism $\phi : F \to F$ is surjective, then every irreducible polynomial $f \in F[x]$ is separable.*

Proof. Let $f \in F[x]$ be irreducible and suppose that f is inseparable over F. By Proposition 16.58, $f = \sum_{i=0}^{n} b_i x^{ip}$ for $b_i \in F$. Now since the Frobenius is surjective, every element of F is a pth power. In particular, $b_i = (a_i)^p$ for some $a_i \in F$. Then

$$f = \sum_{i=0}^{n} b_i x^{ip} = \sum_{i=0}^{n} (a_i)^p (x^i)^p = \sum_{i=0}^{n} (a_i x^i)^p = (\sum_{i=0}^{n} a_i x^i)^p.$$

But now f factors as a product of p copies of the polynomial $g = \sum_{i=0}^{n} a_i x^i \in F[x]$, so f is not irreducible, a contradiction. □

When F is a field, we write the image of the Frobenius map $\phi : F \to F$ as $F^p = \{a^p | a \in F\}$. Since the Frobenius is a homomorphism of fields, it is injective and so F^p is a subfield of F which is isomorphic to F. The Frobenius need not be surjective, however.

Definition 16.61. A field F is *perfect* if either $\operatorname{char} F = 0$ or else $\operatorname{char} F = p > 0$ and $F^p = F$.

The following result justifies including these two very different cases in one definition.

Proposition 16.62. *A field F is perfect if and only if every irreducible polynomial in $F[x]$ is separable.*

Proof. We have seen that if either $\operatorname{char} F = 0$ or $\operatorname{char} F = p$ and $F^p = F$, then every irreducible $f \in F[x]$ is separable; see Proposition 16.58 and Proposition 16.60.

Conversely, suppose that F is not perfect. Thus there is a prime p such that $\operatorname{char} F = p > 0$ and $F^p \neq F$. thus we can pick $a \in F$ such that a is not a pth power of an element in F. Now let $f = x^p - a \in F[x]$ and find a field extension $F \subseteq K$ such that f has a root in K, say $\alpha \in K$. This means that $\alpha^p - a = 0$, so α is a pth root of a in K. Now $(x-\alpha)^p = x^p - \alpha^p = x^p - a = f$, so $f = (x-\alpha)^p$ already splits in $K[x]$ as a pth power. Thus f is inseparable.

On the other hand, we claim that f is irreducible over F. If not, then $f = gh$ with $g, h \in F[x]$ where $\deg g \geq 1$, $\deg h \geq 1$. Now because of the factorziation of f in $K[x]$ is a pth power of a degree 1 irreducible, we must have $g = (x-\alpha)^i$ and $h = (x-\alpha)^j$ in $K[x]$, where $i + j = p$ and $i, j \geq 1$. But now the constant term of g is $\pm\alpha^i$, so $\alpha^i \in F$. Since $\gcd(i, p) = 1$ and $\alpha^p = a \in F$, writing $1 = mi + np$ we get $\alpha^1 = (\alpha^i)^m(\alpha^p)^n \in F$, a contradiction since a has no pth root in F. Thus f is irreducible over F as claimed. □

Example 16.63. Let F be a field with $\operatorname{char} F = p > 0$, and suppose that F is finite. Then F must be perfect. Indeed, the Frobenius map $\phi : F \to F$ is always injective (since F is a field). Then since F is a finite set this forces ϕ to be surjective as well.

Example 16.64. Let $F = \mathbb{F}_p(y)$ be a field of fractions functions in one variable over $\mathbb{F}_p$. Then F is not perfect. In fact, this must be true, we already saw Example 16.54 that $F[x]$ has an irreducible inseparable polynomial when $p = 2$, and a similar example works for arbitrary p.

But we can also check it directly from the definition of perfect, by showing that $y \in F$ has no pth root. Indeed, suppose that $f/g \in F$, where $f, g \in \mathbb{F}_p[y]$, and that $(f/g)^p = y$. Then $f^p = yg^p$ in $\mathbb{F}_p[y]$. But then considering degrees we have $p \deg f = p \deg g + 1$, which is absurd.

There are infinite fields of characteristic p which are perfect, as well. An easy example is any algebraically closed field of characteristic p (we will see later that these exist). Over an algebraically closed field there are no irreducible polynomials except those of degree 1, which are trivially separable.

16.6. **Finite fields and the Theorem of the Primitive Element.** In the first part of this section, we give the basic structure theory of fields with finitely many elements.

We start with an easy group theory result and its application to the structure of the multiplicative group of a field.

Lemma 16.65. *Let G be a finite abelian group of order n. Suppose that for each divisor d of n that G has at most d elements of order dividing d. Then G is cyclic.*

Proof. We use the classification of finite abelian groups in the invariant factor form. This tells us that G is isomorphic to an additive group $\mathbb{Z}/(a_1) \oplus \cdots \oplus \mathbb{Z}/(a_m)$, where $a_1 | a_2 | \dots | a_m$ are integers greater than 1. Suppose that $m \geq 2$. Then since $a_{m-1} | a_m$, every element of the form $g = (0, 0, \dots, 0, b, c)$ satisfies $a_m g = 0$. There are $(a_m)(a_{m-1}) > a_m$ such elements. In other words, G has more than a_m elements of order dividing a_m, contradicting the hypothesis. Thus $m = 1$ and $G \cong \mathbb{Z}/(a_1)$ is cyclic. □

Corollary 16.66. *Let F a be a field and $F^{\times} = F - \{0\}$ its multiplicative group. If G is a finite subgroup of $F^{\times}$ then G is cyclic. In particular, if F is a finite field then $F^{\times}$ is cyclic.*

Proof. Using multiplicative notation, the elements in G of order dividing d are $\{g \in G | g^d = 1\}$. In other words, these are roots in G of the polynomial $x^d - 1 \in F[x]$. This polynomial can have at most d roots in F by Corollary 16.2. Thus the hypotheses of the lemma are satisfied and we conclude that G is cyclic. The last statement is clear. □

Now let us prove the basic structural results of finite fields. Note that if F is a finite field, then certainly 1 has finite additive order and so F has positive characteristic, say p for a prime p. Then the additive subgroup generated by 1 is a subfield isomorphic to $\mathbb{F}_p$, so we can think of F as an extension of $\mathbb{F}_p$.

Theorem 16.67. *Let p be prime. For each $n \geq 1$, setting $q = p^n$ the splitting field of $x^q - x$ over $\mathbb{F}_p$ is a field $\mathbb{F}_q$ with $|\mathbb{F}_q| = q$. Conversely, if F is any finite field of characteristic p then $F \cong \mathbb{F}_q$ for $q = p^n$, some $n \geq 1$.*

Proof. Let F be the splitting field of $f = x^{p^n} - x$ over $\mathbb{F}_p$. Note that $f' = p^n x^{p^n - 1} - 1 = -1$ since we are in characteristic p, so we must have $\gcd(f, f') = 1$, and by Lemma 16.57 f is separable. Thus f has p^n distinct roots in F. Let E be the set of these roots in F, so $E = \{\alpha \in F | \alpha^{p^n} = \alpha\}$. Note that $\phi : F \to F$ given by $\phi(x) = x^{p^n}$ is the nth power of the Frobenius homomorphism; in particular, ϕ is a homomorphism of fields. Then $E = \{x \in F | \phi(x) = x\}$ is automatically a subfield of F. Now since E contains all of the roots of f, the polynomial f already splits in $E[x]$, and obviously the subfield of F generated over $\mathbb{F}_p$ by these roots must be E. Hence $F = E$ and the spliting field of f is actually equal to the set of roots of f. In particular $|F| = p^n$. Letting $q = p^n$ and writing $\mathbb{F}_q = F$, we have $|\mathbb{F}_q| = q$.

Conversely, let F be a finite field of characteristic p. As we remarked above, F contains a copy of $\mathbb{F}_p$. Thus we have a field extension $\mathbb{F}_p \subseteq F$. The degree $[F : \mathbb{F}_p] = n$ is certainly finite, since F is. Moreover, a vector space of dimension n over $\mathbb{F}_p$ clearly has precisely p^n elements, because this is the number of distinct $\mathbb{F}_p$-linear combinations of n basis vectors. Now $F^{\times} = F - \{0\}$ is a multiplicative group of order $p^n - 1$; so for any $\alpha \in F^{\times}$ we have $\alpha^{p^n - 1} = 1$. Then $\alpha^{p^n} = \alpha$. This latter equation is also true for $\alpha = 0$. Hence every element of F is a root of $f = x^{p^n} - x \in \mathbb{F}_p[x]$. Since F consists of $p^n = \deg f$ distinct roots of f, the polynomial f must already split over F as $x^{p^n} - x = \prod_{\alpha \in F}(x - \alpha) \in F[x]$. So F is a splitting field of f over $\mathbb{F}_p$. Now by the uniqueness of splitting fields up to isomorphism, setting $q = p^n$ we have that $F \cong \mathbb{F}_q$ for the field $\mathbb{F}_q$ constructed above. □

Thinking of the field $\mathbb{F}_q$ for $q = p^n$ as the splitting field of $x^{p^n} - x$ was useful for theoretical reasons, but to actually construct a field $\mathbb{F}_q$, in practice it is helpful to find it as a splitting field of a polynomial of smaller degree. This can be done with the help of our result that the multiplicative group of a finite field is cyclic.

Lemma 16.68. *Let p be prime. For each $n \geq 1$ there is at least one irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n; and for any such f, $\mathbb{F}_p[x]/(f)$ is a field isomorphic to $\mathbb{F}_{p^n}$.*

Proof. Consider the field $F = \mathbb{F}_{p^n}$ as an extension of its prime subfield $\mathbb{F}_p$. We know that $F^\times$ is a cyclic group by Corollary 16.66, say with generator γ. Since the powers of γ fill up $F^\times$, clearly $\mathbb{F}_p(\gamma) = F$, so F is a simple extension of $\mathbb{F}_p$. Consequently $F \cong \mathbb{F}_p[x]/(f)$ where $f = \text{minpoly}_{\mathbb{F}_p}(\gamma)$. But we also know that $[F : \mathbb{F}_p] = n = \deg f$, so f is an irreducible polynomial over $\mathbb{F}_p$ of degree n.

Conversely, it is clear that for any irreducible polynomial $g \in \mathbb{F}_p[x]$ of degree n, then $K = \mathbb{F}_p[x]/(g)$ is a field with $[K : \mathbb{F}_p] = n$ and hence $|K| = p^n$. □

Example 16.69. Irreducible polynomials of low degree over $\mathbb{F}_p$ can be found explicitly as described in Example 16.6. For example, as shown there, $f = x^4 + x + 1$ is irreducible over $\mathbb{F}_2$ and thus $\mathbb{F}_{16} \cong \mathbb{F}_2[x]/(x^4+x+1)$. This gives an explicit description of $\mathbb{F}_{16}$ that allows one to do calculations in this field, by writing its elements as $\{a_0 + a_1x + a_2x^2 + a_3x^3 + (f) | a_i \in \mathbb{F}_2\}$ and doing arithmetic of such elements modulo (f).

For example, consider $x \in \mathbb{F}_{16}$ (ommitting the $+(f)$ and just remembering to do calculations modulo f). The relation tells us that $x^4 = -x - 1 = x + 1$. Then $x^5 = x^2 + x$ and $x^3 = x^3$ are not equal to 1 (since the representative of a coset with degree ≤ 3 is uniquely determined). This shows that the order of x in $\mathbb{F}_{16}^\times$ is 15 and so x is a generator of the cyclic group $\mathbb{F}_{16}^\times$.

Corollary 16.66 is also relevant to the proof of a basic theorem often called the "Theorem of the Primitive Element". It gives a useful criterion for when a finite degree extension is a simple extension, i.e. generated by one element. The name arises from the fact that in a simple extension $F \subseteq F(\gamma)$ the element γ is sometimes called a primitive element for the extension.

Theorem 16.70. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$. The following are equivalent:*

(i) *$K = F(\gamma)$ for some $\gamma \in K$.*

(ii) *There are finitely many subfields E with $F \subseteq E \subseteq K$.*

Proof. Different arguments are required here depending on whether F is finite or infinite. The finite case follows quickly from results we have already proved. If $|F| < \infty$ then $|K| = [K : F]|F| < \infty$

also. Then $K^\times$ is a cyclic group by Corollary 16.66, generated by some γ, say. Thus $K = F(\gamma)$, so every finite degree extension of a finite field is simple and (i) is automatic. Condition (ii) also trivially holds when F and hence K is finite, since K has finitely many distinct subsets.

Assume for the rest of the proof that $|F| = \infty$. Suppose as in (ii) that there are finitely many subfields E with $F \subseteq E \subseteq K$ (these are called *intermediate* fields for the extension $F \subseteq K$). Suppose that $\alpha, \beta \in K$ and consider the fields $E_a = F(\alpha + a\beta)$ as a ranges over elements of F. Since each E_a is an intermediate field, of which there are only finitely many, yet $|F| = \infty$, we must have $E_a = E_b$, for some $a \neq b$. Then E_a contains $\alpha + a\beta - (\alpha + b\beta) = (a-b)\beta$ and since $0 \neq a - b \in F$, $\beta \in E_a$. But then $a\beta \in E_a$ and so $\alpha \in E_a$ also. It follows that $E_a = F(\alpha + a\beta) = F(\alpha, \beta)$. This shows that the subfield of K generated over F by any two elements can actually be generated by one element. Since $[K : F] < \infty$, K/F is certainly finitely generated. By iteratively replacing pairs of generators by a single generator we obtain that $K = F(\gamma)$ for some γ.

Conversely, suppose that $K = F(\gamma)$ for some $\gamma \in K$. Let $F \subseteq E \subseteq K$ where E is an intermediate field. Since $[K : F] < \infty$, certainly $[K : E] < \infty$, so γ is algebraic over E. Let $f = \text{minpoly}_E(\gamma) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in E[x]$. Since all $a_i \in E$, we have $E' = F(a_0, \dots, a_{n-1}) \subseteq E$. Since f is irreducible in $E[x]$, and also $f \in E'[x]$, certainly f is irreducible in $E'[x]$. Thus $f = \text{minpoly}_{E'}(\gamma)$ also. But this means that $[K : E'] = [E'(\gamma) : E'] = \deg f = [E(\gamma) : E] = [K : E]$. This forces $[E : E'] = 1$ and hence $E' = E$. In particular, E is generated over F by the coefficients of $f = \text{minpoly}_E(\gamma)$. Now let $g = \text{minpoly}_F(\gamma)$. Then since $g(\gamma) = 0$, we have $f | g$ in $E[x]$, hence in $K[x]$. Since there are only finitely many monic polynomials that divide g in $K[x]$, there are finitely many possible f and so finitely many intermediate fields E. □

We will see later that if $F \subseteq K$ is a finite degree extension such that every $\alpha \in K$ has a separable minimal polynomial over F, then condition (ii) above holds and hence $K = F(\gamma)$ is a simple extension. For example, an arbitrary finite degree extension of $\mathbb{Q}$ can be generated by one element, which is quite surprising.

17. Galois Theory

17.1. **Separable and normal extensions.** In this section we study two special properties of field extensions, separability and normality, and their relations to automorphisms of fields. This will lay the main groundwork for the fundamental theorem of Galois Theory in the next section. This theory is named for the French mathematician Évariste Galois (pronounced "gal-wah"), who developed its main ideas at a young age before his life was tragically cut short in a duel. At the time, there was no notion of a "group"—the groups that occured in Galois's work were explicit subsets of the group of permutations of the roots of a polynomial. These ideas nonetheless helped to point the way to the idea of an abstract group which was formulated later in the 1800's.

