18.702 — Algebra II

Class by Roman Bezrukavnikov

Notes by Sanjana Das

Spring 2022

Notes for the MIT class **18.702** (Algebra II), taught by Roman Bezrukavnikov. All errors are my responsibility.

Contents

§1 Introduction

In **18.701**, we saw the concept of a *group action*, and how group actions relate to the symmetry of the set being acted on. In *representation theory*, we'll look at how a group can act on a *vector space* — combining ideas of both symmetry and linearity.

Another main topic we'll see is *ring theory*, where we can both add *and* multiply in abstract spaces.

We'll also see *Galois theory*, and how it relates to the symmetries of the solutions to a polynomial equation. For example, we're familiar with the quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

In this formula, we have a $\pm$ — there's two possible solutions, and when we're trying to write one down, we have this ambiguity around which square root to extract. We can think of this as a *symmetry* on the set of solutions.

In this specific situation, the symmetry isn't very rich — changing the sign of the square root just swaps the roots. But in equations of higher degree, this is more interesting. We'll see that symmetries (such as swapping the two square roots in the quadratic case) control the existence of a formula for the roots of the polynomial, as well as what such a formula can look like (when it exists).

§2 Group Representations

§2.1 Definitions

Definition 2.1. Let G be a group. Then a *complex representation* of G of dimension n is a homomorphism $R : G \to \mathrm{GL}_n(\mathbb{C})$.

We can define a representation over any field F similarly. But in this class, we'll mostly only study complex representations of finite groups.

Note that matrices act on column vectors — a matrix A describes a linear transformation from $\mathbb{C}^n$ to itself sending $v \mapsto Av$. If $A \in \mathrm{GL}_n(\mathbb{C})$, then this linear transformation is bijective. So $\mathrm{GL}_n(\mathbb{C})$ is the group of linear automorphisms of $\mathbb{C}$. This means a dimension-n representation is the same as a linear action of G on $\mathbb{C}^n$ (an action of G where each element acts by a linear map on $\mathbb{C}^n$).

To write down a representation R, for each $g \in G$ we can define an invertible matrix R_g. The condition that R is a homomorphism means we must have

$$R_{gh} = R_g R_h$$

for all $g, h \in G$, and R_1 must be the identity matrix.

Equivalently, we could write down R by thinking of the elements R_g as *linear maps* on $\mathbb{C}^n$ rather than matrices — for each $g \in G$ and each $v \in \mathbb{C}^n$, we can define the vector $R_g(v) \in \mathbb{C}^n$. For each g, the map $R_g(v)$ should be linear in v, meaning that

$$R_g(v_1 + v_2) = R_g(v_1) + R_g(v_2),$$
$$R_g(cv) = cR_g(v).$$

Meanwhile, the different maps R_g should be compatible with composition, meaning that for all v,

$$R_{gh}(v) = R_g(R_h(v)).$$

Example 2.2

Some examples of representations are the following:

1. Any group has a trivial 1-dimensional representation, where $R_g = 1$ for all g.
2. S_n has the 1-dimensional *sign representation*, where $R_\sigma = 1$ if σ is even, and -1 if σ is odd.
3. S_n has a n-dimensional representation: the *permutation representation* where we send a permutation to the corresponding permutation matrix. For example, if $\sigma = (123)$ then
$$R_\sigma = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$
4. $\mathbb{Z}/m\mathbb{Z}$ has a 2-dimensional real representation, as the group of rotational symmetries of a regular m-gon. In this representation, for each residue $\overline{a}$ (used to denote a mod m), we map
$$\overline{a} \mapsto \begin{bmatrix} \cos 2\pi a/m & -\sin 2\pi a/m \\ \sin 2\pi a/m & \cos 2\pi a/m \end{bmatrix}.$$

Note that if G is given by generators $x_1, \ldots, x_n$ and some relations, then to define a representation of G, it's enough to specify the images $x_1 \mapsto \gamma_1, \ldots, x_n \mapsto \gamma_n$, and make sure that the γ_i satisfy all the same relations as the x_i.

Example 2.3

Since $\mathbb{Z}/m\mathbb{Z} = \langle x \mid x^m = 1\rangle$ (in other words, $\mathbb{Z}/m\mathbb{Z}$ is generated by $\overline{1}$, with the relation $m \cdot \overline{1} = 0$), in order to define a representation of $\mathbb{Z}/m\mathbb{Z}$, it's enough to define the image of $\overline{1}$, and check that it has order dividing m. In our above 2-dimensional representation, we have

$$\overline{1} \mapsto A = \begin{bmatrix} \cos 2\pi/m & -\sin 2\pi/m \\ \sin 2\pi/m & \cos 2\pi/m \end{bmatrix},$$

and we can check that $A^m = 1$.

Example 2.4

Let D_n be the dihedral group, the group of symmetries of a regular n-gon (which are all rotations and reflections). Then D_n is generated by a rotation and reflection:

$$D_n = \langle r, s \mid r^m = 1, srs^{-1} = r^{-1}\rangle.$$

Then D_n has a 2-dimensional real representation where

$$r \mapsto \begin{bmatrix} \cos 2\pi/m & -\sin 2\pi/m \\ \sin 2\pi/m & \cos 2\pi/m \end{bmatrix} \text{ and } s \mapsto \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

§2.1.1 A Basis-Free Definition

As we saw when studying linear algebra in **18.701**, we often want to be able to do things that don't depend on coordinates. If we have a matrix representation R, then by looking at the same linear transformations in a different basis of $\mathbb{C}^n$, we get a new matrix representation — if P is the change of basis matrix, then our new representation has

$$R'_g = P^{-1} R_g P$$

for all g. Then we say R and R' are *conjugate representations*. Since these two representations are essentially the same representation, just in different coordinates, we only want to study representations up to conjugacy — or equivalently, we actually want to study the *conjugacy classes* of representations.

In fact, we can give the definition of a representation without referencing any matrices. Recall that for any vector space V, the notation $\mathrm{GL}(V)$ denotes the group of linear automorphisms (or in other words, bijective linear transformations) of V.

Definition 2.5. A *linear representation* of G on a vector space V is a homomorphism $\rho : G \to \mathrm{GL}(V)$.

Given a linear representation, we can choose a basis for V in order to turn it into a matrix representation.

We still want to keep track of whether two representations are really the same. If we have a representation $\rho : G \to V$ and a vector space $W \cong V$, of course there's a representation $\rho' : G \to W$ as well, which is essentially the same as ρ — if I is the isomorphism $V \to W$, then we can intertwine the actions of I and ρ, defining ρ' such that

$$\rho'_g(I(v)) = I(\rho_g(v)).$$

We use this as the definition of when two representations are isomorphic.

Definition 2.6. Two linear representations $\rho : G \to \mathrm{GL}(V)$ and $\rho' : G \to \mathrm{GL}(W)$ are *isomorphic* if there exists an isomorphism $I : V \to W$ such that

$$\rho'_g(I(v)) = I(\rho_g(v))$$

for all $v \in V$ and all $g \in G$.

Studying linear representations up to isomorphism is the same as studying matrix representations up to conjugacy; we then want to describe representations up to isomorphism.

§2.2 Characters

When studying representations, we'd like the properties we study to not depend on the choice of basis. The determinant and trace of a matrix are two such properties, which motivates the definition of the *character*.