Definition 17.1. Let $F \subseteq K$ be a field extension. We define

$$\mathrm{Aut}(K) = \{\sigma : K \to K \,|\, \sigma \text{ is an automorphism of fields}\},$$

which is a group under composition. The *Galois group* of the field extension K/F is

$$\mathrm{Gal}(K/F) = \{\sigma \in \mathrm{Aut}(K) \,|\, \sigma(a) = a \text{ for all } a \in F\}.$$

It is a subgroup of $\mathrm{Aut}(K)$.

We read $\mathrm{Gal}(K/F)$ as "Galois K over F". We say that the elements of $\mathrm{Gal}(K/F)$ *fix* F. Note that these elements are required to fix F pointwise, not just as an overall set.

One reason it is interesting to consider automorphisms of K that fix F is the following. If $\sigma \in \mathrm{Gal}(K/F)$ and $f \in F[x]$, then if $\alpha \in K$ is a root of f, then $\sigma(\alpha)$ is also a root of f. Explicitly, if $f = \sum a_i x^i$ with $a_i \in F$, then $\sum a_i \alpha^i = 0$, so applying σ we get

$$\sum \sigma(a_i)\sigma(\alpha)^i = \sum a_i \sigma(\alpha)^i = f(\sigma(\alpha)) = 0,$$

since σ fixes F. Thus elements in $\mathrm{Gal}(K/F)$ must permute any roots of $f \in F[x]$ that lie in K. We will use this fact frequently.

It turns out that finite degree extensions which have "enough" automorphisms have particularly good properties and will be the focus of the fundamental theorem later. The relevant definition is the following.

Definition 17.2. A finite degree extension K/F is called *Galois* if $|\,\mathrm{Gal}(K/F)| = [K : F]$.

Here are two examples that show the sort of things that might go wrong and prevent an extension from being Galois.

Example 17.3. Consider $\mathbb{Q} \subseteq K = \mathbb{Q}(\sqrt[3]{2})$ inside $\mathbb{C}$. Since $K \subseteq \mathbb{R}$ and the other two roots of $x^3 - 2$ in $\mathbb{C}$ are not real, $\sqrt[3]{2}$ is the only root of $x^3 - 2$ in K. Now if $\sigma \in G = \mathrm{Gal}(K/\mathbb{Q})$, then by the remark above $\sigma(\sqrt[3]{2}) = \sqrt[3]{2}$ is forced. Since $\sigma \in G$ fixes $\mathbb{Q}$ and the element $\sqrt[3]{2}$ which generates K, it follows that $\sigma = 1_K$. Thus G is trivial and $|\,\mathrm{Gal}(K/\mathbb{Q})| = 1 < [K : \mathbb{Q}] = 3$.

Example 17.4. Consider $F = \mathbb{F}_2(y)$. As we have already seen in Example 16.54, the polynomial $f = x^2 - y \in F[x]$ is inseparable and irreducible. If K is a splitting field of f over F then there is $\alpha \in K$ such that $f(\alpha) = 0$, that is $\alpha^2 = y$, and where $f = (x - \alpha)^2 \in K[x]$. So $K = F(\alpha)$. Again if

$\sigma \in \text{Gal}(K/F)$ then $\sigma(\alpha) = \alpha$ because σ permutes the roots of $f \in F[x]$ (and α is the only root). This implies that $\sigma = 1_K$ and so $|\text{Gal}(K/F)| = 1 < [K : F] = 2$.

In the next results we will see that the things that happened in the two examples above are the *only* things that can go wrong in an extension K/F that is not Galois—either there is an irreducible polynomial in $F[x]$ that does not entirely split in $K[x]$, so it doesn't have enough roots in K, or a polynomial in $F[x]$ that splits in $K[x]$ but with indistinct roots, so again there are not enough different places for an automorphism to send the roots. This leads to the following two definitions.

Definition 17.5. Let $F \subseteq K$ be an algebraic field extension. We say the extension K/F is *separable* if for all irreducible polynomials $f \in F[x]$, if f has a root in K then f is separable.

Definition 17.6. Let $F \subseteq K$ be an algebraic field extension. We say the extension K/F is *normal* if for all irreducible polynomials $f \in F[x]$, if f has a root in K then f splits over K.

An alternative way to define these notions is using minimal polynomials: given an extension K/F, it is normal if for all $\alpha \in K$, $\text{minpoly}_F(\alpha)$ splits in $K[x]$, and it is separable if for all $\alpha \in K$, $\text{minpoly}_F(\alpha)$ is a separable polynomial. This follows immediately from the fact that any irreducible polynomial in $F[x]$ that has α as a root must be the minimal polynomial of α.

Note that in Example 17.3, the extension $\mathbb{Q} \subseteq K$ is not normal, because the irreducible polynomial $f = x^3 - 2 \in \mathbb{Q}[x]$ has the root $\sqrt[3]{2} \in K$ but does not split over K. In Example 17.4, the extension $F \subseteq K$ fails to be separable, because the minimal polynomial of α is the irreducible inseparable polynomial $x^2 - y \in F[x]$.

It is useful to see how the separable and normal properties pass to smaller extensions.

Lemma 17.7. *Let $F \subseteq E \subseteq K$ be field extensions where K/F is algebraic.*

(1) *If K/F is separable, then E/F and K/E are separable.*

(2) *if K/F is normal, then K/E is normal.*

Proof. (1) It is obvious that E/F is separable from the definition–we are just checking $\text{minpoly}_F(\alpha)$ is separable for those $\alpha \in E$, rather than for all $\alpha \in K$. Now if $\alpha \in K$, consider $g = \text{minpoly}_E(\alpha)$. If $f = \text{minpoly}_F(\alpha)$, then since $f(\alpha) = 0$ and $f \in F[x] \subseteq E[x]$, we have $g|f$ in $E[x]$. But now since f has distinct roots in a splitting field, so does its factor g. So K/E is also separable.

(2) Similarly as in part (1), for $\alpha \in K$, $g = \text{minpoly}_E(\alpha)$ divides $f = \text{minpoly}_F(\alpha)$. Now since f splits over K, so does its factor g. □

It turns out that finite degree normal extensions are the same as splitting fields of polynomials. Both points of view are useful, since the notion of normality doesn't depend on a choice of polynomial, so it is easier to work with abstractly; while thinking in terms of the splitting field of a particular polynomial is important in calculations.

Lemma 17.8. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$. Then K/F is normal if and only if there is a polynomial $f \in F[x]$ such that K is the splitting field of f over F.*

Proof. Suppose first that K is the splitting field over F of $f \in F[x]$. Thus $f = a(x-\alpha_1)\dots(x-\alpha_m)$ in $K[x]$, with $K = F(\alpha_1, \dots, \alpha_m)$. Let $g \in F[x]$ be irreducible over F and assume that $g(\beta_1) = 0$ with $\beta_1 \in K$. Let $K \subseteq L$ where L is the splitting field of g over K. So the polynomial g splits as $g = a(x - \beta_1)(x - \beta_2)\dots(x - \beta_n)$ in $L[x]$, and $L = K(\beta_1, \dots, \beta_n)$.

Now note that f and g both split over L, and that $L = F(\alpha_1, \dots, \alpha_m, \beta_1, \dots \beta_n)$. This implies that L is a splitting field of $h = fg \in F[x]$ over F. By Corollary 16.50, since g is an irreducible factor of h, for any root β_i of g there is an automorphism σ of L such that σ satisfies $\sigma|_F = 1_F$, in other words σ fixes F, and $\sigma(\beta_1) = \beta_i$. On the other hand, since f has coefficients in F, σ must permute the roots $\{\alpha_j\}$ of f as well. Since K is generated by the $\{\alpha_j\}$ over F, $\sigma(K) = K$. In particular, since $\beta_1 \in K$, we get $\sigma(\beta_1) = \beta_i \in K$ for all i. This shows that $L \subseteq K$ and hence $K = L$. Thus g already splits over K. We see that any irreducible polynomial in $F[x]$ with a root in K splits over K, so K/F is normal.

The converse is easier. If K/F is normal, then since $[K : F] < \infty$ the extension K/F is certainly finitely generated; say $K = F(\gamma_1, \dots, \gamma_r)$. Let $g_i = \text{minpoly}_F(\gamma_i)$. By the normality condition, each g_i must split over K. But then $f = g_1 g_2 \dots g_r$ is a polynomial that splits over K, and the set of all of its roots generates K over F since these roots are contained in K and include all of the γ_i. So K is the splitting field of f over F. □

The lemma allows for an alternate proof of Lemma 17.7: If $F \subseteq E \subseteq K$ with K/F normal, then we know that K is the splitting field over F over some $f \in F[x]$. It is easy to see then that K is also the splitting field over E of the same f, so K/E must also be normal.

Example 17.9. We can also now give an example showing that if $F \subseteq E \subseteq K$ and K/F is normal, then E/F need not be normal. Consider $F = \mathbb{Q} \subseteq E = \mathbb{Q}(\sqrt[3]{2}) \subseteq K = \mathbb{Q}(\sqrt[3]{2}, \zeta)$ where $\zeta = e^{2\pi i/3}$. We have seen in Example 16.45 that K is the splitting field over F of $x^3 - 2$. On the other hand, we saw that E is not normal over F in Example 17.3. Thus normality of K/F does not pass in general to the subextension E/F.

We now give the main result of this section. It shows that the failure of an extension to be Galois is always because there are two few automorphisms, not too many; and this always happens essentially because of one of the two problems exhibited in Examples 17.3 and 17.4, namely lack of normality or lack of separability.

Theorem 17.10. *Let $F \subseteq K$ be a field extension with $[K : F] < \infty$.*

(1) $|\operatorname{Gal}(K/F)| \leq [K : F]$.

(2) *The following are equivalent:*

  (i) *K/F is Galois, i.e. $|\operatorname{Gal}(K/F)| = [K : F]$.*

  (ii) *K/F is separable and normal.*

  (iii) *K is the splitting field over F of a separable polynomial $f \in F[x]$.*

Proof. (1) If $K = F$ the result is vacuous, so assume that $[K : F] \geq 2$. Pick any $\alpha_1 \in K - F$ and let $g = \operatorname{minpoly}_F(\alpha_1)$, so g is irreducible over F and $n = \deg g \geq 2$, with $[F(\alpha_1) : F] = n$. Let $\{\alpha_1, \ldots, \alpha_m\}$ be the set of all elements in K that are roots of g, so $m \leq n$.

If $\sigma \in \operatorname{Gal}(K/F)$, then $\sigma(\alpha_1) = \alpha_i$ for some i, because $g \in F[x]$. Let

$$S = \{i \in \{1, 2, \ldots, m\} | \text{there exists an automorphism } \sigma \in \operatorname{Gal}(K/F) \text{ such that } \sigma(\alpha_1) = \alpha_i\}.$$

For each $i \in S$ fix an automorphism $\sigma_i \in \operatorname{Gal}(K/F)$ such that $\sigma_i(\alpha_1) = \alpha_i$.

Now let us prove that $|\operatorname{Gal}(K/F)| \leq [K : F]$ by induction on the degree $[K : F]$. We have $[K : F(\alpha_1)] < [K : F]$. By the induction hypothesis, $|\operatorname{Gal}(K/F(\alpha_1))| \leq [K : F(\alpha_1)]$. Now if $\tau \in \operatorname{Gal}(K/F)$ is arbitrary, then $\tau(\alpha_1) = \alpha_i$ for some $i \in S$; then $\rho = (\sigma_i)^{-1} \circ \tau \in \operatorname{Gal}(K/F)$ and $\rho(\alpha_1) = \alpha_1$. Thus ρ fixes $F(\alpha_1)$ pointwise, and so $\rho \in \operatorname{Gal}(K/F(\alpha_1))$. Let $H = \operatorname{Gal}(K/F(\alpha_1))$. We conclude that $\tau = \sigma_i \circ \rho \in \sigma_i H$. Thus

$$|\operatorname{Gal}(K/F)| \leq |S||H| \leq m|H| \leq n|\operatorname{Gal}(K/F(\alpha_1))| \leq |F(\alpha_1) : F||K : F(\alpha_1)| = |K : F|. \tag{17.11}$$

(2) Consider the argument in part (1). In order to have equality in (17.11), so that $|\operatorname{Gal}(K/F)| = |K : F|$, it is necessary and sufficient that $|S| = m = n$ and $|\operatorname{Gal}(K/F(\alpha_1))| = [K : F(\alpha_1)]$.

(i) $\implies$ *(ii)*. The case $K = F$ is trivial. Assume that K/F is Galois with $[K : F] > 1$. Choose any $\alpha_1 \in K - F$ and let $g = \operatorname{minpoly}_F(\alpha_1)$. As just remarked, we must have the number m of elements in K that are roots of g equal to the degree n of g in the argument of part (1). In particular, this forces g to split over K, and with distinct roots, so that g is a separable polynomial. By definition, K/F is normal and separable.

(ii) $\implies$ *(iii)*. Since K/F is normal, K is the splitting field over F of some polynomial $f \in F[x]$ by Lemma 17.8. We may assume that f is monic. Write $f = (g_1)^{e_1} \ldots (g_m)^{e_m}$ where the g_i are

monic irreducibles in $F[x]$ such that $g_1, \dots, g_m$ are all distinct, and $e_i \geq 1$. It is clear that K is also the splitting field of $h = g_1 \dots g_m$ over F. Now if g_i and g_j have a common root $\alpha \in K$, then $g_i = \text{minpoly}_F(\alpha) = g_j$, a contradiction. Also, since K/F is separable, each irreducible polynomial g_i is separable and thus splits with distinct roots in K. We conclude that h has distinct roots in K and so is a separable polynomial, and K is the splitting field of h over F.

$(iii) \implies (i)$. The proof is by induction on $[K : F]$, with the case $F = K$ trivial as usual. Assume that K is the splitting field of a separable polynomial f over F, where f does not split over F (otherwise we are back to the trivial case $F = K$). Let g be an irreducible factor of f in $F[x]$ with $\deg g \geq 2$, and apply the argument in (1) to a root $\alpha_1 \in K$ of g. Since K is a splitting field of f, the polynomial g splits over K, and since f is separable, so is g, so g has $m = n = \deg g$ distinct roots in K. In addition, for any root α_i of g, we can find an automorphism $\sigma_i \in \text{Gal}(K/F)$ such that $\sigma_i(\alpha_1) = \alpha_i$, by Corollary 16.50. So $|S| = m = n$ in the argument of part (1). Finally, K is also the splitting field of the separable polynomial f over $F(\alpha_1)$. Since $[K : F(\alpha_1)] < [K : F]$, by induction we may assume that $|\text{Gal}(K/F(\alpha_1))| = [K : F(\alpha_1)]$. Now (17.11) implies that $|\text{Gal}(K/F)| = [K : F]$. □

Corollary 17.12. *Let K/F be a finite degree Galois extension. Then for every intermediate field E with $F \subseteq E \subseteq K$, the extension K/E is also Galois.*

Proof. By Theorem 17.10, we know that for a finite degree extension being Galois is equivalent to being normal and separable. But both properties pass from K/F to K/E by Lemma 17.7. □

Example 17.9 is an example of a Galois extension K/F such that E/F is not Galois.

Suppose $F \subseteq K$ is a finite-degree extension that is not necessarily normal. Then we can embed it canonically in a normal extension, as follows.

Proposition 17.13. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$.*

(1) *There is an extension $K \subseteq L$ with $[L : K] < \infty$ such that L/F is normal, and where L is minimal in the sense that if $K \subseteq E \subseteq L$ with E/F normal, then $E = L$.*
(2) *If K/F is separable in (1), then L/F is Galois.*

Proof. (1) K/F is certainly finitely generated, say $K = F(\alpha_1, \dots, \alpha_m)$. Let $g_i = \text{minpoly}_F(\alpha_i) \in F[x]$. Some of the g_i may be equal; let f be the product of the distinct g_i, each once.