Definition 2.7. The *character* of a representation R is the function on G where

$$\chi_R(g) = \mathrm{tr}(R_g).$$

We can define character in the basis-free definition as well. This works because trace is *conjugation-invariant*, meaning that

$$\mathrm{tr}(A) = \mathrm{tr}(P^{-1}AP).$$

So although we defined the character using a matrix, the characters of conjugate matrix representations coincide — so the character of a linear representation (when we *don't* have a fixed basis) is still well-defined.

Proposition 2.8

The character χ_ρ is constant on the conjugacy classes of G — for all g and x, we have

$$\chi_\rho(xgx^{-1}) = \chi_\rho(g).$$

This follows directly from the fact that trace is conjugation-invariant — $\rho(xgx^{-1}) = \rho(x)\rho(g)\rho(x)^{-1}$ is conjugate to $\rho(g)$.

If a function is fixed on each conjugacy class of G, then we call it a *class function* — so χ_ρ is a class function for any representation ρ.

Example 2.9

Find the character of the permutation representation $\rho : S_3 \to \mathbb{C}^3$.

Solution. It suffices to take one representative from each conjugacy class, and the conjugacy classes of S_n are dependent on the cycle decomposition pattern.

For a permutation matrix, the entries are all 0 or 1, and 1's on the diagonal correspond to fixed points. So the trace of a permutation matrix is just the number of fixed points of the permutation, and we get the following character table:

(1)	(12)	(123)
3	1	0

For example,

$$(12) \mapsto \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which has trace 1 — the one fixed point is 3. □

Remark 2.10. If the dimension of ρ is 1, then the character is the same as the representation — so one-dimensional characters are exactly homomorphisms $\chi : G \to \mathbb{C}^\times$ (meaning $\chi(gh) = \chi(g)\chi(h)$). But in higher dimensions, characters are not necessarily homomorphisms.

Example 2.11

Find the one-dimensional representations of $\mathbb{Z}/n\mathbb{Z}$.

Solution. We saw that a 1-dimensional representation of $\mathbb{Z}/n\mathbb{Z}$ is determined by the image of $\overline{1}$, which must be a nth root of unity. Let this root of unity be

$$\zeta_a = \cos\frac{2\pi a}{n} + i\sin\frac{2\pi a}{n} = \exp\frac{2\pi i a}{n}.$$

Then the representation is

$$\overline{x} \mapsto \zeta_a^x = \exp\frac{2\pi i a x}{n},$$

where $\overline{x}$ denotes x mod n. □

§2.3 Building Blocks of Representations

We'll now see how to build representations from smaller ones.

§2.3.1 The Direct Sum

Definition 2.12. If $\psi : G \to \mathrm{GL}(U)$ and $\eta : G \to \mathrm{GL}(W)$ are representations of G, then their *direct sum* $\rho = \psi \oplus \eta$ is the representation on the space $V = U \oplus W$ where

$$\rho_g(u, w) = (\psi_g(u), \eta_g(w)).$$

In terms of matrices, for all g, the matrix of ρ_g is the block-diagonal matrix

$$\rho_g = \left[\begin{array}{c|c} \psi_g & 0 \\ \hline 0 & \eta_g \end{array}\right].$$

Proposition 2.13

If $\rho = \psi \oplus \eta$, then $\chi_\rho = \chi_\psi + \chi_\eta$.

Proof. This is clear from writing all representations in terms of matrices — since ρ is a block-diagonal matrix with blocks ψ and η, the diagonal entries of ρ are exactly the diagonal entries of ψ and those of η. □

Question 2.14. Given a representation, can we split it as the sum of two smaller representations?

Example 2.15

Consider the 2-dimensional representation of $\mathbb{Z}/2\mathbb{Z}$ where

$$\overline{1} \mapsto \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

(This can be thought of as the permutation representation of S_2.)

Solution. This is not a block-diagonal matrix as written. But we can take a new basis, with $v_1 = (1,1)^t$ and $v_2 = (1,-1)^t$. Then our matrix would be written as

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

which *is* block-diagonal. So this representation does split as a sum. □

Remark 2.16. A bit more generally, if $G = \mathbb{Z}/m\mathbb{Z}$ and ρ is a representation of dimension n, then it's determined by one matrix $A = \rho(\overline{1})$. Then the question of decomposing ρ into a direct sum corresponds to *diagonalization* — if we can diagonalize A, then we've found an isomorphism between ρ and a sum

$$\rho = \psi_1 \oplus \cdots \oplus \psi_n,$$

where the ψ_i are 1-dimensional representations.

But $A^n = 1_n$ (where 1_n denotes the identity matrix of dimension n), and it's possible to show that this means A cannot have any Jordan blocks of size greater than 1. So it follows that A can necessarily be diagonalized.

This means every representation of $\mathbb{Z}/m\mathbb{Z}$ is a sum of 1-dimensional representations $\rho = \psi_1 \oplus \cdots \oplus \psi_n$, where ψ_i is the representation sending $\overline{1}$ to the eigenvalue λ_i (which must be a mth root of unity). Our above example is a special case of this.

§2.3.2 Irreducible Representations

The building blocks of representations will be the *irreducible* representations.

Definition 2.17. A *G-invariant subspace* is a subspace $W \subset V$ which is preserved by G, meaning that ρ_g sends W to itself for all $g \in G$.

Definition 2.18. A representation $\rho : G \to \mathrm{GL}(V)$ is *irreducible* if the only G-invariant subspaces of V are $\{0\}$ and V itself.

If we have a G-invariant subspace $W \subset V$, then we can consider the action of G on W instead — this gives a smaller representation inside the original representation ρ. (*A priori* this gives a map $G \to \mathrm{End}(W)$, but each of the linear maps in $\mathrm{End}(W)$ we use must be invertible — $\rho(g^{-1})$ and $\rho(g)$ are inverses for all g — so we actually get a map $G \to \mathrm{Aut}(W) = \mathrm{GL}(W)$, and therefore a representation on W.)

Of course, if $\rho : G \to \mathrm{GL}(V)$ can be written as a direct sum $\rho = \psi \oplus \eta$ (where ψ acts on U and η on W, with $V = U \oplus W$), then ρ is not irreducible — the set of pairs $(u, 0)$ in V (which is isomorphic to U) is a G-invariant subspace. So irreducible representations can't be split as a direct sum of representations. (It's not clear whether the converse is true, but as we'll see later, for complex representations of finite groups, the converse *is* true!)

All 1-dimensional representations are trivially irreducible; but there are usually other irreducible representations as well.

Example 2.19

Consider the representation of S_3 acting on the space

$$V = \{(x, y, z) \in \mathbb{C}^3 \mid x + y + z = 0\}$$

by permuting coordinates. Show that this representation is irreducible.

Proof. Suppose $W \subset V$ is a nonzero subspace of V. Pick a nonzero vector $v = (x, y, z) \in W$; we want to then use the permutations to generate all of V.

Two coordinates must be different (otherwise $x = y = z = 0$), so without loss of generality, $x \neq y$. Then $(12)v$ and v must be in W, so

$$(12)v - v = (y - x, x - y, 0) \in W.$$

By scaling, we get that $(1, -1, 0) \in W$. Now we can permute the coordinates to get that $(1, 0, -1) \in W$ as well. But $(1, -1, 0)$ and $(1, 0, -1)$ span V, so we must have $W = V$. This means V has no S_3-invariant subspaces (except for $\{0\}$ and itself), so this representation is irreducible. □

This argument generalizes — for *any* n, we can consider the permutation representation of S_n acting on the set $V \subset \mathbb{C}^n$ consisting of points with $x_1 + \cdots + x_n = 0$. The same argument shows that this representation is always irreducible.

Remark 2.20. In the case $n = 3$, we can think of this argument geometrically — imagine drawing an equilateral triangle, for example with vertices at $(2, -1, -1)$, $(-1, 2, -1)$, and $(-1, -1, 2)$. Then S_3 is the set of symmetries of this equilateral triangle. By considering the reflections, we can see that there are no non-obvious subspaces of V fixed by all three reflections.

Example 2.21

Consider the same set V as a representation of $\mathbb{Z}/3\mathbb{Z} = \{1, (123), (132)\}$, instead of the entire group S_3. Is this representation irreducible?

Proof. The answer is no. We have to be careful if thinking geometrically — geometrically, the elements of $\mathbb{Z}/3\mathbb{Z}$ correspond to *rotations* of the triangle, and it may appear that there are no subspaces of the plane fixed by a rotation. But this only works over reals. Over reals, the representation really is irreducible. But it *can't* be irreducible as a complex representation — we've already seen the general statement that every representation of a cyclic group is the sum of one-dimensional representations.

By taking two basis vectors for the plane V (which are perpendicular and have the same length), this representation is isomorphic to one on $\mathbb{C}^2$, where

$$\overline{1} \mapsto \begin{bmatrix} \cos 2\pi/3 & -\sin 2\pi/3 \\ \sin 2\pi/3 & \cos 2\pi/3 \end{bmatrix}.$$

Let $\alpha = -\frac{1}{2}$ and $\beta = \frac{\sqrt{3}}{2}$. Then we have

$$\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} \alpha - \beta i \\ \beta + \alpha i \end{bmatrix} = (\alpha - \beta i) \begin{bmatrix} 1 \\ i \end{bmatrix},$$

which means $(1, i)^t$ is an eigenvector of our linear map. So if we take the basis $v_1 = (1, i)^t$ and $v_2 = (1, -i)^t$ for $\mathbb{C}^2$ instead, then we get the matrix

$$\begin{bmatrix} \alpha - \beta i & 0 \\ 0 & \alpha + \beta i \end{bmatrix},$$

which gives a decomposition of ρ as a direct sum. □

§2.3.3 Maschke's Theorem

Question 2.22. Can any reducible representation be split as a direct sum of smaller representations?

Let $\rho : G \to \mathrm{GL}(V)$ be a representation which is *not* irreducible. Then there is a G-invariant subspace $W \subset V$. Now if we pick a basis $v_1, \ldots, v_n$ for V, such that $v_1, \ldots, v_m$ form a basis for W, then all elements of G act by matrices of the form

$$g \mapsto \left[\begin{array}{c|c} \psi_g & * \\ \hline 0 & \eta_g \end{array}\right]$$

(where the top-left block is $m \times m$), since all elements of W must be sent to elements of W.

So we can write down the maps $\psi : G \to \mathrm{GL}(W)$ and $\eta : G \to \mathrm{GL}(V/W)$, corresponding to the blocks ψ_g and η_g. These are homomorphisms — we have $\psi_{gh} = \psi_g \psi_h$ and $\eta_{gh} = \eta_g \eta_h$ (note that we're ignoring the "junk" in the top-right corner, since in V/W we treat all elements of W as 0).

Then we can write $\rho \cong \psi \oplus \eta$ if and only if the "junk" is 0 (the part denoted $*$). The values of $*$ may depend on our choice of basis, but it's actually possible to phrase this without referencing a basis — $*$ is 0 if and only if the vector space $U = \mathrm{Span}(v_{m+1}, \ldots, v_n)$ is G-invariant as well. So it's necessary to choose basis vectors whose span U is a G-invariant subspace (different choices of $v_{m+1}, \ldots, v_n$ may provide different spans), but given U, which basis of U we take isn't relevant. So we have the following conclusion:

Lemma 2.23

We can decompose $\rho \cong \psi \oplus \eta$ if and only if W has a complement (a subspace U with $V \cong W \oplus U$) which is G-invariant as well.

So to see whether we can split our representation as a direct sum, we want to see whether an invariant complement exists.

Example 2.24 (A non-example)

The group $\mathbb{Z}$ has the 2-dimensional complex representation (acting on $V = \mathbb{C}^2$) given by

$$1 \mapsto \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

which means

$$n \mapsto \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}.$$

The span W of $(1, 0)^t$ is invariant, and the representation is trivial on V/W. But our representation is not isomorphic to the sum of two trivial ones.

So there exist situations where there may be an invariant subspace which does not have any invariant complement. But it turns out that this never happens for complex representations of finite groups — we'll show that there *always* exists an invariant complement.

In order to prove this, we'll use Hermitian forms.

Theorem 2.25

If $\rho : G \to \mathrm{GL}(V)$ is a complex representation of a finite group, then V has a G-invariant positive Hermitian form.

Recall that a *Hermitian form* is a pairing $V \times V \to \mathbb{C}$, written as $\langle -, - \rangle$, which is linear in the first variable and satisfies $\langle w, v \rangle = \overline{\langle w, v \rangle}$. Then $\langle v, v \rangle$ is always real; we say the Hermitian form is *positive* if $\langle v, v \rangle > 0$ for all $v \neq 0$.

Meanwhile, we say the Hermitian form is G-invariant if $\langle gv, gw \rangle = \langle v, w \rangle$ for all $g \in G$ and all $v, w \in V$ — so the G-action doesn't affect the pairing of any two elements. (Here gv denotes $\rho_g(v)$.)

Proof. Start with *any* positive Hermitian form $\langle -, - \rangle'$. Now use the *averaging trick* — take the form

$$\langle v, w \rangle = \sum_{g \in G} \langle gv, gw \rangle'.$$

It's clear that this is a positive Hermitian form. Meanwhile, for any $h \in G$, we have

$$\langle hv, hw \rangle = \sum_{g \in G} \langle (gh)v, (gh)w \rangle' = \sum_{g \in G} \langle gv, gw \rangle' = \langle v, w \rangle,$$

since for a fixed h, as g runs over the entire group, so does gh. So $\langle -, - \rangle$ is G-invariant as well. □

Theorem 2.26

If $\rho : G \to \mathrm{GL}(V)$ is a complex (finite-dimensional) representation with a G-invariant positive Hermitian form, then every G-invariant subspace has an invariant complement.

Proof. If $W \subset V$ is a subspace, then we saw in **18.701** that

$$W^\perp = \{v \mid \langle v, w \rangle = 0 \text{ for all } w \in W\}$$

is a complement subspace. Now if $\langle -, - \rangle$ is G-invariant, and W is also G-invariant, then we claim $W^\perp$ is G-invariant as well — for any $g \in G$, we have $\langle gv, gw \rangle = 0$ if and only if $\langle v, w \rangle = 0$, so then $gv \in (gW)^\perp$ if and only if $v \in W^\perp$. But $gW = W$, so $gv \in W^\perp$ if and only if $v \in W^\perp$, which means $W^\perp$ is indeed G-invariant. □

Combining these two theorems, we get the following result.