Now take L to be a splitting field of f over K. We define the splitting field over K, not over F, because we need to ensure the resulting field L contains K. However, if $\beta_1, \dots, \beta_n$ are the roots of

f in L then

$$L = K(\beta_1, \ldots, \beta_n) = F(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n) = F(\beta_1, \ldots, \beta_n)$$

because the elements α_i are roots of f and hence lie among the β_j. Thus L is in fact the splitting field of f over F as well. By Lemma 17.8, L/F is now a normal extension.

To see that L is minimal, note that if $K \subseteq E \subseteq L$ is an extension where E/F is normal, then each irreducible polynomial g_i has a root in K and so must split in E. But then all of the roots β_j of f in L are already in E, so $E = L$ since L is generated over K by the β_j.

(2) Now suppose that K/F is separable. Then each g_i is a separable polynomial, by definition. As we saw previously, distinct monic irreducible polynomials in $F[x]$ cannot have a common root. Since each g_i has distinct roots in the splitting field L, the product f of the distinct g_i is a separable polynomial. This L is a splitting field over F of a separable polynomial f. By Theorem 17.10, L is Galois over F. □

The extension L/F constructed in (1) above is called the *normal closure* of K/F. It is unique up to isomorphism, as the reader may check. When K/F is separable, so that the normal closure L/F is Galois as in part (2), then it is called the *Galois closure*.

17.2. **The Fundamental Theorem of Galois Theory.** The fundamental theorem we will prove in this section gives a surprisingly tight connection between group theory and field theory. Namely, the set of intermediate fields E such that $F \subseteq E \subseteq K$ where K/F is a finite degree Galois extension will be in one-to-one correspondence with the set of subgroups of the group $G = \mathrm{Gal}(K/F)$. Moreover, the correspondence will have many other special properties.

The way the correspondence is set up is not particularly complicated to describe. Let $F \subseteq K$ be an extension of fields. Let $G = \mathrm{Gal}(K/F)$, which we know is a group under composition. If E is an intermediate field with $F \subseteq E \subseteq K$, then we can define a subgroup H of G by $H = \mathrm{Gal}(K/E)$. In other words, these are the automorphisms of K that fix (pointwise) the larger field E, not just F. This is obviously a subset of G, and since the elements fixing E are closed under products and inverses, H is a subgroup of G.

Conversely, if we start with a subgroup H of $G = \mathrm{Gal}(K/F)$, then we define an intermediate field E by $E = \mathrm{Fix}(H) = \{\alpha \in K | \sigma(\alpha) = \alpha \text{ for all } h \in H\}$. This is called the *fixed field of H*. Because the elements in H are automorphisms of K it is clear that $\mathrm{Fix}(H)$ is a subfield of K; and $F \subseteq \mathrm{Fix}(H)$ since every element of G fixes F. So $F \subseteq \mathrm{Fix}(H) \subseteq K$ and $\mathrm{Fix}(H)$ is an intermediate field.

Thus for a fixed field extension $F \subseteq K$ we have the following setup and notation:

(17.14)

$$\begin{tikzcd} \left\{\text{intermediate fields } E \text{ with } F \subseteq E \subseteq K\right\} \arrow[r, shift left, "\Gamma=\mathrm{Gal}(K/-)"] & \left\{\text{subgroups } H \text{ of } G = \mathrm{Gal}(K/F)\right\} \arrow[l, shift left, "\Phi=\mathrm{Fix}(-)"] \end{tikzcd}$$

There are some basic properties of the maps Γ and Φ that we follow directly from the definitions. First, if E is an intermediate field, then $E \subseteq \Phi\Gamma(E) = \operatorname{Fix}\operatorname{Gal}(K/E)$; this is clear since every element of $\mathrm{Gal}(K/E)$ fixes E at least (but there could be a larger field of elements fixed by $\mathrm{Gal}(K/E)$). Similarly, if H is a subgroup of $G = \mathrm{Gal}(K/F)$ then $H \subseteq \Gamma\Phi(H) = \mathrm{Gal}(K/\operatorname{Fix}(H))$; because certainly H is contained in the set of those elements of G that fix everything in $\mathrm{Fix}(H)$ (though there could be more elements that do). Second, both Φ and Γ are *inclusion reversing*. Namely, if $F \subseteq E \subseteq L \subseteq K$, then $\Gamma(L) = \mathrm{Gal}(K/L) \subseteq \Gamma(E) = \mathrm{Gal}(K/E)$, since an automorphism that fixes L pointwise certainly also fixes E. Similarly, if $H \subseteq J \subseteq G$ are subgroups of G, then $\Phi(J) = \mathrm{Fix}(J) \subseteq \Phi(H) = \mathrm{Fix}(H)$, since the elements fixed pointwise by everything in J are certainly also fixed by everything in H.

While the set up above in (17.14) makes sense for any field extension $F \subseteq K$, we will prove it has especially good properties when $[K : F] < \infty$ and K/F is Galois. In that case we will see shortly that Γ and Φ are inverse bijections of sets, so they will define a one-to-one (inclusion reversing) correspondence between intermediate fields and the subgroups of the Galois group.

Understanding the action of $\Phi\Gamma$ on intermediate subfields of a Galois extension is easy from what we have already done.

Lemma 17.15. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$ and K/F Galois. For every intermediate field E with $F \subseteq E \subseteq K$, we have $\Phi\Gamma(E) = \operatorname{Fix}\operatorname{Gal}(K/E) = E$.*

Proof. By Corollary 17.12, K/E is a Galois extension with $[K : E] < \infty$. Let $E' = \operatorname{Fix}\operatorname{Gal}(K/E)$, so $E \subseteq E' \subseteq K$. Now by the inclusion reversing property of $\mathrm{Gal}(K/-)$ we have $\mathrm{Gal}(K/E') \subseteq \mathrm{Gal}(K/E)$. On the other hand, if $\sigma \in \mathrm{Gal}(K/E)$ then by definition σ fixes pointwise everything in $E' = \operatorname{Fix}\operatorname{Gal}(K/E)$, and so $\sigma \in \mathrm{Gal}(K/E')$. Hence $\mathrm{Gal}(K/E) = \mathrm{Gal}(K/E')$. Also, K/E' is a Galois extension by Corollary 17.12. It follows that $[K : E] = |\operatorname{Gal}(K/E)| = |\operatorname{Gal}(K/E')| = [K : E']$. But this forces $[E' : E] = 1$ and hence $E = E'$. □

An immediate corollary is the following version of the theorem of the primitive element.

Corollary 17.16. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$. If K/F is separable, then it has a primitive element; in other words $K = F(\gamma)$ for some $\gamma \in K$.*

Proof. Using Proposition 17.13, we can take a Galois closure L/F of K/F; so $F \subseteq K \subseteq L$ and L/F is Galois. Now by Lemma 17.15, for any $F \subseteq E \subseteq L$ we have $E = \operatorname{Fix} \operatorname{Gal}(K/E)$. If $G = \operatorname{Gal}(L/F)$ then G is finite and so has finitely many subgroups. Since every E is the fixed field of the subgroup $\operatorname{Gal}(K/E)$ of G, there are finitely many intermediate fields E. Since the extension L/F has finitely many intermediate fields, of course the smaller extension K/F also has this property. Now apply Theorem 16.70. □

We note that a finite degree inseparable extension might well have infinitely many intermediate fields.

To understand the action of $\Gamma\Phi$ on subgroups of $\operatorname{Gal}(K/F)$ we need one more new idea, which is a way of determining the minimal polynomial of an element using the action of the Galois group.

Lemma 17.17. *Let K/F be a finite degree Galois extension. Suppose that H is a subgroup of $\operatorname{Gal}(K/F)$ such that $\operatorname{Fix}(H) = F$.*

(1) *For any $\alpha \in K$, let $\mathcal{O}_\alpha = \{\sigma(\alpha) | \sigma \in H\}$. Then $\operatorname{minpoly}_F(\alpha)$ is equal to $\prod_{\beta \in \mathcal{O}_\alpha}(x - \beta)$.*

(2) *$H = \operatorname{Gal}(K/F)$.*

Proof. (1) Given any automorphism $\sigma \in \operatorname{Gal}(K/F)$, $\sigma : K \to K$ extends to an automorphism σ of $K[x]$ in the usual way, by acting on the coefficients of a polynomial. Apriori the polynomial $f = \prod_{\beta \in \mathcal{O}_\alpha}(x - \beta)$ just lies in $K[x]$, but we claim it is actually in $F[x]$.

Since $\mathcal{O}_\alpha$ is an orbit of the action of H on K, it is clear that for any $\sigma \in H$, if $\mathcal{O}_\alpha = \{\beta_1, \dots, \beta_m\}$ then $\{\sigma(\beta_1), \dots, \sigma(\beta_m)\} = \mathcal{O}_\alpha$ as well. Now we have $\sigma(f) = \prod_{\beta \in \mathcal{O}_\alpha}(x - \sigma(\beta)) = \prod_{\beta \in \mathcal{O}_\alpha}(x - \beta) = f$. This is true for all $\sigma \in H$. But that means that every coefficient of f is fixed by all $\sigma \in H$. In other words the coefficients of f lie in $\operatorname{Fix}(H) = F$ and so $f \in F[x]$ as claimed.

Since $\alpha \in \mathcal{O}_\alpha$ it is clear that $f(\alpha) = 0$. Let $g = \operatorname{minpoly}_F(\alpha)$. Then for each $\sigma \in H \subseteq \operatorname{Gal}(K/F)$, we know that σ permutes the roots of g in K. In particular $\sigma(\alpha)$ must be a root of g for all $\sigma \in H$. But now every element of $\mathcal{O}_\alpha$ is a root of g, and so $(x - \beta)$ is a factor of g for all $\beta \in \mathcal{O}_\alpha$. This forces $f | g$ and hence $f = g$ since g is irreducible.

(2) By Corollary 17.16, we know there is $\gamma \in K$ such that $K = F(\gamma)$. Now $|\operatorname{Gal}(K/F)| = [K : F] = \deg \operatorname{minpoly}_F(\gamma)$. By part (1), $\deg \operatorname{minpoly}_F(\gamma) = |\mathcal{O}_\gamma| \leq |H|$. This shows that $|\operatorname{Gal}(K/F)| \leq |H|$ and since H is a subgroup of $\operatorname{Gal}(K/F)$ we have $H = \operatorname{Gal}(K/F)$. □

Corollary 17.18. *Let $G = \operatorname{Gal}(K/F)$ for a finite degree Galois extension K/F. If H is a subgroup of G then $\Gamma\Phi(H) = \operatorname{Gal}(K/\operatorname{Fix}(H)) = H$.*

Proof. Since K/F is Galois, we know that $K/\operatorname{Fix}(H)$ is Galois by Corollary 17.12. Now applying Lemma 17.17 to the subgroup H and the extension $K/\operatorname{Fix}(H)$ yields that $H = \operatorname{Gal}(K/\operatorname{Fix}(H))$. □

We also now get an additional characterization of Galois extensions to add to those we found in Theorem 17.10. In fact, the property in the next theorem is sometimes taken to be the definition of a Galois extension.

Theorem 17.19. *Let $F \subseteq K$ be an extension with $[K : F] < \infty$. Then K/F is Galois if and only if $F = \operatorname{Fix}(\operatorname{Gal}(K/F))$.*

Proof. If K/F is Galois, then we saw that $F = \operatorname{Fix}(\operatorname{Gal}(K/F))$ in Lemma 17.15. On the other hand, if $F = \operatorname{Fix}(\operatorname{Gal}(K/F))$ then Lemma 17.17(1) applies with $H = \operatorname{Gal}(K/F)$. For any $\alpha \in K$ the minimal polynomial $\operatorname{minpoly}_F(\alpha)$ calculated there clearly has distinct roots and splits over K. This shows that K/F is normal and separable, and hence it is Galois by Theorem 17.10. □

We now have all of the ingredients we need to prove the fundamental theorem.

Theorem 17.20 (Fundamental Theorem of Galois Theory)**.** *Let $F \subseteq K$ be a finite degree Galois extension of fields. Let $G = \operatorname{Gal}(K/F)$.*

1. *The functions Γ, Φ defined as in (17.14) by*

$$\begin{tikzcd} \left\{ \textit{intermediate fields } E \textit{ with } F \subseteq E \subseteq K \right\} \arrow[r, shift left, "\Gamma=\operatorname{Gal}(K/-)"] & \left\{ \textit{subgroups } H \textit{ of } G = \operatorname{Gal}(K/F) \right\} \arrow[l, shift left, "\Phi=\operatorname{Fix}(-)"] \end{tikzcd}$$

   are inverse inclusion-reversing bijections between the indicated sets.
2. *For $F \subseteq E \subseteq K$, we have $[K : E] = |\operatorname{Gal}(K/E)|$ and $[E : F] = |G : \operatorname{Gal}(K/E)|$.*
3. *For $F \subseteq E \subseteq K$, E/F is normal (and hence Galois) if and only if $H = \operatorname{Gal}(K/E)$ is a normal subgroup of G, and in this case $\operatorname{Gal}(E/F) \cong G/H$ as groups.*

Proof. (1) We have done almost all of the work needed for the proof in the preceding lemmas. We saw that Γ and Φ make sense for a general field extension and that they always reverse inclusions. Now using that K/F is finite degree Galois, by Lemma 17.15 we have $\Phi\Gamma$ is the identity function on intermediate fields. By Lemma 17.18, $\Gamma\Phi$ is the identity function on subgroups of $G = \operatorname{Gal}(K/F)$. Thus Φ and Γ are inverse bijections.

(2) We also saw in Corollary 17.12 that for any intermediate field E, K/E is also Galois. Thus by definition $[K : E] = |\operatorname{Gal}(K/E)|$. Of course we have in particular that $[K : F] = |G| = |\operatorname{Gal}(K/F)|$. Now we have $[E : F] = [K : F]/[K : E] = |G|/|\operatorname{Gal}(K/E)| = |G : \operatorname{Gal}(K/E)|$.

(3) Since K/F is Galois, it is separable and so E/F is separable. This is why if E/F is normal it is automatically Galois as commented.

Let $H = \operatorname{Gal}(K/E)$. It is easy to see that the conjugate $\sigma H \sigma^{-1}$ is equal to $\operatorname{Gal}(K/\sigma(E))$. We see that H is normal in G if and only if $\operatorname{Gal}(K/\sigma(E)) = \operatorname{Gal}(K/E)$ for all $\sigma \in G$. But since Γ is a bijection this is if and only if $\sigma(E) = E$ for all $\sigma \in G$.

Now we claim that $\sigma(E) = E$ for all $\sigma \in G$ if and only if E/F is a normal extension. If E/F is normal and $\sigma \in G$, then for any $\alpha \in E$, $f = \operatorname{minpoly}_F(\alpha)$ splits in $E[x]$, so every root of f in K is already in E. On the other hand, σ must permute the roots of f, so $\sigma(\alpha) \in E$ and thus $\sigma(E) \subseteq E$; applying this argument with σ^{-1} yields $\sigma^{-1}(E) \subseteq E$ and so $E \subseteq \sigma(E)$; thus $\sigma(E) = E$ for all $\sigma \in G$. Conversely, suppose that $\sigma(E) = E$ for all $\sigma \in G$. For $\alpha_1 \in E$, let $f = \operatorname{minpoly}_F(\alpha_1)$. Now f splits in $K[x]$, say with roots $\alpha_1, \ldots, \alpha_m \in K$. By Corollary 16.50, for any i there is $\sigma \in \operatorname{Gal}(K/F)$ such that $\sigma(\alpha_1) = \alpha_i$. By hypothesis since $\alpha_1 \in E$, $\alpha_i \in \sigma(E) = E$. So f splits over E and E/F is normal. This proves the claim.