Theorem 2.27 (Maschke's Theorem)

Every complex representation of a finite group is a sum of irreducible representations.

This is sometimes referred to as *complete reducibility*.

Proof. Induct on dimension. For the base case, all one-dimensional representations are irreducible. Now suppose $\rho : G \to \mathrm{GL}(V)$ is a representation. If it's irreducible then we're done; otherwise we can pick an invariant subspace W, which is neither 0 nor V. Let $\langle -, - \rangle$ be an invariant Hermitian form, so $V = W \oplus W^\perp$.

If we let $\eta : G \to \mathrm{GL}(W)$ and $\psi : G \to \mathrm{GL}(W^\perp)$, then we have

$$\rho = \eta \oplus \psi.$$

But these have strictly smaller dimension than ρ, so by the inductive hypothesis, they are the sum of irreducible representations. So ρ is the sum of irreducibles as well. □

Remark 2.28. We need positivity in order to get that $W^\perp$ is a complement subspace of W. But we also use positivity to guarantee that the new form we get when averaging isn't just 0.

The averaging trick is useful in many other contexts as well: in **18.701**, we used it to show that a finite group of isometries of $\mathbb{R}^n$ has a fixed point (by taking the orbit of a point, and the center of mass of that orbit).

We can also use it, given a representation $\psi : G \to \mathrm{GL}(W)$ and a vector $v \in W$, to produce an invariant vector

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

This vector is invariant because if we act on it by h, then we get

$$\psi_h \sum_{g \in G} \psi_g(v) = \sum_{g \in G} \psi_{hg}(v),$$

which is the same sum. (This idea will again be useful later.)

Remark 2.29. The existence of an invariant positive Hermitian form is equivalent to saying that every representation of a finite group is conjugate to a unitary representation — one where all the elements ρ_g are in $U_n \subset \mathrm{GL}_n(\mathbb{C})$, the set of matrices $\rho_g \in \mathrm{GL}_n(\mathbb{C})$ with $\rho_g \rho_g^* = 1$.

§2.4 Main Theorem

We'll see that the character of a representation stores quite a lot of information — in fact, it's possible to recover the representation from looking at its character! Before we state the theorem that allows us to do this, we'll start by looking at a few useful properties of these characters.

§2.4.1 More on Characters

We can prove a few basic properties of characters.

Proposition 2.30

If $\rho : G \to \mathrm{GL}(V)$ is a complex representation, then:

(a) $\chi_\rho(g)$ is a sum of roots of unity for all g.

(b) $\chi_\rho(g^{-1}) = \overline{\chi_\rho(g)}$ for all g.

(c) $\overline{\chi_\rho}$ is the character of another representation of the same dimension.

Proof. For (a), the trace of a matrix is the sum of its eigenvalues (with multiplicity). Since $\rho(g)$ must have finite order, all eigenvalues are roots of unity.

For (b), we have

$$\chi_\rho(g^{-1}) = \operatorname{tr}\rho(g)^{-1}.$$

But the eigenvalues of $\rho(g)^{-1}$ are the inverses of those of $\rho(g)$, and $\zeta^{-1} = \overline{\zeta}$ for any root of unity ζ.

For (c), given a representation ρ acting on V, we can define an action of G on the *dual space* V^*, the set of all linear maps from V to $\mathbb{C}$. If we write $f(v) = \langle f, v\rangle$ as a pairing, then take

$$\langle \rho_g^*(f), v\rangle = \langle f, \rho_{g^{-1}}(v)\rangle.$$

In other words, we take a representation of G acting on V^* such that for all $v \in V$ and $f \in V^*$, we have

$$\langle f, v\rangle = \langle \rho_g^*(f), \rho_g(v)\rangle.$$

For a matrix presentation of the same construction, fix a basis to get a matrix representation $R : G \to \mathrm{GL}_n(\mathbb{C})$. Then take $R_g^* = R_{g^{-1}}^t$. This is a valid representation because

$$((AB)^t)^{-1} = (B^tA^t)^{-1} = (A^t)^{-1}(B^t)^{-1}.$$

Since $\operatorname{tr} A^t = \operatorname{tr} A$, we have

$$\chi_{R^*}(g) = \chi_R(g^{-1}) = \overline{\chi_R(g)}$$

for all g, as desired. □

§2.4.2 The Theorem

Theorem 2.31

Let G be a finite group. Then:

(a) If $\rho_1, \ldots, \rho_n$ is a full list of irreducible representations (up to isomorphism), then their characters $\chi_1, \ldots, \chi_n$ form a basis of the space of class functions on G.

(b) This basis is orthonormal with respect to the Hermitian pairing on the space of class functions defined as

$$\langle f_1, f_2\rangle = \frac{1}{|G|}\sum_{g\in G} f_1(g)\overline{f_2(g)}.$$

(c) If the representations have dimensions $d_1, \ldots, d_n$, then

$$d_1^2 + \cdots + d_n^2 = |G|\,,$$

and d_i divides $|G|$ for each i.

We will prove this later; first we will look at a few implications.

Corollary 2.32

The isomorphism class of a representation is determined by its character.

Proof. By Maschke's Theorem, we know that every representation can be decomposed as a sum of irreducibles, as

$$\rho = \bigoplus_i \psi_i^{n_i}$$

(where ψ^n is the direct sum of n copies of ψ). But then

$$\chi_\rho = \sum_i n_i \chi_{\psi_i},$$

which means the n_i are the coefficients of the decomposition of χ_ρ in the basis given by the ψ_i — so the n_i are uniquely determined from χ_ρ. □

Corollary 2.33

The number of irreducible representations is the number of conjugacy classes of G.

Proof. Both are the dimension of the space of class functions. □

§2.4.3 Examples of Characters

The information about characters can be put into a *character table*, where columns correspond to conjugacy classes and rows to irreducible representations.

Example 2.34

Find the character table of $\mathbb{Z}/4\mathbb{Z}$.

Solution. Since $\mathbb{Z}/4\mathbb{Z}$ is an abelian group, every conjugacy class has one element. We saw earlier that all irreducible representations of $\mathbb{Z}/n\mathbb{Z}$ are one-dimensional, and are of the form $\rho_a : \overline{x} \mapsto e^{2\pi i a x/n}$. So we get the following table:

	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
χ_0	1	1	1	1
χ_2	1	-1	1	-1
χ_1	1	i	-1	$-i$
χ_3	1	$-i$	-1	i

Here χ_a is the character of ρ_a. □

Note that $\chi_\rho(1)$ is always $\dim \rho$, where 1 is the identity element of G — since $\rho(1)$ is the identity matrix.

If χ and χ' are 1-dimensional characters, then so is $\chi\chi'$. (In general, the product of two characters is still a character, but it usually won't be irreducible.) This means we can check orthogonality quite directly for one-dimensional characters:

Example 2.35

One-dimensional characters are orthogonal.