We have seen that E/F is a normal extension if and only if $H = \operatorname{Gal}(K/E)$ is a normal subgroup of G. Now assume this is the case. The map

$$\phi : \operatorname{Gal}(K/F) \longrightarrow \operatorname{Gal}(E/F)$$

$$\sigma \longmapsto \sigma|_E$$

is well defined because $\sigma(E) = E$ for all $\sigma \in G$. It is easy to see that ϕ is a homomorphism of groups since it is simply defined by restriction. Also, $\ker \phi = H = \operatorname{Gal}(K/E)$ by definition. Then $G/H \cong \phi(G)$ as groups by the 1st isomorphism theorem. The orders satisfy

$$|\phi(G)| = |G|/|H| = [G : \operatorname{Gal}(K/E)] = [E : F] = |\operatorname{Gal}(E/F)|$$

since E/F is Galois, and this forces ϕ to be surjective, so $G/H \cong \operatorname{Gal}(E/F)$. □

17.3. **Examples of the fundamental theorem.**

Example 17.21. Let K be the splitting field over $\mathbb{Q}$ of $f = x^3 - 2 \in \mathbb{Q}[x]$. We have seen that $[K : \mathbb{Q}] = 6$. since K is the splitting field of a separable polynomial, $K/\mathbb{Q}$ is Galois and so $|\operatorname{Gal}(K/\mathbb{Q})| = 6$. In fact, in Example 16.51 we already constructed 6 different automorphisms, but let us revisit this yet again from another perspective.

We have $K = \mathbb{Q}(\alpha, \zeta)$, where $\alpha = \sqrt[3]{2}$ and $\zeta = e^{2\pi i/3}$. Any $\sigma \in G = \operatorname{Gal}(K/\mathbb{Q})$ must permute the roots $\{\zeta, \zeta^2\}$ of $x^2 + x + 1$ and the the roots $\{\alpha, \alpha\zeta, \alpha\zeta^2\}$ of $x^3 - 2$. Moreover, since α and ζ generate K over $\mathbb{Q}$, any automorphism in G is determined by its action on α and ζ. Since there are 3 possible things to send α to and 2 possible things to send ζ to, there are at most 6 different automorphisms in G. Since we know $|G| = 6$, all of these possibilities occur. Thus by applying our general results which tell us that $K/\mathbb{Q}$ is Galois in advance, we do not have to directly construct these automorphisms using the theory of splitting fields, as in Example 16.51.

In particular, there is an automorphism σ with $\sigma(\alpha) = \alpha\zeta$ and $\sigma(\zeta) = \zeta$, and an automorphism ρ with $\rho(\alpha) = \alpha$ and $\rho(\zeta) = \zeta^2$. Then $\sigma\rho(\alpha) = \sigma(\alpha) = \alpha\zeta$ while $\rho\sigma(\alpha) = \rho(\alpha\zeta) = \alpha\zeta^2$. It follows that $\sigma\rho \neq \rho\sigma$ and hence G is non-abelian. The only non-abelian group of order 6 is S_3, so $G \cong S_3$.

Now σ clearly has order 3, so $\langle\sigma\rangle$ is the unique subgroup of order 3 in G. Since this group has index 2 in G, the corresponding intermediate field $\operatorname{Fix}\langle\sigma\rangle$ has degree 2 over $\mathbb{Q}$. Clearly this must be $\operatorname{Fix}\langle\sigma\rangle = \mathbb{Q}(\zeta)$.

The elements of order 2 in the group are then $\rho, \sigma\rho$, and $\sigma^2\rho$. The corresponding fixed fields must have degree 3 over $\mathbb{Q}$. It is now easy to check that $\operatorname{Fix}\langle\rho\rangle = \mathbb{Q}(\alpha)$, $\operatorname{Fix}\langle\sigma\rho\rangle = \mathbb{Q}(\alpha\zeta^2)$, and $\operatorname{Fix}\langle\sigma\rho\rangle = \mathbb{Q}(\alpha\zeta)$. The fields we have found must be all of the intermediate fields strictly between $\mathbb{Q}$ and K, by the fundamental theorem.

Now let us work through a more elaborate example of the fundamental theorem.

Example 17.22. Consider the splitting field of $f = x^4 - 2$ over $\mathbb{Q}$. The 4th roots of 1 in $\mathbb{C}$ are $\{\pm 1, \pm i\}$. Let $\alpha = \sqrt[4]{2}$ be the positive real 4th root of 2. Then the roots of f in $\mathbb{C}$ are $\{\alpha, \alpha i, -\alpha, -\alpha i\}$. It is clear that the splitting field K of f is equal to $\mathbb{Q}(\alpha, i)$.

Now $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$, since f is irreducible by Eisenstein and thus $f = \operatorname{minpoly}_{\mathbb{Q}}(\alpha)$. Since $\mathbb{Q}(\alpha) \subseteq \mathbb{R}$, $\mathbb{Q}(\alpha) \neq K$; since i is a root of the degree 2 polynomial $x^2 + 1 \in \mathbb{Q}[x]$, the only possibility is $[\mathbb{Q}(\alpha, i) : \mathbb{Q}(\alpha)] = 2$ and so $[K : \mathbb{Q}] = 8$.

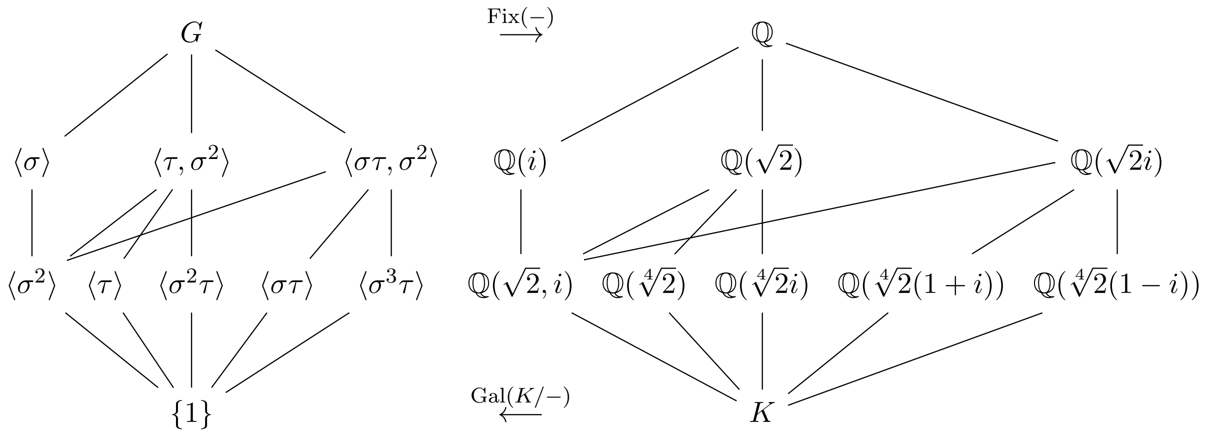
$K/\mathbb{Q}$ is certainly a Galois extension, being the splitting field of a separable polynomial, so $|\operatorname{Gal}(K/\mathbb{Q})| = 8$. Let $G = \operatorname{Gal}(K/\mathbb{Q})$ and let us determine the isomorphism type of G. If $\sigma \in G$, then σ sends α to one of the 4 roots of f and sends i to one of the 2 roots of $\operatorname{minpoly}_{\mathbb{Q}}(i) = x^2 + 1$. Since α and i generate K over $\mathbb{Q}$, any $\sigma \in G$ is determined by where it sends α and i. Since there are only $(4)(2) = 8$ possibilities they must all occur.

Let us call σ the automorphism in G with $\sigma(\alpha) = \alpha i$ and $\sigma(i) = i$. We let τ be the automorphism in G with $\tau(\alpha) = \alpha$ and $\tau(i) = -i$. Now it is easy to see that $|\sigma| = 4$ and $|\tau| = 2$. It is clear that $\langle\sigma\rangle \cap \langle\tau\rangle = \{1\}$ and so we must have $\langle\sigma\rangle\langle\tau\rangle = G$ as this product has order 8. Thus $G = \{1, \sigma, \sigma^2, \sigma^3, \tau, \sigma\tau, \sigma^2\tau, \sigma^3\tau\}$.

Now one easily calculates that $\tau\sigma = \sigma^3\tau = \sigma^{-1}\tau$. We recognize from this relation that G is isomorphic to the dihedral group D_8. From our knowledge of D_8 we can write down all of the subgroups of G. Of course we have the trivial subgroup and all of G; there is the reflection subgroup

$\langle\sigma\rangle$ of order 4, which has a subgroup $\langle\sigma^2\rangle$ of order 2; each reflection generates a subgroup of order 2, and these are the subgroups $\langle\tau\rangle$, $\langle\sigma\tau\rangle$, $\langle\sigma^2\tau\rangle$, and $\langle\sigma^3\tau\rangle$. Any missing subgroups have order 4 and so must intersect $\langle\sigma\rangle$ in a subgroup of order 2, thus containing σ^2. This leads to two further subgroups $\langle\sigma^2,\tau\rangle$ and $\langle\sigma^2,\sigma\tau\rangle$ which are isomorphic to the Klein 4-group.

We display the lattice diagram of these subgroups on the left below, with a line drawn when one is included in the other. For each subgroup H we write the subfield $\mathrm{Fix}(H)$ on the right in the same position; the resulting diagram of all intermediate fields E looks the same, but due to the inclusion-reversing nature of the correspondence, the larger fields are below the fields they contain. For each intermediate field we have written elements that generate that field over $\mathbb{Q}$.


G
$\langle\sigma\rangle$
$\langle\tau,\sigma^2\rangle$
$\langle\sigma\tau,\sigma^2\rangle$
$\langle\sigma^2\rangle$
$\langle\tau\rangle$
$\langle\sigma^2\tau\rangle$
$\langle\sigma\tau\rangle$
$\langle\sigma^3\tau\rangle$
$\{1\}$
$\xrightarrow{\mathrm{Fix}(-)}$
$\xleftarrow{\mathrm{Gal}(K/-)}$
$\mathbb{Q}$
$\mathbb{Q}(i)$
$\mathbb{Q}(\sqrt{2})$
$\mathbb{Q}(\sqrt{2}i)$
$\mathbb{Q}(\sqrt{2},i)$
$\mathbb{Q}(\sqrt[4]{2})$
$\mathbb{Q}(\sqrt[4]{2}i)$
$\mathbb{Q}(\sqrt[4]{2}(1+i))$
$\mathbb{Q}(\sqrt[4]{2}(1-i))$
K

The verification that the fixed fields of each subgroup are what we have displayed is largely routine. For example, the subgroup $H=\langle\sigma\tau\rangle$ has order 2 and hence index 4 in G. Thus $E=\mathrm{Fix}(H)$ has degree 4 over $\mathbb{Q}$. To find elements in $\mathrm{Fix}(H)$, one method is to take any $\gamma\in K$ and note that $\sum_{\rho\in H}\rho(\gamma)$ is fixed by H. (This idea already appeared in the proof of Lemma 17.17.) Applying this to $\alpha=\sqrt[4]{2}$ gives that $\sqrt[4]{2}+\sqrt[4]{2}i=\sqrt[4]{2}(1+i)\in\mathrm{Fix}(\langle\sigma\tau\rangle)$. On the other hand, $(\sqrt[4]{2}(1+i))^4=2(-4)=-8$ and so $\sqrt[4]{2}(1+i)$ is a root of x^4+8. One may check that this polynomial is irreducible over $\mathbb{Q}$, and so $[\mathbb{Q}(\sqrt[4]{2}(1+i)):\mathbb{Q}]=4$. Thus $\mathbb{Q}(\sqrt[4]{2}(1+i))=\mathrm{Fix}(H)$. We leave the other verifications to the reader.

We know by the theorem of the primitive element that every intermediate field can be generated by one element over $\mathbb{Q}$; for most of the intermediate fields above this has already been done. Suppose we want to write $K=\mathbb{Q}(\sqrt[4]{2},i)$ in the form $\mathbb{Q}(\gamma)$ for some γ. It suffices to choose a γ which is not contained in any of the 5 displayed fields that have index 4 over $\mathbb{Q}$. This could be done by showing it is not fixed by any of the 5 elements generating order 2 subgroups of G. For example, $\sigma\tau(\sqrt[4]{2}+i) = \sqrt[4]{2}i - i = (\sqrt[4]{2}-1)i$, which has 0 real part and so certainly is not equal to $\sqrt[4]{2}+i$. Similarly, σ^2, τ, $\sigma^2\tau$, and $\sigma^3\tau$ do not fix $\sqrt[4]{2}+i$, so $K = \mathbb{Q}(\sqrt[4]{2}+i)$.

We may also easily determine which subfields are normal and hence Galois over $\mathbb{Q}$. The normal subgroups of G include the three subgroups of index 2 and the subgroup $\langle\sigma^2\rangle$ (which is actually the center of D_8). The other 4 subgroups of order 2 generated by the reflections are not normal. Thus the intermediate fields which are normal over $\mathbb{Q}$ are $\mathbb{Q}(i)$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{2}i)$ and $\mathbb{Q}(\sqrt{2}, i)$. It is also easy to see directly that these are all splitting fields over $\mathbb{Q}$. On the the other hand, for example, $\mathbb{Q}(\sqrt[4]{2})$ is not normal over $\mathbb{Q}$ since it contains one root of x^4-2 but this polynomial does not split over that field. Similarly, $\mathbb{Q}(\sqrt[4]{2}(1+i))$ contains one root of x^4+8 but that polynomial does not split over it. In fact, it is easy to see that K is also the splitting field over $\mathbb{Q}$ of x^4+8.

Finally, for any field E which is normal over $\mathbb{Q}$, the fundamental theorem tells us that $\mathrm{Gal}(E/\mathbb{Q}) \cong \mathrm{Gal}(K/\mathbb{Q})/\mathrm{Gal}(K/E)$. For instance, taking $E = \mathbb{Q}(\sqrt{2}, i)$ we must have $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, i)/\mathbb{Q}) \cong D_8/\langle\sigma^2\rangle$. It is also easy to see directly that both sides are Klein 4-groups.

We observe that it was easy to compute the subgroups of the Galois group above, and relatively easy then to find the fixed fields. Without Galois theory it is extremely unclear how one would go about finding all of the intermediate fields in the extension, or even why there should be finitely many.

Let us give an example of a Galois group over $\mathbb{Q}$ where determining the group is a bit less straightforward.

Example 17.23. Let $\alpha = \sqrt{2+\sqrt{2}}$. First let us calculate its minimal polynomial over $\mathbb{Q}$. We have $\alpha^2 = 2+\sqrt{2}$ so $(\alpha^2-2)^2 = 2$. Then α is a root of $f = (x^2-2)^2 - 2 = x^4 - 4x^2 + 2$. This polynomial is irreducible by the Eisenstein criterion, so $f = \mathrm{minpoly}_{\mathbb{Q}}(\alpha)$. Then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$.

We claim that $K = \mathbb{Q}(\alpha)$ is already the splitting field of f over $\mathbb{Q}$. First, applying the quadratic formula to $x^2 - 4x + 2 = 0$ yields roots $2 \pm \sqrt{2}$. Thus $x^4 - 4x^2 + 2 = 0$ has roots $\pm\sqrt{2 \pm \sqrt{2}}$, where all of these roots are real. Write $\beta = \sqrt{2-\sqrt{2}}$, so the roots are $\pm\alpha, \pm\beta$.

Now notice that $\sqrt{2-\sqrt{2}}\sqrt{2+\sqrt{2}} = \sqrt{(2-\sqrt{2})(2+\sqrt{2})} = \sqrt{4-2} = \sqrt{2}$. Moreover, $\sqrt{2} = \alpha^2 - 2$ is already in K. Hence $\beta = \sqrt{2-\sqrt{2}} \in K$ already and so all of the roots of f are in K. Thus K is the splitting field of f as claimed.

Let $G = \mathrm{Gal}(K/\mathbb{Q})$. Since K is a splitting field of a polynomial over $\mathbb{Q}$, $K/\mathbb{Q}$ is Galois and so $|G| = 4$. Either $G \cong \mathbb{Z}_4$ or $G \cong \mathbb{Z}_2 \times \mathbb{Z}_2$ and we would like to determine which occurs.

We know from our results on splitting fields that we can find $\sigma \in G$ that sends α to any other root of f. Let us choose $\sigma \in G$ with $\sigma(\alpha) = \beta$. How does σ act on the other roots of f? Well,

$\sigma(\alpha^2) = \beta^2$. This says $\sigma(2+\sqrt{2}) = 2-\sqrt{2}$. Then clearly $\sigma(\sqrt{2}) = -\sqrt{2}$. Now since $\alpha\beta = \sqrt{2}$, $\sigma(\alpha\beta) = -\alpha\beta$ which implies $\sigma(\beta) = -\alpha$. We see from this that $\sigma^2(\alpha) = -\alpha$ and thus σ cannot have order 2. Hence σ has order 4 and thus G is cyclic of order 4.

Now G has only one proper subgroup, which is $\langle\sigma^2\rangle$. The corresponding intermediate field is $\mathbb{Q}(\sqrt{2}) = \mathrm{Fix}\langle\sigma^2\rangle$.

For any splitting field K of a separable polynomial f over F, it is possible to visualize the Galois group as a subgroup of the permutation group of the roots of f in K; indeed, this is how Galois originally thought about it. This is because any automorphism in $\mathrm{Gal}(K/F)$ must permute these roots and is determined by its action on these roots. For instance, in the previous example one notes that σ acts as a 4-cycle on the roots of $x^4 - 4x^2 + 2$, which is another way of seeing that it is cyclic of order 4.

17.4. **Cyclotomic extensions.** The set $\mu_n = \{e^{2m\pi i/n} | 0 \le m \le n-1\}$ consists of the n distinct nth roots of 1 in $\mathbb{C}$. This can also be described as the set of roots in $\mathbb{C}$ of $x^n - 1 \in \mathbb{Q}[x]$. The group μ_n is a subgroup of the multiplicative group $\mathbb{C}^\times$.

The map $(\mathbb{Z}_n, +) \to \mu_n$ given by $\overline{m} \mapsto e^{2m\pi i/n}$ is easily seen to be an isomorphism of groups. Thus μ_n is cyclic, and any generator of μ_n is called a *primitive* nth root. From our knowledge of cyclic groups we know that the generators of $\mathbb{Z}_n$ are $\{\overline{m} | \gcd(m,n) = 1\}$ and so the primitive nth roots are $\{e^{2m\pi i/n} | \gcd(m,n) = 1\}$. There are $\varphi(n)$ of them, where φ is the Euler φ-function.

Definition 17.24. Let $P_n = \{e^{2m\pi i/n} | \gcd(m,n) = 1\}$ be the set of primitive nth roots of 1. Let

$$\Phi_n(x) = \prod_{\alpha \in P_n} (x - \alpha) \in \mathbb{C}[x].$$

$\Phi_n(x)$ is called the nth *cyclotomic polynomial.* Clearly $\deg \Phi_n(x) = \varphi(n)$.

If $\alpha \in \mu_n$, then α has order d in $\mathbb{C}^\times$ for some $d|n$, and then α is a primitive dth root. Thus the following formula is clear.

Lemma 17.25. *Let $n \ge 1$. Then $x^n - 1 = \prod_{d|n} \Phi_d(x)$.*

Using this lemma we can compute the polynomials $\Phi_n(x)$ inductively. This is easy to do by hand if n is small.

Example 17.26. 1 is the only primitive 1st root, so $\Phi_1(x) = x - 1$. Similarly, -1 is the lone primitive 2nd root and $\Phi_2(x) = x + 1$. By Lemma 17.25 we have $x^3 - 1 = \Phi_3(x)\Phi_1(x)$ and so

$\Phi_3(x) = (x^3 - 1)/(x - 1) = x^2 + x + 1$. Next,

$$\Phi_4(x) = (x^4 - 1)/(\Phi_2(x)\Phi_1(x)) = (x^4 - 1)/((x + 1)(x - 1)) = (x^4 - 1)/(x^2 - 1) = x^2 + 1.$$

Of course, $x^2 + 1 = (x - i)(x + i)$, and the roots of $\Phi_4(x)$ are the two primitive 4th roots of unity, i and $-i$. We leave it to the reader to check that $\Phi_6(x) = x^2 - x + 1$.

If p is prime then $\Phi_p(x) = (x^p - 1)/(x - 1) = x^{p-1} + \cdots + x + 1$. We proved this polynomial is irreducible in $\mathbb{Q}[x]$ using the Eisenstein criterion with substitution. In fact, all of the polynomials $\Phi_n(x)$ lie in $\mathbb{Z}[x]$ and are irreducible over $\mathbb{Q}$. We give the proof of irreducibility now, though as it is a bit technical the reader might wish to simply assume this fact and move on.

Theorem 17.27. *For any $n \geq 1$, $\Phi_n(x) \in \mathbb{Z}[x]$, it is monic, and it is irreducible in $\mathbb{Q}[x]$.*

Proof. By definition it is obvious that $\Phi_n(x)$ is monic. We have $x^n - 1 = \Phi_n(x) \prod_{d|n, d<n} \Phi_d(x)$ by Lemma 17.25. We prove that $\Phi_n(x) \in \mathbb{Z}[x]$ by induction on n. Thus we can assume that $\Phi_d(x) \in \mathbb{Z}[x]$, for all divisors d of n with $d < n$, by the induction hypothesis. Then $g = \prod_{d|n, d<n} \Phi_d(x)$ is also monic. By Gauss's lemma, there is $\lambda \in \mathbb{Q}$ such that $\lambda^{-1}\Phi_n(x)$ and λg are in $\mathbb{Z}[x]$. Since g and $\Phi_n(x)$ are monic, this forces λ and $\lambda^{-1} \in \mathbb{Z}$, so $\lambda = \pm 1$ and $\Phi_n(x) \in \mathbb{Z}[x]$ already.

Let f be one of the irreducible factors of $\Phi_n(x)$ in $\mathbb{Q}[x]$ and write $\Phi_n(x) = fg$. By Gauss's Lemma again, we can choose this factorization so f and g are in $\mathbb{Z}[x]$ and are monic. Now pick any prime p with $\gcd(p, n) = 1$. The idea of the proof is to consider this factorization of $\Phi_n(x)$ modulo p.

Let ζ be a primitive nth root of 1; since $\gcd(p, n) = 1$, we also have ζ^p is a primitive nth root. Thus ζ and ζ^p are roots of $\Phi_n(x)$ and each is a root of either f or g. Suppose that $f(\zeta) = 0$, while $g(\zeta^p) = 0$.

Now $g(\zeta^p) = 0$ means that ζ is a root of $g(x^p) \in \mathbb{Z}[x]$. Since f is irreducible, $f = \text{minpoly}_{\mathbb{Q}}(\zeta)$. Hence $f|g(x^p)$ in $\mathbb{Q}[x]$, say $g(x^p) = fh$. As above, h is monic and by Gauss's lemma we have $h \in \mathbb{Z}[x]$.

Let $\phi : \mathbb{Z}[x] \to (\mathbb{Z}/p\mathbb{Z})[x] = \mathbb{F}_p[x]$ be the reduction mod p homomorphism which sends each coefficient $a \in \mathbb{Z}$ to $\overline{a} = a + p\mathbb{Z}$. For $f \in \mathbb{Z}[x]$ write $\overline{f}$ for $\phi(f)$. Now applying ϕ we have $\overline{g}(x^p) = \overline{f}\,\overline{h}$. If we write $\overline{g}(x) = \sum_{i=0}^m a_i x^i$, with $a_i \in \mathbb{F}_p$, then since the pth power map is a ring homomorphism of $\mathbb{F}_p[x]$, we have

$$\overline{g}(x^p) = \sum_{i=0}^m a_i x^{ip} = \sum_{i=0}^m a_i^p x^{ip} = \sum_{i=0}^m (a_i x^i)^p = (\sum_{i=0}^m a_i x^i)^p = (\overline{g}(x))^p,$$

where we have used that $a^p = a$ for all $a \in \mathbb{F}_p$ by Fermat's little theorem.

We now see that $\overline{f} \,|\, (\overline{g})^p$ in $\mathbb{F}_p[x]$. This implies that $\overline{f}$ and $\overline{g}$ have a common irreducible factor in $\mathbb{F}_p[x]$. But since $\overline{\Phi_n} = \overline{f}\,\overline{g}$ this means that $\overline{\Phi_n}$ has a repeated irreducible factor in $\mathbb{F}_p[x]$, and so it is not a separable polynomial. On the other hand, $\overline{\Phi_n}$ divides $\overline{x^n - 1}$, which is a separable polynomial in $\mathbb{F}_p[x]$: its derivative is $\overline{n}x^{n-1} \neq 0$ (because $\gcd(p,n) = 1$), and so $\gcd(x^n - 1, \overline{n}x^{n-1}) = 1$ in $\mathbb{F}_p[x]$. This is a contradiction.

The contradiction implies that for all roots ζ of f and all primes p with $\gcd(p,n) = 1$, ζ^p must also be a root of f. Now if $\gcd(i,n) = 1$ for some integer i, then factorizing $i = p_1 p_2 \dots p_k$ where each p_j is prime, then $\gcd(p_j, n) = 1$ for all j and so by induction we get for any root ζ of f that ζ^i is also a root of f. However, any root ζ of f is by definition a generator of the group μ_n of nth roots of 1, and the other generators are equal to ζ^i for $0 < i < n$ with $\gcd(i,n) = 1$. Hence every primitive nth root of 1 is a root of f. This shows that $\Phi_n(x) = f$ and hence $\Phi_n(x)$ is irreducible over $\mathbb{Q}$. □

Theorem 17.28. *Let $n \geq 1$. Consider the nth cyclotomic field $K = \mathbb{Q}(\zeta)$ where ζ is a primitive nth root of 1. Then $[K : \mathbb{Q}] = \varphi(n)$, and $K/\mathbb{Q}$ is Galois with $\mathrm{Gal}(K/\mathbb{Q}) \cong \mathbb{Z}_n^*$, where $\mathbb{Z}_n^*$ is the multiplicative group of units mod n.*

Proof. We know that K is the splitting field of $x^n - 1$ over $\mathbb{Q}$ by Example 16.44. We also know from Theorem 17.27 that $\Phi_n(x)$ is irreducible over $\mathbb{Q}$; hence $\Phi_n(x) = \mathrm{minpoly}_{\mathbb{Q}}(\zeta)$, which implies $[\mathbb{Q}(\zeta) : \mathbb{Q}] = \deg \Phi_n(x) = \varphi(n)$. $K/\mathbb{Q}$ is Galois since it is a splitting field of a separable polynomial.

Now any $\sigma \in G = \mathrm{Gal}(K/\mathbb{Q})$ is determined by its action on ζ, and $\sigma(\zeta)$ must be another root of $\Phi_n(x)$. Since this polynomial has $\varphi(n)$ roots and $|G| = \varphi(n)$, all possibilities occur. For each $0 \leq i \leq n-1$ with $\gcd(i,n) = 1$, we have an automorphism $\sigma_i \in G$ where $\sigma_i(\zeta) = \zeta^i$, so $G = \{\sigma_i \,|\, 0 \leq i \leq n-1,\ \gcd(i,n) = 1\}$. Now define a function $\phi : G \to \mathbb{Z}_n^*$ by $\phi(\sigma_i) = \overline{i}$. Since $\sigma_i\sigma_j(\zeta) = \sigma_i(\zeta^j) = \sigma_i(\zeta)^j = (\zeta^i)^j = \zeta^{ij} = \zeta^m$, where m is unique integer with $0 \leq m \leq n-1$ such that $m \equiv ij \mod n$, we have $\sigma_i\sigma_j = \sigma_m$ where $\overline{ij} = \overline{m}$. This implies that ϕ is a homomorphism of groups, and it is clearly bijective. □

Recall from our study of groups that the structure of the group $\mathbb{Z}_n^*$ is well-understood. First, if $n = p_1^{e_1} \dots p_m^{e_m}$ for distinct primes p_i, then $\mathbb{Z}_n^* \cong \prod_i \mathbb{Z}_{p_i^{e_i}}^*$. If p is an odd prime, then $\mathbb{Z}_{p^e}^*$ is cyclic of order $\varphi(p^e) = p^{e-1}(p-1)$; while $\mathbb{Z}_{2^e}^* \cong \mathbb{Z}_2 \times \mathbb{Z}_{2^{e-2}}$.

Example 17.29. Let $n = 9$ and consider the splitting field K of $x^n - 9$ over $\mathbb{Q}$. We know that $[K : Q] = \varphi(9) = 6$ and if $G = \mathrm{Gal}(K/\mathbb{Q})$ then $G \cong \mathbb{Z}_9^*$, which is cyclic order 6.

Let ζ be a primitive 9th root of unity, so $K = \mathbb{Q}(\zeta)$. One may check that $\overline{2}$ is a generator of $\mathbb{Z}_9^*$. It follows that the automorphism $\sigma \in G$ with $\sigma(\zeta) = \zeta^2$ generates G. So $G = \langle \sigma \rangle$, and by the stucture of cyclic groups of order 6, the subgroups of G are $\langle \sigma^2 \rangle$, $\langle \sigma^3 \rangle$, and the trivial subgroup.

We have a diagram of subgroups and and corresponding diagram of fixed subfields as follows, where the numbers indicate the index of one subgroup in another (on the left) or the degree of one subfield over another (on the right).

$$\begin{tikzcd}
& \langle\sigma\rangle \arrow[dl, dash, "2"'] \arrow[dr, dash, "3"] & \\
\langle\sigma^2\rangle \arrow[dr, dash, "3"'] & & \langle\sigma^3\rangle \arrow[dl, dash, "2"] \\
& \langle\sigma^6\rangle = \{1\} &
\end{tikzcd}$$

$$\begin{tikzcd}
& \mathbb{Q} \arrow[dl, dash, "2"'] \arrow[dr, dash, "3"] & \\
\mathbb{Q}(\alpha) \arrow[dr, dash, "3"'] & & \mathbb{Q}(\beta) \arrow[dl, dash, "2"] \\
& \mathbb{Q}(\zeta) = K &
\end{tikzcd}$$

Let us find elements α and β which generate the indicated extensions on the right. Note that since ζ is a primitive 9th root, ζ^3 is a primitive 3rd root. Thus K also contains the cyclotomic extension $\mathbb{Q}(\zeta^3)$, where the minimal polynomial of ζ^3 is $x^2 + x + 1$. Thus $\mathbb{Q}(\zeta^3)$ is a field of degree 2 over $\mathbb{Q}$, so we can take $\alpha = \zeta^3$ in the picture, as there is only one subfield of degree 2 over $\mathbb{Q}$. For the other extension we note that $\zeta + \sigma^3(\zeta) = \zeta + \zeta^8 = \zeta + \zeta^{-1}$ is fixed by σ^3, so $\zeta + \zeta^{-1} \in \operatorname{Fix}(\langle \sigma^3 \rangle)$. $\zeta + \zeta^{-1}$ is not in $\mathbb{Q}$ (for if it was, $\zeta + \zeta^{-1} = q \in \mathbb{Q}$ would give $\zeta^2 + 1 = q\zeta$ and hence ζ would satisfy a degree 2 polynomial in $\mathbb{Q}[x]$, while we know that $\deg \operatorname{minpoly}_{\mathbb{Q}}(\zeta) = 6$). Thus we can take $\beta = \zeta + \zeta^{-1}$ in the picture.