Proof. If χ and χ' are one-dimensional characters, then

$$\sum_{g\in G} \chi'(g)\overline{\chi(g)} = \sum_{g\in G} \chi'(g)\chi(g^{-1}).$$

Let $\psi(g) = \chi'(g)\overline{\chi(g)}$, so $\psi = 1$ if and only if $\chi = \chi'$. Then the proof reduces to the following claim:

Claim — For any one-dimensional character, we have

$$\sum_{g\in G}\psi(g)=\begin{cases}|G| & \psi=1\\ 0 & \text{otherwise.}\end{cases}$$

Proof. To prove this, if ψ is not trivial, we can let the sum be s, and pick some g_0 such that $\psi(g_0)=\lambda\neq 1$. Then

$$\lambda s=\sum_{g\in G}\psi(g_0)\psi(g)=\sum_{g\in G}\psi(g_0g)=\sum_{h\in G}\psi(h)=s,$$

since multiplication by g_0 permutes the group. So $\lambda s=s$, which means $s=0$. ■

Now since ψ is a 1-dimensional character, we're done. □

Remark 2.36. If the group is $\mathbb{Z}/n\mathbb{Z}$, this claim states that the sum of all nth roots of unity is 0, and our proof corresponds to rotating the regular n-gon.

Corollary 2.37

If G is abelian, every irreducible representation is one-dimensional.

There are many ways to prove this — some which use the theorem, and some which don't.

Proof 1. Use Theorem 2.31 — if $d_1,\dots,d_n$ are the dimensions of the irreducible representations, then we must have $n=|G|$, since G has n conjugacy classes. But then we know

$$d_1^2+d_2^2+\cdots+d_n^2=n.$$

Since the d_i are positive integers, they must all be 1. □

Sketch of Proof 2. We've already seen that this is true for cyclic groups. Later in this course, we'll see that every finite abelian group is a product of cyclic groups, which can be used to deduce this result as well. □

Proof 3. Let ρ be a representation; then the elements ρ_g must all commute, since G is abelian.

Claim — If we have a collection of diagonalizable matrices which commute, then there is a basis that diagonalizes all of them.

Proof. If $AB=BA$ and v is a λ-eigenvector of A, then

$$A(Bv)=BAv=B\lambda v=\lambda Bv,$$

so Bv is a λ-eigenvector of A as well. Then the λ-eigenspace of A for each λ is B-invariant (and is invariant under all matrices in the collection for the same reason). We can then split the vector space as a direct sum of the eigenspaces of A. This automatically diagonalizes A (as it is a scalar matrix on each eigenspace); meanwhile, since these eigenspaces are invariant under all other matrices, it then suffices to diagonalize our matrices (or rather, their corresponding linear operators) on each of the smaller space, which we can do by induction on the number of matrices. ■

In our situation, this means we can diagonalize all the ρ_g simultaneously, which means ρ is the direct sum of one-dimensional representations. □

Example 2.38

Compute the character table of S_3.

Solution. We've already seen 3 irreducible representations: 1, sgn, and the representation τ which permutes coordinates on $V = \{(x_1, x_2, x_3) \mid x_1 + x_2 + x_3 = 0\}$. But there are three conjugacy classes — 1, (12), and (123) — so these must be *all* the irreducible representations. So we get the following table:

	1	(12)	(123)
1	1	1	1
sgn	1	-1	1
τ	2	0	-1

In order to compute χ_τ, we used the fact that $\mathbb{C}^3 = V \oplus \mathrm{Span}((1,1,1)^t)$, so the permutation representation of S_3 is $\tau \oplus 1$. This means the character of τ is one less than the character of the permutation representation (which we saw is the number of fixed points).

We can check that these rows are orthonormal (note that the conjugacy classes have 1, 3, and 2 elements, respectively): we have

$$\langle \tau, \tau \rangle = \frac{1}{6}(2 \cdot 2 \cdot 1 + (-1) \cdot (-1) \cdot 2) = 1,$$

for example, and

$$\langle \mathrm{sgn}, \tau \rangle = \frac{1}{6}(2 \cdot 1 + (-1) \cdot 1 \cdot 2) = 0.$$

We can check other pairs similarly. □

Example 2.39

Compute the character table of S_4.

Solution. There are five conjugacy classes (corresponding to the cycle types), so there are five irreducible representations. We know that 1, sgn, τ, and $\mathrm{sgn} \cdot \tau$ are irreducible representations, so there must be one more irreducible representation ρ.

	(1)	(12)	(12)(34)	(123)	(1234)
1	1	1	1	1	1
sgn	1	-1	1	1	-1
τ	3	1	-1	0	-1
$\mathrm{sgn} \cdot \tau$	3	-1	-1	0	1
ρ	2	0	2	-1	0

To fill in the last row, we can find the dimension of ρ using $\sum d_i^2 = 24$, which gives $\chi_\rho(1) = 2$. Then we can compute the rest of the character by using the orthogonality relations

$$\langle \chi_\rho, \chi_1 - \chi_{\mathrm{sgn}} \rangle = \langle \chi_\rho, \chi_\tau - \chi_{\mathrm{sgn}\,\tau} \rangle = 0.$$

It is also possible to construct ρ explicitly using the fact that $S_4/K \cong S_3$, where K is the Klein 4-group. More precisely, this gives a homomorphism $S_4 \to S_3$, and the representation τ of S_3 gives a homomorphism $S_3 \to \mathrm{GL}_2(\mathbb{C})$; composing these two homomorphisms produces ρ. □

Now we will prove the theorem.

§2.4.4 Schur's Lemma

Definition 2.40. Let $\rho : G \to \mathrm{GL}(V)$ and $\psi : G \to \mathrm{GL}(W)$ be (not necessarily irreducible) representations. Then we define

$$\mathrm{Hom}_G(\rho, \psi) := \{f : V \to W \mid f \text{ is a linear map such that } f(\rho_g(v)) = \psi_g(f(v))\}.$$

If $f \in \mathrm{Hom}_G(\rho, \psi)$ we say f is G-equivariant.

We may also write this as $\mathrm{Hom}_G(V, W)$. In particular, $\mathrm{Hom}_G(\rho, \rho) = \mathrm{End}_G(\rho)$ is the set of G-equivariant endomorphisms (where $f(\rho_g(v)) = \rho_g(f(v))$ for all g).

Theorem 2.41 (Schur's Lemma)
Let $\rho : G \to \mathrm{GL}(V)$ and $\psi : G \to \mathrm{GL}(W)$ be irreducible representations. Then

$$\dim \mathrm{Hom}_G(\rho, \psi) = \begin{cases} 0 & \rho \ncong \psi \\ 1 & \rho \cong \psi. \end{cases}$$

If $\rho \cong \psi$, then

$$\mathrm{End}_G(\rho) = \mathbb{C} \cdot \mathrm{id}$$

(meaning any G-equivariant endomorphism is scalar).

The first statement means that if $\rho \ncong \psi$, then the only G-equivariant homomorphism is the zero map; the second means that if $\rho \cong \psi$, the only G-equivariant maps are the isomorphism we already know between the two representations, and its scalar multiples — equivalently, $\mathrm{End}_G(\rho) = \mathbb{C} \cdot \mathrm{Id}$ (meaning any G-equivariant endomorphism is scalar).

Proof. Suppose $f : V \to W$ is a nonzero G-equivariant map. Then $\mathrm{im}(f)$ is a G-invariant subspace of W, as

$$\psi_g(f(v)) = f(\rho_g(v)) \in \mathrm{im}(f).$$

But $\mathrm{im}(f) \neq 0$, and ψ is irreducible. So then we must have $\mathrm{im}(f) = W$ — meaning f is surjective.