17.5. **More on finite fields.** Recall that we have seen that there is a unique field $\mathbb{F}_{p^n}$ with p^n elements up to isomorphism, for any prime p and $n \geq 1$. It can be defined as the splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$. We can now easily calculate the Galois group of this field as an extension of $\mathbb{F}_p$.

Theorem 17.30. *Let K be a field with $|K| = p^n$ for a prime p and $n \geq 1$. Then $\mathbb{F}_p \subseteq K$ and $K/\mathbb{F}_p$ is Galois with $[K : \mathbb{F}_p] = n$ and $G = \operatorname{Aut}(K) = \operatorname{Gal}(K/\mathbb{F}_p) \cong \mathbb{Z}_n$. In particular, G is generated as a group by the Frobenius automorphism $\sigma : K \to K$ given by $\sigma(a) = a^p$.*

Proof. We know that K has characteristic p, and we have seen that the additive subgroup of K generated by 1 is a copy of the field $\mathbb{F}_p$. Any automorphism of K sends 1 to itself and so will fix the elements in $\mathbb{F}_p$, which are sums of 1. Thus $\operatorname{Aut}(K) = \operatorname{Gal}(K/\mathbb{F}_p)$.

We know that $K^\times$ is a cyclic group of order $p^n - 1$. Let $\gamma \in K$ be a generator of this group. Then $\gamma^{p^n - 1} = 1$ and so $\gamma^{p^n} = \gamma$. Conversely, if $\gamma^j = \gamma$ for some $j > 1$ then $\gamma^{j-1} = 1$ and hence $j \geq p^n$ since γ has order $p^n - 1$.

Now the Frobenius map σ is an automorphism of K, as in Example 16.63. We have $\sigma^i(\gamma) = \gamma^{p^i}$ and this cannot equal γ for $0 < i < n$; so $\sigma^i \neq 1_K$. On the other hand $\sigma^n(\gamma) = \gamma$ and since $K = \mathbb{F}_p(\gamma)$, $\sigma^n = 1$. It follows that σ is an element of order n in G. Since $|G| = [K : \mathbb{F}_p] = n$, the Frobenius map σ must generate G and hence $G \cong (\mathbb{Z}_n, +)$. □

Corollary 17.31. *Let K be a field with $|K| = p^n$ for some prime p and $n \geq 1$. Then K has a unique subfield with p^d elements for each positive divisor d of n. These are the only subfields of K.*

Proof. $G = \mathrm{Gal}(K/\mathbb{F}_p) = \langle\sigma\rangle$, where σ, the Frobenius map, has order n in G. Since G is cyclic of order n, for each divisor d of n there is a unique subgroup $H_d = \langle\sigma^d\rangle$ with $[G : H_d] = d$. Then the fields E_d where $\mathbb{F}_p \subseteq E_d = \mathrm{Fix}(H_d) \subseteq K$ are the only intermediate fields of the extension $K/\mathbb{F}_p$, where $[E_d : \mathbb{F}_p] = d$ and hence $|E_d| = p^d$. In fact these E_d are all of the subfields of K, because every subfield of K must contain the prime subfield $\mathbb{F}_p$. There is one for each divisor d of n. □

We can describe the subfield E_d of order p^d inside a field K of order p^n more explicitly: Since E_d has order p^d, all $a \in E_d$ must satisfy $a^{p^d} = a$, because we saw in our original study of finite fields that the elements of E_d are all roots of $x^{p^d} - x$. Since that polynomial only has p^d roots, the elements in E must be all of its roots. Thus $E = \{a \in K | a^{p^d} = a\}$.

One elegant consequence of our results so far is the following description of the factorization of $x^{p^n} - x$ over $\mathbb{F}_p$.

Proposition 17.32. $x^{p^n} - x \in \mathbb{F}_p[x]$ *is the product of all monic irreducible polynomials of degree d over $\mathbb{F}_p$, for all divisors d of n.*

Proof. If $f \in \mathbb{F}_p[x]$ is monic and irreducible of degree d, where $d|n$, then $K = \mathbb{F}_p[x]/(f)$ is a field with p^d elements. We know then by Corollary 17.31 that the field $\mathbb{F}_{p^n}$ has a subfield isomorphic to this field. Since every element of $\mathbb{F}_{p^n}$ satisfies $a^{p^n} = a$, the same must be true of the elements of K. In particular, $(x + (f))^{p^n} = x^{p^n} + (f) = x + (f)$ in K, which means $x^{p^n} - x \in (f)$. In other words, f divides $x^{p^n} - x$.

Conversely, if $g \in \mathbb{F}_p[x]$ is any monic irreducible factor of $x^{p^n} - x$, then we know that g splits over $\mathbb{F}_{p^n}$ since this field is the splitting field of $x^{p^n} - x$. If $\alpha \in \mathbb{F}_{p^n}$ is a root of g, then $\mathbb{F}_p(\alpha) \subseteq \mathbb{F}_{p^n}$ where $[\mathbb{F}_p(\alpha) : \mathbb{F}_p] = \deg g$ because $g = \mathrm{minpoly}_{\mathbb{F}_p}(\alpha)$. It follows that $|\mathbb{F}_p(\alpha)| = p^{\deg g}$. Since we have seen that all subfields of a field with p^n elements have p^d elements for some divisor d of n, $\deg g = d$ for some divisor d of n.

Finally, we know that $x^{p^n} - x$ is separable over $\mathbb{F}_p$, as we showed that its roots in $\mathbb{F}_{p^n}$ are the p^n distinct elements of $\mathbb{F}_{p^n}$. It follows that $x^{p^n} - x$ is a product of distinct irreducible polynomials. By the arguments above the irreducibles occurring are exactly those of degree d where $d|n$. □

The proposition can be used to give a precise count of the number of irreducible polynomials of each degree n over $\mathbb{F}_p$, by induction. We omit the exact formula here, but demonstrate the idea in an example.

Example 17.33. Consider $x^{81} - x \in \mathbb{F}_3[x]$. Here $81 = 3^4$. By the proposition, $x^{81} - x$ is the product of all monic irreducible polynomials in $\mathbb{F}_3[x]$ of degree $1, 2$, or 4. We know the degree 1 monic polynomials are $(x - 1)$, $(x - 2)$, and $(x - 4)$. The degree 2 monic irreducibles are those without a root in $\mathbb{F}_3$; these are $x^2 + 1$, $x^2 + 2x + 2$, and $x^2 + x + 2$ by direct calculation. The product of these 6 polynomials has degree 9. That means there is a polynomial of degree $81 - 9 = 72$ left over in the factorization of $x^{81} - x$, which is a product of all distinct monic degree 4 irreducibles. There are thus $72/4 = 18$ distinct such irreducibles over $\mathbb{F}_3$.

17.6. **Root Extensions.** A very common kind of field extension is "adding an nth root of an existing element". We have seen many examples of this already. In other words, one has a field F and an element $a \in F$ such that $f = x^n - a \in F[x]$ does not split already over F. In a splitting field K for f there is a root $\alpha \in K$ of f and we can consider the extension $F \subseteq F(\alpha)$ inside K. Since $\alpha^n = a \in F$, we think of α as an nth root of a, and might loosely write $\alpha = \sqrt[n]{a}$, though this notation is not uniquely defined, as a may have as many as n different nth roots. We are going to see in the next results that extensions of this kind are closely related to cyclic Galois groups. Technically, extensions of this form are simplest when $x^n - 1$ already splits in the base field with distinct roots, and we will concentrate on that case.

We first see that adding an nth root gives a cyclic Galois group (when the base field already has enough roots of 1).

Proposition 17.34. *Let F be a field such that $x^n - 1 \in F[x]$ splits with distinct roots in F. Suppose that $F \subseteq K$ is a field extension and $\alpha \in K$ is a root of $f = x^n - a \in F[x]$. Then $F(\alpha)/F$ is a Galois extension and* $\mathrm{Gal}(F(\alpha)/F)$ *is cyclic of order d for some divisor d of n.*

Proof. The set of roots of $x^n - 1$ in F is a finite multiplicative subgroup of $F^\times$. Since we are assuming this polyonomial splits with distinct roots in F, this is a subgroup of order n. We have seen that a finite subgroup of the multplicative group of a field is always cyclic in Corollary 16.66, so this group is cyclic generated by some ζ, say; thus $\{1, \zeta, \zeta^2 \dots, \zeta^{n-1}\}$ is the set of roots of $x^n - 1$ in F.

Now $\alpha^n = a$ and so clearly $(\alpha\zeta^i)^n = a$ also for all i; thus $x^n - a$ has the n distinct roots $\{\alpha, \alpha\zeta, \dots, \alpha\zeta^{n-1}\}$ in K. Since by assumption $\zeta \in F$, we see that all of these roots are contained in $F(\alpha)$ already. Thus $F(\alpha)$ is the splitting field of f over F; and f has distinct roots and so is a separable polynomial. Thus $F(\alpha)/F$ is a Galois extension.

Now consider any $\sigma \in G = \mathrm{Gal}(F(\alpha)/F)$. Since σ permutes the roots of $x^n - a \in F[x]$, we have $\sigma(\alpha) = \alpha\zeta^i$. This allows us to define a map $\phi : G \to (\mathbb{Z}_n, +)$ defined by $\phi(\sigma) = \bar{i}$, where i is any integer such that $\sigma(\alpha) = \alpha\zeta^i$. Note that if we express $\sigma(\alpha)$ as $\alpha\zeta^j$ for some other j, we will have $\zeta^i = \zeta^j$ and thus since ζ has order n, $\bar{i} = \bar{j} \in \mathbb{Z}_n$, so that ϕ is well-defined.

Now to see that ϕ is a homomorphism, we just note that if $\sigma, \tau \in G$, with $\sigma(\alpha) = \alpha\zeta^i$ and $\tau(\alpha) = \alpha\zeta^j$, then

$$\tau\sigma(\alpha) = \tau(\alpha\zeta^i) = \tau(\alpha)\tau(\zeta)^i = \tau(\alpha)\zeta^i = \alpha\zeta^j\zeta^i = \alpha\zeta^{i+j},$$

where we use that $\zeta \in F$ and that $\tau \in \mathrm{Gal}(F(\alpha)/F)$ fixes F. This shows that $\phi(\tau\sigma) = \overline{i+j} = \bar{i} + \bar{j} = \phi(\tau) + \phi(\sigma)$ and thus ϕ is a homomorphism.

Since any $\sigma \in G$ is determined by where it sends α, it follows that ϕ is injective. Thus G is isomorphic to a subgroup of the cyclic group $\mathbb{Z}_n$, and thus from our classification of subgroups of cyclic groups, we conclude that G is cyclic of order d for some divisor d of n. □

Example 17.35. One really can get a proper subgroup of $\mathbb{Z}_n$ in the previous theorem, because there is no requirement that $f = x^n - a$ be irreducible over F. (If f does happen to be irreducible, then $[F(\alpha) : F] = n = |G|$ and this does force $G \cong \mathbb{Z}_n$).

Here is an explicit example. Let K be the splitting field of $f = x^8 - 2$ over $\mathbb{Q}$. Let ζ be a primitive 8th root of 1 in $\mathbb{C}$, and let $\alpha = \sqrt[8]{2}$ be the positive 8th root of 2. Then the roots of f in $\mathbb{C}$ are $\{\alpha, \alpha\zeta, \dots, \alpha\zeta^7\}$ and $K = \mathbb{Q}(\zeta, \alpha)$.

Of course $\mathbb{Q}$ does not contain 8 distinct roots of $x^8 - 1$, but $F = \mathbb{Q}(\zeta)$ does, and so Proposition 17.34 applies to the extension $F \subseteq F(\alpha) = \mathbb{Q}(\zeta, \alpha) = K$. By that proposition, $G = \mathrm{Gal}(K/F)$ is cyclic of order dividing 8.

Let us now calculate $[K : \mathbb{Q}]$. Explicitly we have $\zeta = e^{2\pi i/8} = \sqrt{2}/2 + (\sqrt{2}/2)i \subseteq \mathbb{Q}(\sqrt{2}, i)$, and it is easy to see that $[\mathbb{Q}(\sqrt{2}, i) : \mathbb{Q}] = 4$. On the other hand, we know that $[\mathbb{Q}(\zeta) : \mathbb{Q}] = \varphi(8) = 4$ by the theory of cyclotomic extensions. (In fact one may easily calculate that $\mathrm{minpoly}_{\mathbb{Q}}(\zeta) = \Phi_8(x) = x^4 + 1$). This shows that $\mathbb{Q}(\zeta) = \mathbb{Q}(\sqrt{2}, i)$.

We do know that $x^8 - 2$ is irreducible over $\mathbb{Q}$ by the Eisenstein criterion, and so $x^8 - 2 = \operatorname{minpoly}_{\mathbb{Q}}(\alpha)$ and hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 8$. Now

$$K = \mathbb{Q}(\alpha, \zeta) = \mathbb{Q}(\sqrt[8]{2}, i, \sqrt{2}) = \mathbb{Q}(\sqrt[8]{2}, i)$$

from which it is straightforward to see that $[K : \mathbb{Q}] = 16$, since $i \not\in \mathbb{Q}(\sqrt[8]{2})$. This implies that $[K : F] = 4$ and so $\operatorname{Gal}(K/F) \cong \mathbb{Z}_4$.

In retrospect, it is easy to see explicitly that $f = x^8 - 2$ is not irreducible over $F = \mathbb{Q}(\zeta)$. Since $\sqrt{2} \in F$, f factors as $(x^4 - \sqrt{2})(x^4 + \sqrt{2})$ in $F[x]$.

The perhaps more surprising fact is that when a Galois extension has a cyclic Galois group, and the base field has enough roots of 1 already, then the extension must be given by adjoining an nth root. The standard proof relies on a result known as "linear independence of group characters", a result with other applications in field theory which we would certainly present as well if we had more time. Here, for simplicity we give a proof which relies on some of the techniques of modules over PIDs we already have on hand.

Proposition 17.36. *Let $F \subseteq K$ be a an extension with $[K : F] < \infty$. Assume that $x^n - 1$ splits in F with distinct roots. Suppose that K/F is Galois with* $\operatorname{Gal}(K/F)$ *cyclic of order dividing n. Then $K = F(\alpha)$ for some $\alpha \in K$ such that $\alpha^n \in F$.*

Proof. By assumption $G = \operatorname{Gal}(K/F)$ is cyclic of order d where d divides n. Since K/F is Galois, $d = [K : F]$ also. Let σ be a generator of G, so σ has order d. Note that σ preserves addition and if $\lambda \in F, a \in K$, then $\sigma(\lambda a) = \sigma(\lambda)\sigma(a) = \lambda\sigma(a)$, since σ fixes F. It follows that σ is an F-linear transformation of the n-dimensional vector space K over F. As such, there is an associated $F[x]$-module structure on K, where x acts as σ, and we now investigate this module using our results on modules over a PID.

Since $\sigma^d = 1$, σ satisfies the polynomial $x^d - 1$. Thus the minimal polynomial of σ divides $x^d - 1$. Since d divides n and F has n distinct nth roots of 1, F has d distinct dth roots of 1 already as well. Thus $x^d - 1$ factors as $(x - 1)(x - \rho) \dots (x - \rho^{d-1})$ for some primitive dth root ρ of 1 in F, where $1, \rho, \dots \rho^{d-1} \in F$ are all distinct.