Meanwhile, $\ker(f)$ is a G-invariant subspace of V, as if $f(v) = 0$, then

$$f(\rho_g(v)) = \psi_g(f(v)) = \psi_g(0) = 0.$$

But since f is nonzero, $\ker(f) \neq V$, so we must have $\ker(f) = 0$ — meaning f is injective.

So f is an isomorphism, which means $\rho \cong \psi$.

Now suppose $\rho \cong \psi$. Then we may assume $\rho = \psi$, so f is a map from V to itself. Let f have eigenvalue λ. Then $f - \lambda \,\mathrm{Id}$ must be G-equivariant as well. But since its kernel is nonzero (it contains an eigenvector of λ), it must be the entire vector space — so $f - \lambda \,\mathrm{Id} = 0$, and f is scalar. □

Remark 2.42. The first case — when $\rho \ncong \psi$ — works for any field of coefficients, since we didn't use anything specific to $\mathbb{C}$. But the second doesn't — we used the existence of an eigenvalue, which isn't necessarily true over $\mathbb{R}$. (In fact, there exist irreducible representations over $\mathbb{R}$ for which the second statement *isn't* true, and there are more G-equivariant endomorphisms.)

Corollary 2.43

Let $\rho : G \to \mathrm{GL}(V)$ be a representation. Let $\rho_1, \ldots, \rho_n$ be a list of all irreducible representations. Then

$$\rho \cong \bigoplus_{i=1}^{n} \rho_i^{d_i},$$

where for each i,

$$d_i = \dim \mathrm{Hom}_G(\rho_i, \rho).$$

Proof. From Maschke's Theorem, we know ρ can be written in this form for some d_i. But we have

$$\mathrm{Hom}_G(\rho_k, \bigoplus \rho_i^{d_i}) = \bigoplus \mathrm{Hom}_G(\rho_k, \rho_i)^{d_i} = \mathbb{C}^{d_k},$$

since all summands with $i \neq k$ are 0. □

Remark 2.44. In this proof, we used the fact that

$$\mathrm{Hom}_G(\rho, \psi_1 \oplus \psi_2) = \mathrm{Hom}_G(\rho, \psi_1) \oplus \mathrm{Hom}_G(\rho, \psi_2).$$

To see why this is true, suppose ρ, ψ_1, and ψ_2 act on V, W_1, and W_2, respectively. Then $\psi_1 \oplus \psi_2$ acts on $W_1 \oplus W_2$, the space (w_1, w_2) with $w_1 \in W_1$ and $w_2 \in W_2$. An element of $\mathrm{Hom}_G(\rho, \psi_1 \oplus \psi_2)$ is then a linear map from V to $W_1 \oplus W_2$ that is G-equivariant. But specifying a linear map $V \to W_1 \oplus W_2$ is the same as specifying a linear map $V \to W_1$ (for the first coordinate) and $V \to W_2$ (for the second), and since G acts separately on the two spaces W_1 and W_2, our linear map is G-equivariant if and only if the two components are.

§2.4.5 Proof of Orthonormality

We can rewrite Schur's Lemma in matrices. Choose a basis for V and W, so we can write a linear map $V \to W$ as a $m \times n$ matrix, where $n = \dim V$ and $m = \dim W$, mapping $v \in V \mapsto Av \in W$.

Let our representations ρ and ψ have matrix representations $R : G \to \mathrm{GL}_n(\mathbb{C})$ and $S : G \to \mathrm{GL}_m(\mathbb{C})$. Then $A \in \mathrm{Mat}_{m\times n}(\mathbb{C})$ corresponds to $f \in \mathrm{Hom}_G(V, W)$ if and only if

$$AR_g = S_g A \text{ for all } g \in G.$$

We can rewrite this as

$$A = S_g A R_g^{-1}.$$

The space $M = \mathrm{Mat}_{m\times n}(\mathbb{C})$ is a vector space, and we can make it a representation of G, as

$$C_g : A \mapsto S_g A R_g^{-1}.$$

We can check

$$C_{gh}(A) = S_{gh} A R_{gh}^{-1} = S_g S_h A R_h^{-1} R_g^{-1} = C_g C_h(A),$$

so this is a valid representation. Then $\mathrm{Hom}_G(V, W)$ is the space of invariant vectors in that representation (where by vectors, we here mean matrices in $\mathrm{Mat}_{m\times n}(\mathbb{C})$).

We can describe this construction without matrices as well: take

$$M = \mathrm{Hom}_{\mathbb{C}}(V, W)$$

(the space of all homomorphisms $V \to W$, which corresponds exactly to $\mathrm{Mat}_{m\times n}(\mathbb{C})$). Then our representation is

$$\gamma_g : E \mapsto \psi_g E \rho_g^{-1}.$$

Proposition 2.45

For the representation γ above, we have

$$\chi_\gamma = \chi_\psi \overline{\chi_\rho}.$$

The proposition quickly reduces to a statement about matrices:

Lemma 2.46

Let A and B be $n \times n$ and $m \times m$ matrices. Consider the map from $\mathrm{Mat}_{m\times n}(\mathbb{C})$ to itself defined as

$$A \otimes B : E \mapsto BEA.$$

Then we have

$$\mathrm{tr}(A \otimes B) = \mathrm{tr}(A) \cdot \mathrm{tr}(B).$$

Proof. The space of matrices has a basis consisting of the matrices E_{ij} with a 1 at (i, j) and 0's everywhere else. But we can check that $BE_{ij}A$ has $b_{ii}a_{jj}$ in position (i, j). So the trace of $A \otimes B$ is the sum of the diagonal entries in the basis of E_{ij}, which is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} b_{ii}a_{jj} = \sum_{i=1}^{m} b_{ii} \sum_{j=1}^{n} a_{jj} = \mathrm{tr}(B) \cdot \mathrm{tr}(A),$$

as desired. □

Proof of Proposition 2.45. Using the above lemma, we have

$$\chi_\gamma(g) = \mathrm{tr}(\rho_{g^{-1}} \otimes \psi_g) = \mathrm{tr}(\rho_{g^{-1}}) \cdot \mathrm{tr}(\psi_g) = \chi_\rho(g^{-1}) \cdot \chi_\psi(g).$$

We saw earlier that $\chi_\rho(g^{-1}) = \overline{\chi(g)}$, so

$$\chi_\gamma(g) = \overline{\chi_\rho(g)} \cdot \chi_\psi(g),$$

as desired. □

This is enough to prove that the χ_i are orthonormal:

Proof of orthonormality. Let ρ and ψ be irreducible representations, and consider the expression

$$\langle \chi_\psi, \chi_\rho \rangle = \frac{1}{|G|} \sum_{g\in G} \overline{\chi_\rho(g)}\chi_\psi(g).$$

We saw a representation γ whose character is exactly this sum, so

$$\langle \chi_\psi, \chi_\rho \rangle = \frac{1}{|G|} \sum_{g\in G} \chi_\gamma(g) = \mathrm{tr}\, \frac{1}{|G|} \sum_{g\in G} \gamma_g$$

(using the fact that $\mathrm{tr}(A + B) = \mathrm{tr}\, A + \mathrm{tr}\, B$).

But this sum is the averaging operator, so applying it to any vector gives a G-invariant vector: we saw earlier that

$$\gamma_h \sum_{g\in G} \gamma_g v = \sum_{g\in G} \gamma_{hg} v = \sum_{g\in G} \gamma_g v.$$

In γ, vectors are actually homomorphisms $V \to W$, and we saw that the G-invariant vectors are exactly $\mathrm{Hom}_G(V, W)$ — which we described using Schur's Lemma.

So if $\rho \not\cong \psi$, then our sum must map every vector to 0, since 0 is the only G-invariant vector (as by Schur's Lemma, $\mathrm{Hom}_G(V, W)$ only contains 0). Therefore the sum is the 0 matrix, which has trace 0.

Meanwhile, note that if v is G-invariant, then

$$\frac{1}{|G|}\sum_{g\in G}\gamma_g v = \frac{1}{|G|}\sum_{g\in G} v = v.$$

If $\rho \cong \psi$, then there is exactly one invariant vector up to scaling (since the invariant vectors are exactly scalar homomorphisms, again by Schur's Lemma).

Now choose a basis (for the space of homomorphisms) where that invariant vector is the first element of the basis.

Then our operator $\frac{1}{|G|}\sum \gamma_g$ sends that vector to itself, and everything else to a scalar multiple of that vector (because they are sent to *some* invariant vector, and these are the only invariant vectors). So in this basis, it must be of the form

$$\begin{bmatrix} 1 & * & \cdots & * \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

which means its trace is 1. □

Corollary 2.47

Any representation $\rho : G \to \mathrm{GL}(V)$ can be split as a sum of irreducibles as

$$\rho \cong \bigoplus \rho_i^{n_i},$$

where for each i,

$$n_i = \langle \chi_\rho, \chi_i \rangle.$$

Proof. We know ρ can be written in this form for some n_i, by Maschke's Theorem. But then

$$\chi_\rho = \sum n_i \chi_i,$$

which means

$$\langle \chi_\rho, \chi_j \rangle = \sum n_i \langle \chi_i, \chi_j \rangle = n_j$$

by orthonormality. □

Unlike the previous formula (using Schur's Lemma), this is a very concrete expression that we can evaluate.

Example 2.48

Describe the dimension of the space of invariant vectors in a representation ρ.

Solution. This dimension is the multiplicity n_1 of the trivial representation in the above decomposition. But we have

$$n_1 = \langle \chi_\rho, \chi_1 \rangle = \frac{1}{|G|}\sum_{g\in G} \chi_\rho(g),$$

since $\chi_1(g) = 1$ for all g. □

Corollary 2.49

We have

$$\langle \rho, \rho \rangle = \sum n_i^2.$$

In particular, ρ is irreducible if and only if $\langle \rho, \rho \rangle = 1$.

Proof. Write $\rho = \bigoplus \rho_i^{n_i}$. Then we can expand

$$\langle \chi_\rho, \chi_\rho \rangle = \sum_i \sum_j n_i n_j \langle \chi_i, \chi_j \rangle = \sum n_i^2$$

by orthonormality. □

§2.4.6 The Regular Representation

We've shown that the characters χ_i are orthonormal, which means they are linearly independent. But to show that they form a basis, we also need to show that they span the space of class functions.

To do this (and prove the sum of squares formula), we'll introduce the regular representation.

If G acts on a finite set X, then we can form a matrix representation of G of dimension $n = \dim(X)$, where G acts by permutation matrices: every element $g \in G$ is mapped to the permutation matrix of how it permutes the elements of X.

For example, S_n acts on the set of n elements, which we used to find a n-dimensional representation of S_n.

But any group acts on itself by left multiplication. So we can take X to be G itself.

Definition 2.50. Let V be a vector space with basis v_h indexed by elements $h \in G$. Then the *regular representation* of G is the representation $\rho : G \to \mathrm{GL}(V)$ such that

$$\rho_g(v_h) = v_{gh}$$

for all g and h.

So in the regular representation, $g \in G$ is sent to the permutation matrix of how left multiplication by g permutes the elements of G.

Example 2.51

Find the regular representation of $\mathbb{Z}/3$.

Solution. Adding $\overline{1}$ corresponds to the permutation (123), so then

$$\overline{1} \mapsto \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Since $\overline{1}$ generates the group, this suffices to determine the representation. □

Example 2.52

Find the regular representation of S_3.

Solution. A transposition $g = (12)$ swaps pairs of permutations, so if we order the rows and columns so that permutations swapped by (12) are consecutive, then it acts as a block-diagonal matrix with three blocks of the form

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

We can find the matrices corresponding to the other permutations similarly. □

The regular representation has exactly one invariant vector up to scaling — the sum of all basis elements. (If $\sum a_h v_h$ is an invariant vector, then we have $a_h = a_{gh}$ for all g and h, which means all a_h must be equal.)

Remark 2.53. A vector in V has the form

$$\sum_{h \in G} a_h v_h.$$

So we can also think of such vectors as $\mathbb{C}$-valued functions on G, which map $h \mapsto a_h$.

Now use ρ to denote the regular representation; we want to describe the character and decomposition of ρ.

Lemma 2.54

The character of the regular representation is

$$\chi_\rho(g) = \begin{cases} |G| & g = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Proof. The trace of a permutation matrix is its number of fixed points. But $h \mapsto gh$ has $|G|$ fixed points if $g = 1$, and no fixed points otherwise: more explicitly, if h were a fixed point under ρ_g, then

$$v_h = \rho_g(v_h) = v_{gh} \implies h = gh,$$

which means $g = 1$. □

So we can decompose ρ into a sum of irreducibles pretty easily:

Lemma 2.55

We have

$$\rho \cong \bigoplus \rho_i^{d_i},$$

where $d_i = \dim \rho_i$ for each i.

Proof. We know $\rho = \bigoplus \rho_i^{n_i}$, where

$$n_i = \langle \chi_i, \chi_\rho \rangle = \frac{1}{|G|} \sum_{g \in G} \chi_i(g) \overline{\chi_\rho(g)} = \frac{1}{|G|} \chi_i(1) \cdot |G| = \chi_i(1).$$

But $\rho_i(1) = 1_{d_i}$ is the identity matrix of dimension d_i, so $\chi_i(1) = d_i$. □

For example, in an abelian group, all dimensions are 1, so $\rho \cong \bigoplus \rho_i$.

Corollary 2.56

$|G| = \sum d_i^2$, where the d_i are the dimensions of all irreducible representations.

Proof. Compare the dimensions of the two sides. We have $\dim \rho = |G|$. Meanwhile, $\bigoplus \rho_i^{d_i}$ contains d_i copies of a d_i-dimensional representation for each i, so its dimension is $\sum d_i^2$. □

§2.4.7 Proof of Span

Finally, we will prove the first part of the main theorem — that the characters of irreducible representations span the space of class functions. We already know that they are linearly independent (because they are orthonormal). So it suffices to show that for any class function f, we have

$$f = \sum \langle f, \chi_i \rangle \chi_i.$$

(We already know this holds for the characters of representations.)

In order to show this, it's clear that the space of class functions with this pairing is a Hermitian space. So it suffices to check the following claim:

Lemma 2.57

If f is a class function and $\langle f, \chi_i \rangle = 0$ for all i, then f is identically 0.