Let $a_1, \dots, a_m \in F[x]$ be the invariant factors of K as an $F[x]$-module, so that

$$K \cong F[x]/(a_1) \oplus \dots \oplus F[x]/(a_m)$$

as $F[x]$-modules, where $a_i | a_{i+1}$ for all i. We have seen that the largest invariant factor a_m is the minimal polynomial of σ. Now a_m divides $x^d - 1$ and thus a_m factors in $F[x]$ as a product of distinct linear factors. Then the same is true of all a_i. Now the elementary divisors of the module are found by factoring each invariant factor as a product of powers of distinct irreducibles. So in this case we see that all elementary divisors have degree 1. It follows that σ has a Jordan form over F, and moreover this Jordan form is diagonal, with diagonal entries which are dth roots of 1. Since σ has order d and not smaller order, one of the diagonal entries has to be a primitive dth root of 1. Without loss of generality we can assume ρ is one of the entries. This tells us that σ has an eigenvector in K, say $\alpha \in K$, with eigenvalue ρ. Thus $\sigma(\alpha) = \rho\alpha$.

Up until now we have not used that σ is an automorphism of K, i.e. that σ preserves multiplication. Now we note that $\sigma(\alpha^i) = (\sigma(\alpha))^i = \rho^i\alpha^i$. Thus $\alpha^i \in K$ is an eigenvector of σ with eigenvalue ρ^i, and we conclude that all powers of ρ, and hence all dth roots of 1, are eigenvalues of σ. Also, since $1, \alpha, \alpha^2, \dots \alpha^{d-1}$ are eigenvectors with distinct eigenvalues, they are linearly independent over F. Since $[K : F] = d$, these powers of α form a basis of K over F. With respect to this basis σ has diagonal entries $1, \rho, \dots, \rho^{d-1}$ and so the minimal polynomial of σ is in fact $x^d - 1$. (Note that the minimal polynomial of a vector space map is not necessarily irreducible, unlike the minimal polynomial of an algebraic element in a field extension.)

Since the powers of α give a basis of K over F, certainly $F(\alpha) = K$. And since $\sigma(\alpha^d) = \rho^d\alpha^d = \alpha^d$, we have $\alpha^d \in \operatorname{Fix}\langle\sigma\rangle = \operatorname{Fix} G = F$ since the extension is Galois. Certainly then $\alpha^n \in F$ as well. □

Putting together the previous two results we get the following very nice theorem.

Theorem 17.37. *Let $F \subseteq K$ be a field extension and suppose that $x^n - 1$ has n distinct roots in F. Then the following are equivalent:*

(1) *K/F is Galois with $\operatorname{Gal}(K/F)$ cyclic of order dividing n.*
(2) *$K = F(\alpha)$ for some $\alpha \in K$ with $\alpha^n \in F$.*

We now single out those extensions that can be formed by iterating the procedure of adjoining a root.

Definition 17.38. A field extension $F \subseteq K$ is called a *root extension* if there is a chain of subfields

$$F = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K$$

such that for all $i \geq 0$ there is $\alpha_i \in K_{i+1}$ and $n_i \geq 1$ such that $K_{i+1} = K_i(\alpha_i)$ and $\alpha_i^{n_i} \in K_i$. A polynomial $f \in F[x]$ is *solvable by radicals* if there exists a root extension $F \subseteq K$ such that f splits in $K[x]$.

In other words, at each step in a root extension, we get the next field by adjoining some root of an element we already have. Speaking loosely, the elements in a root extension K of F are those that can be expressed using only elements in F, field operations, and nested root symbols. Thus a polynomial is solvable by radicals if all of its roots can be expressed in such a way. For example, $\sqrt[3]{(1/10)+\sqrt{2}}/\sqrt[5]{3}$ lies in a root extension of $\mathbb{Q}$, namely

$$\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2}) \subseteq \mathbb{Q}(\sqrt{2})(\sqrt[3]{(1/10)+\sqrt{2}}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt[3]{(1/10)+\sqrt{2}})(\sqrt[5]{3}).$$

A major preoccupation of mathematicians in the Renaissance period in Europe was to find solutions for a general polynomial equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

Really, what was desired was a formula, or method, that would give the solutions to any equation in terms of manipulations of its coefficients, including algebraic operations and taking roots. Solutions of this type for quadratic equations had been long known (although not exactly in the form of the quadratic formula as we know it today). Formulas for solving polynomial equations of degree 3 and 4 were eventually successfully developed. These were first published by Cardano in 1545, but the solutions are now attributed to Tartaglia and del Ferro (for the cubic) and Ferrari (for the quartic). Interestingly, complex numbers arise naturally in these formulas even when one is only seeking real roots, and this probably helped spur the eventual acceptance of complex numbers in mathematics. No solution to the general degree 5 (quintic) equation could be found in teh following years, and eventually, hundreds of years later in 1824, Abel proved that no solution of this kind could exist (building on earlier work of Ruffini). Galois's theory, which came just a bit later in 1830, then put this theorem into a more general context.

Technically, what Abel and Galois proved is that there can exist no formula for the solution of a quintic which depends only on field operations and nested root signs. In our terminology, we can state the key theorem as follows:

Theorem 17.39. *There exists a polynomial $f \in \mathbb{Q}[x]$ of degree 5 such that the splitting field K of f in $\mathbb{C}$ is not a root extension of $\mathbb{Q}$.*

Thus the roots of f in $\mathbb{C}$ are "not expressible in terms of radicals". In fact Galois's work showed a much stronger result, which today we state in terms of solvable groups although that concept wasn't formalized at the time.

Theorem 17.40. *Let F be a field of characteristic 0. Then $f \in F[x]$ is solvable by radicals if and only if $\mathrm{Gal}(K/F)$ is a solvable group, where K is the splitting field of f over F.*

Then to give the required example in Theorem 17.39, it just suffices to find a particular polynomial whose splitting field has a non-solvable Galois group. Since the splitting field of a polynomial of degree d has degree at most $d!$ over $\mathbb{Q}$, this explains why no such example could exist when $d \leq 4$, since all groups of order at most 24 are solvable, and thus all polynomials of degree at most 4 in $\mathbb{Q}$ are solvable by radicals. On the other hand, we will see that there does exist a polynomial of degree 5 such that the Galois group of its splitting field is the non-solvable group S_5.

We have decided to omit the complete proof of Theorem 17.40, which is a bit technical. The result itself does not have the importance it once did, as solving polynomial equations explicitly is no longer a central topic in algebra. But from the results we have already presented the main idea of Theorem 17.40 is easy to grasp, and so we briefly discuss this. First, recall that in our study of groups we defined a group G to be solvable if it has a series of subgroups

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

such that H_{i+1}/H_i is abelian for all i. However, if G is finite and solvable, then in fact it has such a series where each H_{i+1}/H_i is cyclic. This follows easily from the fact that a finite abelian group is a direct product of cyclic groups, which allows one to add additional terms to any series as above in order to make the factor groups cyclic. Thus finite solvable groups are exactly those groups which are built out of cyclic groups in this sense. By definition, root extensions are field extensions which are built out of extensions where one adjoins a single root. Finally, Theorem 17.37 showed that an extension where one adds a root has Galois group which is cyclic (under certain hypotheses). These facts, together with the fundamental theorem of Galois theory, already suggest that root extensions should correspond roughly to solvable groups under the Galois correspondence.

In order to write down a precise proof of Theorem 17.40, one has to deal with some additional technical details. First, in Theorem 17.40 there is no assumption that F contains any roots of 1, so to successfully apply Theorem 17.37 one first has to adjoin a bunch of roots of 1 to F, and show this doesn't change the property of being a root extension. Second, a root extension is not necessarily Galois, since at each stage we just add one root of a polynomial, so one may need to pass to a larger root extension which is Galois in order to apply the fundamental theorem.

From now on we simply assume Theorem 17.40. We will, however, show how to use it to prove Theorem 17.39.

Proof of Theorem 17.39. Let $f = 2x^5 - 10x + 5 \in \mathbb{Q}[x]$. We claim that f is not solvable by radicals over $\mathbb{Q}$.

Let K be the splitting field of f over $\mathbb{Q}$. We claim that $\mathrm{Gal}(K/\mathbb{Q}) \cong S_5$. Once this proved, then since S_5 is not solvable, we will know that f is not solvable by radicals by Theorem 17.40.

The polynomial f is irreducible over $\mathbb{Q}$ by an application of the Eisenstein criterion with prime 5. Thus if $\alpha \in K$ is a root of f, we will have $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 5$. In particular, if $G = \mathrm{Gal}(K/\mathbb{Q})$ then $|G| = |K : \mathbb{Q}|$ is divisible by 5. Thus G has an element of order 5 by Cauchy's theorem.

Next, if $\alpha_1, \ldots, \alpha_5$ are the roots of f in K (they are distinct since we are in characteristic 0), then every $\sigma \in G$ sends each root of f to a root of f, and so gives a permutation of these roots. Since these roots also generate K over $\mathbb{Q}$, σ is determined by how it permutes the roots of f. In this way we get an injective homomorphism from G to S_5, which maps $\sigma \in G$ to the corresponding permutation of the roots of f. Now think of G as a subgroup of S_5. The only elements of order 5 in S_5 are 5-cycles, so G contains a 5-cycle.

Now f was chosen to have exactly 3 real roots. This fact can be easily verified using calculus. Namely, we have $f' = 10x^4 - 10$ which has real zeroes at the values ± 1. By the mean value theorem, between any two real roots $r_1 < r_2$ of f there must be $r_1 < s < r_2$ such that $f'(s) = 0$. Since f' has only two real roots we conclude that f has at most 3 real roots. On the other hand, $f(-2) < 0$, $f(-1) > 0$, $f(1) < 0$, and $f(2) > 0$, so an application of the intermediate value theorem shows that f has at least 3 real roots. So f has exactly 3 real roots. Now we have, say, that $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$, while $\alpha_4, \alpha_5 \notin \mathbb{R}$. Since the non-real roots of a polynomial with real coefficients come in conjugate pairs, we must have $\alpha_5 = \overline{\alpha_4}$.

Now the complex conjugation map $z \mapsto \overline{z}$ is an automorphism of $\mathbb{C}$. It permutes the roots α_i of f and thus it restricts to an automorphism of K, that is an element in $G = \mathrm{Gal}(K/\mathbb{Q})$. This automorphism obviously acts on the roots as the 2-cycle (45).

To complete the proof, one checks that any subgroup of S_5 which contains a 5-cycle and a 2-cycle is equal to all of S_5. This is an exercise in group theory which we leave to the reader. This proves the claim that $G \cong S_5$ and finishes the proof. □

17.7. Algebraically closed fields and algebraic closures.

We have used throughout our study of field theory that the field $\mathbb{C}$ of complex numbers is algebraically closed. In this section we will prove that result, as well as studying algebraically closed fields more generally.

Recall the following definition:

Definition 17.41. A field K is algebraically closed if for all $f \in K[x]$, f splits in $K[x]$, that is $f = c(x - \alpha_1)\dots(x - \alpha_n)$ for some $\alpha_1, \dots, \alpha_n \in K$.

This definition can formulated in a number of slightly different ways.

Lemma 17.42. *Let K be a field. The following are equivalent:*

(1) *K is algebraically closed, that is, every polynomial in $K[x]$ splits in $K[x]$.*

(2) *Every nonconstant polynomial $f \in K[x]$ has a root in K.*

(3) *If $K \subseteq L$ is an algebraic extension, then $L = K$.*

(4) *If $K \subseteq L$ is an algebraic extension with $[L : K] < \infty$, then $L = K$.*

Proof. (1) $\implies$ (2) is obvious.

For (2) $\implies$ (3), supppose that L/K is algebraic. Given $\alpha \in L$, let $f = \operatorname{minpoly}_K(\alpha) \in K[x]$. Then f has a root in K, but f is also irreducible over K. This forces $\deg f = 1$ and hence $\alpha \in K$. Thus $K = L$.

(3) $\implies$ (4) is also immediate.

(4) $\implies$ (1): Let $f \in K[x]$. Let L be a splitting field for f over K, so $[L : K] < \infty$ by definition. Then we have $L = K$, so f splits in $K[x]$ already. □

Definition 17.43. If $F \subseteq K$ is a field extensions, K is called an *algebraic closure* of F if K/F is algebraic and K is algebraically closed.

We are going to show shortly that every field has an algebraic closure, and the algebraic closure is unique up to isomorphism.

Here is an alternative way of thinking about the algebraic closure.

Lemma 17.44. *Let $F \subseteq K$ be an algebraic extension. Then K is an algebraic closure of F if and only if every $f \in F[x]$ splits in $K[x]$.*

Proof. One direction is easy: if K is an algebraic closure, then K is algebraically closed. Since every $f \in F[x]$ is also in $K[x]$, of course f splits in $K[x]$ by definition.

Conversely, suppose that every $f \in F[x]$ splits in $K[x]$. Suppose that $K \subseteq L$ is an algebraic extension. Since $F \subseteq K$ is algebraic, we have that $F \subseteq L$ is algebraic, by Theorem 16.39. For $\alpha \in L$, consider $f = \operatorname{minpoly}_F(\alpha)$. Then f splits in $K[x]$. Since α is a root of f, $\alpha \in K$. Thus $K = L$. Now K is algebraically closed by Lemma 17.42. □

While we have only defined the splitting field of a single polynomial, given a set S of polynomials in $F[x]$, one could define an algebraic extension $F \subseteq K$ to be a *splitting field of the set S* if every polynomial in S splits in K and K is generated over F by the roots of all polynomials in S. In this point of view, Lemma 17.44 is telling us that an algebraic closure of F is essentially a splitting field for the set of *all* polynomials in $F[x]$.

Let us note that to find an algebraic closure, it suffices to find some algebraically closed field containing a given one.

Lemma 17.45. *Let $F \subseteq L$ be a field extension and assume that L is algebraically closed. Let $K = \{\alpha \in L | \alpha \text{ is algebraic over } F\}$. Then K is an algebraic closure of F.*

Proof. We have seen previously in Corollary 16.33 that K is a subfield of L such that K/F is algebraic. If $f \in F[x]$, then f splits in $L[x]$ since L is algebraically closed. But the roots of f in L are algebraic over F and so they belong to K. Thus f already splits over K. Now K is an algebraic closure of F by Lemma 17.44. □

We previously defined the field of algebraic numbers $\overline{\mathbb{Q}}$ as the set of elements in $\mathbb{C}$ that are algebraic over $\mathbb{Q}$. We see by Lemma 17.45 that $\overline{\mathbb{Q}}$ is an algebraic closure of $\mathbb{Q}$. It is much smaller than $\mathbb{C}$, for it is not hard to prove that an algebraic closure of a countable field F is again countable, for there are countably many polynomials in $F[x]$, and each has finitely many roots.

We are now ready to prove the first main result of this section.

Theorem 17.46. *Any field F has an algebraic closure.*

Proof. By Lemma 17.45, it suffices to find any algebraically closed field L containing F. Then the set of elements in L that are algebraic over F will be the desired algebraic closure.

All proofs of this result depend on some version of the axiom of choice. The following elegant proof, due to Emil Artin, just uses the fact that any commutative ring has a maximal ideal (which we proved as a consequence of Zorn's lemma). Recall the idea of Lemma 16.41: Given a irreducible polynomial $f \in F[x]$, we can create a larger field containing F in which f has a root by taking $F[x]/(f)$. The idea of this proof is to do this for all polynomials in $F[x]$ at once, adding one new variable for each one. The resulting ring is not a field, but we can just pass to some factor ring that is a field.

For each nonconstant polynomial $f \in F[x]$ we define an intedeterminate x_f. Now we define $R = F[x_f | f \in F[x], \deg(f) > 0]$ to be a polynomial ring generated over F by these (infinitely many) variables. Let I be the ideal of R generated by the set of polynomials $\{f(x_f) | f \in F[x], \deg(f) > 0\}$. Each $f(x_f)$ is a polynomial involving only one of the variables.

We claim that $I \neq R$. Suppose instead that $I = R$ is the unit ideal. Then $1 \in I$, so $1 = \sum_{i=1}^{n} g_i f_i(x_{f_i})$ for some distinct polynomials $f_1, f_2, \ldots, f_n \in R$ with $\deg f_i > 0$ and some $g_i \in R$. Let K be a splitting field over F of $f_1 f_2 \ldots f_n$. Thus each f_i has a root $\alpha_i \in K$. Now we define a homomorphism $\phi : R \to K$ as follows: We let $\phi(a) = a$ for $a \in F$; $\phi(x_{f_i}) = \alpha_i$ for $1 \leq i \leq n$; and $\phi(x_g) = 0$ for all g's not equal to any f_i. Note that by the universal property of a polynomial ring, given a homomorphism $F \to K$, we can specify a unique homomorphism from $R \to K$ by sending each variable x_f to any element of K we please, so there does exist such a homomorphism ϕ.

Now we have $1 = \phi(1) = \sum_{i=1}^{n} \phi(g_i) f_i(\alpha_i) = \sum_{i=1}^{n} \phi(g_i) 0 = 0$, because α_i is a root of f_i by definition. This is a contradiction. Thus $I \neq R$ as claimed.

Since I is a proper ideal, we can choose a maximal ideal M of R with $I \subseteq M \subsetneq R$ (Proposition 4.6). Then $L_1 = R/M$ is a field and we have a homomorphism $\psi : F \to L_1$ which is the composition of the inclusion of F in R followed by the natural homomorphism $R \to R/M$. Since F is a field, ψ is injective and so we can think of F as a subfield of L_1. As such, if $f \in F[x]$ is irreducible, then f has a root $x_f + M \in L_1$, because $f(x_f + M) = f(x_f) + M = 0 + M$, as $f(x_f) \in I \subseteq M$.

To summarize, we have shown that there exists a field extension $F \subseteq L_1$ such that every nonconstant $f \in F[x]$ has a root in L_1. Note that this does not prove that L_1 is an algebraic closure of F, because we do not know that f splits over L_1, only that it has at least one root. However, we can now proceed inductively as follows. By the same argument, there is a field extension $L_1 \subseteq L_2$ such that every nonconstant polynomial in $L_1[x]$ has a root in L_2. Continue in this way, defining a chain of fields $F \subseteq L_1 \subseteq L_2 \subseteq \ldots$. Now let $L = \bigcup_{n \geq 0} L_n$, which as a union of fields is easily seen to be a field. If $g \in L[x]$ is nonconstant, then each coefficient of g lies in some L_i, and so $g \in L_n[x]$ for some n. By construction, g has a root in $L_{n+1} \subseteq L$. Since every nonconstant polynomial in $L[x]$ has a root in L, the field L is algebraically closed by Lemma 17.42. □

Let us also prove now that algebraic closures are unique up to isomorphism. This is similar to our results on uniqueness of splitting fields (as we have already remarked, an algebraic closure is like a splitting field for the set of all polynomials). As in those results, is convenient to work over an isomorphism of base fields rather than a single base field.

Theorem 17.47. *Let $\theta : F \to F'$ be an isomorphism of fields. Let $F \subseteq K$ and $F' \subseteq K'$ be algebraic closures. Then there is an isomorphism $\rho : K \to K'$ such that $\rho|_F = \theta$.*

Proof. Consider the set consisting of triples (E, E', ψ) where E, E' are subfields with $F \subseteq E \subseteq K$, $F' \subseteq E' \subseteq K'$, and $\psi : E \to E'$ is an isomorphism such that $\psi|_F = \theta$. Put a partial order on this set where $(E, E'\psi) \leq (L, L', \rho)$ if $E \subseteq L$, $E' \subseteq L'$, and $\rho|_E = \psi$. In other words, elements of the set are isomorphisms matching up subfields of K and K', and a larger element represents an extension of that isomorphism to one defined on larger subfields. The hypotheses of Zorn's Lemma hold for this set, because given any chain $\{(E_i, E_i', \psi_i) | i \in I\}$ we can can extend the isomorphisms ψ_i to the unions to get an upper bound of the form $(\bigcup_i E_i, \bigcup_i E_i', \psi)$.

By Zorn's Lemma, there is a maximal element (L, L', ρ) in the set. Suppose that $L \neq K$. Then if $\alpha \in K \setminus L$, let $f = \text{minpoly}_L(\alpha)$ and choose any root $\alpha' \in K'$ of $f' = \rho(f)$; such a root exists because K' is algebraically closed. By Lemma 16.46, we can extend the isomorphism ρ to an isomorphism $\delta : L(\alpha) \to L'(\alpha')$, where $\delta|_L = \rho$. But this implies that $(L(\alpha), L'(\alpha'), \delta)$ is a strictly larger element of our set of partial isomorphisms, contradicting that (L, L', ρ) is maximal. We conclude that $L = K$.

Now $\rho : K \to L'$ is an isomorphism. The property of being algebraically closed is preserved by automorphisms, so L' is algebraically closed as well. However, since K'/F' is algebraic, we see that K'/L' is algebraic. Because algebraically closed fields have no proper algebraic extensions by Lemma 17.42, $K' = L'$ and ρ is an isomorphism $K \to K'$ such that $\rho|_F = \theta$, as we wished. □

Of course, taking $\theta = 1_F$ in the result above we get that any two algebraic closures K, K' of F are isomorphic as fields. Given that the algebraic closure is essentially unique by this result, sometimes the algebraic closure of a field F is simply notated $\overline{F}$. The notation $F = \overline{F}$ is used as shorthand to indicate that a field F is itself algebraically closed.

Here is another interesting consequence.

Corollary 17.48. *Let $F \subseteq E$ be an algebraic extension. If $F \subseteq K$ is an algebraic closure, there is an isomorphism $\phi : E \to L$ for some subfield L with $F \subseteq L \subseteq K$, where $\phi|_F = 1_F$.*

In other words, given any algebraic extension, an "isomorphic copy" of it can be found inside any fixed algebraic closure of the base field. Thus when we are studying algebraic extensions of F, we can always fix an algebraic closure of F and make all constructions inside there if we wish.

Proof. Let $E \subseteq K'$ be an algebraic closure of E. Since E/F and K'/F are algebraic, K'/F is also algebraic. Because K' is algebraically closed, we conclude that K' is also an algebraic closure of F. Now choose by Theorem 17.47 an isomorphism $\rho : K' \to K$ such that $\rho|_F = 1_F$. If $L = \rho(E)$, then $\phi = \rho|_E$ is an isomorphism $\phi : E \to L$ with $\phi|_F = 1_F$. □

Example 17.49. Let $\mathbb{F}_p$ be the field with p elements for some prime p. Fix an algebraic closure $\overline{\mathbb{F}_p}$. For each $n \geq 1$ we have a field $\mathbb{F}_{p^n}$ with p^n elements. By the corollary we can find a copy of this field inside the algebraic closure, so we consider $\mathbb{F}_p \subseteq \mathbb{F}_{p^n} \subseteq \overline{\mathbb{F}_p}$.

Now we claim that $\overline{\mathbb{F}_p} = \bigcup_n \mathbb{F}_{p^n}$. To see this, note that if $\alpha \in \overline{\mathbb{F}_p}$, then α is algebraic over $\mathbb{F}_p$, so we can consider $f = \operatorname{minpoly}_{\mathbb{F}_p}(\alpha)$. If f has degree n, then f divides $x^{p^n} - x$ in $\mathbb{F}_p[x]$, so α is a root of $x^{p^n} - x$. The subfield $\mathbb{F}_{p^n}$ must be equal to the set of roots of $x^{p^n} - x$ in $\overline{\mathbb{F}_p}$, so we see that $\alpha \in \mathbb{F}_{p^n}$ for the fixed copy of $\mathbb{F}_{p^n}$. This proves the claim.

This gives a quite explicit picture of $\overline{\mathbb{F}_p}$ as the union of all finite fields of characteristic p. Note, however, that these fields do not form a single chain, as $\mathbb{F}_{p^d}$ is a subset of $\mathbb{F}_{p^n}$ if and only if $d|n$. Rather, we are taking the union of a more complicated partially ordered set.

For our last main result, we will finally prove that the field $\mathbb{C}$ of complex numbers is algebraically closed. Many quite different proofs of this result are known. All use some amount of analysis, which is unavoidable because the real numbers are an analytic object, defined by a limiting process. There is a well-known proof that relies on results in complex analysis, for example.

The proof we give uses Galois theory and reduces the amount of analysis needed to a few elementary facts about the real numbers.

Lemma 17.50. (1) *If $f \in \mathbb{R}[x]$ has odd degree, then f has a root in $\mathbb{R}$.*

(2) *if $g \in \mathbb{C}[x]$ has degree 2, then g splits over $\mathbb{C}$.*

Proof. (1) Let $f = a_n x^n + \cdots + a_0 \in \mathbb{R}[x]$ where $\deg f = n$ is odd. Since we are just trying to show that f has a root in $\mathbb{R}$, without loss of generality we can replace f with $-f$ if necessary and thus assume that $a_n > 0$. Now it is standard that $\lim_{n \to \infty} f(x) = \infty$ and $\lim_{n \to -\infty} f(x) = -\infty$. By the intermediate value theorem, since f is a continuous function f must have a root in $\mathbb{R}$.

(2). If $g = ax^2 + bx + c$ then the quadratic formula tells us that $(-b + \sqrt{b^2 - 4ac})/2a$ is a root of g in $\mathbb{C}$, for any square root $\sqrt{b^2 - 4ac}$ in $\mathbb{C}$. Once g has a root α in $\mathbb{C}$, then $g = (x - \alpha)h$ by the factor theorem, but then h already has degree 1 and so g splits. □

Note that for any complex number $z = re^{i\theta}$ in polar form, with $r \geq 0$, then $\sqrt{r}e^{i\theta/2}$ is a square root of z, where $\sqrt{r}$ is the nonnegative real square root of r. Thus ultimately the existence of square roots in $\mathbb{C}$ is a consequence of the fact that nonegative real numbers have a unique nonnegative square root. This follows easily from the least upper bound property.

The analysis above is all we need to prove our main result.

Theorem 17.51. *$\mathbb{C}$ is algebraically closed.*

Proof. We will show that if $\mathbb{C} \subseteq L$ is a finite degree extension, then $L = \mathbb{C}$. This implies that $\mathbb{C}$ is algebraically closed by Lemma 17.42.

Since $[\mathbb{C} : \mathbb{R}] = 2$, we also have $[L : \mathbb{R}] < \infty$, and of course $L/\mathbb{R}$ is separable since we are in characteristic 0. Thus we can take a Galois closure M of L so that $L \subseteq M$ and $M/\mathbb{R}$ is Galois, by Proposition 17.13.

Now let $G = \mathrm{Gal}(M/\mathbb{R})$. Let P be a Sylow 2-subgroup of G. Let $K = \mathrm{Fix}(P)$. Then $[K : \mathbb{R}] = [G : P]$ is odd. If $\alpha \in K$, then $[\mathbb{R}(\alpha) : \mathbb{R}]$ divides $[K : \mathbb{R}]$ so it is also odd. Then $f = \mathrm{minpoly}_{\mathbb{R}}(\alpha)$ is irreducible and of odd degree. But by Lemma 17.50(1), f has a root in $\mathbb{R}$, and so cannot be irreducible unless it has degree 1, in which case $\alpha \in \mathbb{R}$. This shows that $K = \mathbb{R}$. This implies $P = G$ and so G is a finite 2-group. Also, $[M : \mathbb{R}]$ is a power of 2.

Since $[\mathbb{C} : \mathbb{R}] = 2$ we have $[M : \mathbb{C}]$ is a power of 2 as well. Also, because $M/\mathbb{R}$ is Galois, so is $M/\mathbb{C}$. Suppose that $M \neq \mathbb{C}$. Now $\mathrm{Gal}(M/\mathbb{C})$ is a nontrivial 2-group, so from our earlier results on p-groups we know that it must have a subgroup H of index 2. If $E = \mathrm{Fix}(H)$ then $\mathbb{C} \subsetneq E \subseteq M$ with $[E : \mathbb{C}] = 2$. If $\alpha \in E \setminus \mathbb{C}$, then $\mathrm{minpoly}_{\mathbb{C}}(\alpha)$ has degree 2, but by Lemma 17.50, any degree 2 polynomial in $\mathbb{C}[x]$ splits over $\mathbb{C}$ and cannot be irreducible, which is a contradiction. We conclude that in fact $M = \mathbb{C}$. Then $L = \mathbb{C}$, completing the proof. □

We close with a curious result about the automorphism group of $\mathbb{C}$. In a homework problem, you were asked to show that $\mathrm{Aut}(\mathbb{R}) = 1$, by checking that every automorphism fixes $\mathbb{Q}$ and is continuous. On the other hand, although $\mathbb{C}$ is just a degree 2 extension of $\mathbb{R}$, its automorphism group $\mathrm{Aut}(\mathbb{C})$ is actually very large. This can be seen using the idea of a transcendence basis.

Given any field extension $F \subseteq K$, it is possible to choose a set of elements $\{y_\alpha \in K | \alpha \in I\}$ such that (i) the $\{y_\alpha\}$ are *algebraically independent* in the sense that $T = F(y_\alpha | \alpha \in I)$ is isomorphic to the field of fractions of a polynomial ring $F[y_\alpha | \alpha \in I]$; and (ii) $T \subseteq K$ is algebraic. The set $\{y_\alpha | \alpha \in I\}$ is called a *transcendence basis* for the extension. It can be proved to exist by an application of Zorn's Lemma. The cardinality of a transcendence basis is uniquely determined (though the subfield T it generates isn't). This cardinality is called the *transcendence degree* of the extension.

Consider the particular extension $\mathbb{Q} \subseteq \mathbb{C}$. In this case one can show that the transcendence degree is uncountable. (It this were not the case, but rather the transcendence basis $\{y_\alpha\}$ were countable, then the field of rational functions $\mathbb{Q}(y_\alpha | \alpha \in I)$ would be countable, and then since an algebraic extension of a countable field is countable, $\mathbb{C}$ would be countable, a contradiction.)

Now fix a transcendence basis $\{y_\alpha | \alpha \in I\}$ for $\mathbb{C}$ over $\mathbb{Q}$. Any permutation of the set I gives an automorphism of the polynomial ring $\mathbb{Q}[y_\alpha | \alpha \in I]$ where the variables are permuted in the same way. This then extends to an automorphism of the field $T = \mathbb{Q}(y_\alpha | \alpha \in I)$ with the same permutation of the variables. Finally, since $\mathbb{C}$ is algebraically closed and $\mathbb{C}/T$ is algebraic, $\mathbb{C}$ is an algebraic closure of T. Thus any automorphism of T extends to an automorphism of $\mathbb{C}$, by Theorem 17.47.

In this way we can see that $\mathbb{C}$ has at least as many automorphisms as the number of elements in the permutation group $\mathrm{Sym}(I)$, where I is the index set of a transcendence basis. Since I is uncountable, the set of permutations of I actually has cardinality even bigger than the cardinality of I (as can be seen by a version of Cantor's diagonal argument). Thus $\mathrm{Aut}(\mathbb{C})$ is huge.

On the other hand, while a transcendence basis for $\mathbb{C}$ over $\mathbb{Q}$ exists, it is impossible to write one down explicitly, and so the automorphisms of $\mathbb{C}$ one gets in this way also do not have any kind of explicit description. And in fact they tend to have bizarre properties. It is possible to show that except for the identity map and complex conjugation, any automorphism of $\mathbb{C}$ is discontinuous and maps $\mathbb{R}$ onto a dense subset of $\mathbb{C}$. So these really are hard to picture.

The concept of a transcendence basis is generally useful in commutative ring theory (not just for creating strange automorphisms). We would cover it in more detail if we had more time. You can find a treatment of it in Chapter 24 of Isaacs' book.