This suffices because for any class function f, we can then define $f' = \sum \langle f, \chi_i \rangle \chi_i$. Then we have

$$\langle f - f', \chi_i \rangle = \langle f, \chi_i \rangle - \left\langle \sum \langle f, \chi_j \rangle \chi_j, \chi_i \right\rangle = \langle f, \chi_i \rangle - \langle f, \chi_i \rangle = 0$$

for each i, which means $f = f'$.

Now we will prove this claim.

Definition 2.58. Given any function $f : G \to \mathbb{C}$ and a representation $\rho : G \to \mathrm{GL}(V)$, define

$$\rho(f) = \sum_{g \in G} f(g) \rho_g \in \mathrm{End}(V).$$

This essentially defines ρ on linear combinations of group elements.

Lemma 2.59

We have

$$\operatorname{tr} \rho(f) = \langle \chi_\rho, \overline{f} \rangle.$$

Proof. This follows from linearity of trace: we have

$$\langle \chi_\rho, \overline{f} \rangle = \sum_{g \in G} \chi_\rho(g) f(g) = \sum_{g \in G} f(g) \operatorname{tr} \rho_g = \operatorname{tr} \sum_{g \in G} f(g) \rho_g = \operatorname{tr} \rho(f).$$

□

Lemma 2.60

If f is a class function, then $\rho(f) \in \mathrm{End}_G(\rho)$.

Proof. Recall that class functions are functions which are invariant under conjugation: so they take a fixed value on every conjugacy class. So then $\rho(f)$ must be invariant under conjugation as well: we have

$$\rho_g \rho(f) \rho_g^{-1} = \sum_{h \in G} f(h) \rho_{ghg^{-1}} = \sum_{h \in G} f(h) \rho_h = \rho(f),$$

since $f(ghg^{-1}) = f(h)$ because f is a class function. So $\rho_g(\rho(f)v) = \rho(f)(\rho_g(v))$ for all v. □

Now we prove our initial claim, that if f is a class function with $\langle f, \chi_i \rangle = 0$ for all i, then f is identically 0:

Proof. Since $\langle \chi_i, f \rangle = 0$ for all i, we have

$$\operatorname{tr} \rho_i(\overline{f}) = 0$$

for all i as well. But $\rho_i(\overline{f}) \in \mathrm{End}_G(\rho_i)$, so by Schur's Lemma, it must be a scalar matrix. The only scalar matrix with trace 0 is the 0 matrix, so $\rho_i(\overline{f})$ is the 0 matrix for all i.

But by Maschke's Theorem, every representation is the sum of irreducibles. So since $\rho_i(\overline{f}) = 0$ for all irreducible representations ρ_i, we must have that $\rho(\overline{f}) = 0$ for *any* representation ρ.

But now we can take ρ to be the regular representation. We can actually read off f from its action in the regular representation: we have

$$\rho(f)(v_1) = \sum_g f(g) \rho_g v_1 = \sum_g f(g) v_g.$$

So if $\rho(f)$ is 0, then f is identically 0. □

§2.4.8 Proof of Divisibility

Finally, we'll show that if $\rho : G \to \mathrm{GL}(V)$ is an irreducible representation of dimension d, then $d \mid |G|$. *(This was not proved in class; these notes are from the writeup posted to Canvas.)*

Recall that we defined

$$\rho(f) = \sum_{g \in G} f(g) \rho(g) \in \mathrm{End}(V).$$

Proposition 2.61

We have

$$\rho(\overline{\chi_\rho}) = \frac{|G|}{d} \cdot \mathrm{Id}\,.$$

Proof. Since $\overline{\chi_\rho}$ is a class function, by Schur's Lemma we know $\rho(\overline{\chi_\rho})$ is of the form $\lambda \,\mathrm{Id}$ for some λ. Now we can compute λ by taking the trace: we saw earlier that $\operatorname{tr} \rho(f) = |G| \langle \chi_\rho, \overline{f} \rangle$, so

$$\operatorname{tr} \rho(\overline{\chi_\rho}) = |G| \langle \chi_\rho, \chi_\rho \rangle = |G|$$

by orthonormality. But this trace is $d\lambda$, so we must have $\lambda = \frac{|G|}{d}$. □

So now we have a natural construction where $\frac{|G|}{d}$ comes up. In order to prove it is an integer, we will use algebraic integers:

Definition 2.62. A complex number is a *algebraic integer* if it is the root of a monic polynomial with integer coefficients.

Lemma 2.63

Algebraic integers have the following standard properties:

(a) If α and β are algebraic integers, so are $\alpha+\beta$ and $\alpha\beta$.

(b) If $\alpha \in \mathbb{Q}$ is an algebraic integer, then $\alpha \in \mathbb{Z}$.

We will discuss algebraic integers later in the course.

Proposition 2.64

If f is a function on G such that $f(g)$ is an algebraic integer for every g, and $\rho(f) = r \cdot \mathrm{Id}$ for a rational number r, then r is an integer.

Remark 2.65. In fact, a stronger statement is true — if $f(g)$ is an algebraic integer for all $g \in G$, then every eigenvalue of $\rho(f)$ is an algebraic integer. But this is harder to prove.

Proof. We will show that

$$\operatorname{tr} \rho(f)^n \in \mathbb{Z}$$

for all n, which suffices (as if dr^n is an integer for all n, and $r \in \mathbb{Q}$, then r must be an integer).

When $n = 1$, this is true as

$$\operatorname{tr} \rho(f) = \sum_{g \in G} f(g)\chi_\rho(g),$$

and $f(g)$ and $\chi_\rho(g)$ are both algebraic integers (as $\chi_\rho(g)$ is the sum of roots of unity). So $\operatorname{tr} \rho(f)$ is an algebraic integer and is rational, which means it is an integer.

For $n > 1$, it suffices to find a function f_n with $\rho(f)^n = \rho(f_n)$, where $f_n(g)$ is also always an algebraic integer.

But given two functions on G, we can consider their *convolution*:

$$(\phi * \psi)(g) = \sum_{h \in G} \phi(h)\psi(h^{-1}g).$$

We can check that $\rho(\phi * \psi) = \rho(\phi)\rho(\psi)$, so then

$$f_n = \underbrace{f * f * \cdots * f}_{n}$$

satisfies $\rho(f)^n = \rho(f_n)$, as desired (and f_n consists of sums and products of algebraic integers, so is always an algebraic integer). □

This shows that $\frac{|G|}{d}$ is an integer.

Remark 2.66. Here is a way to think of convolution: the space of functions on G has a basis δ_g mapping g to 1 and all other elements to 0. Then $\delta_g * \delta_h = \delta_{gh}$, and $\rho(\delta_g) = \rho(g)$, so we have

$$\rho(\delta_g * \delta_h) = \rho_{gh} = \rho_g\rho_h = \rho(\delta_g)\rho(\delta_h).$$

The statement for general ϕ and ψ then follows from linearity.

§2.5 Final Remarks

We can also study *continuous* representations of compact subgroups of $\mathrm{GL}_n(\mathbb{C})$ — for example, $U(n)$ and $O(n)$. A lot of the theory still goes through, with some modification: the set of irreducible representations is not finite, but it is discrete. Their characters are orthonormal under the pairing

$$\langle \chi, \psi \rangle = \int_G \chi(g)\overline{\psi(g)}\, dg.$$

For example, consider $U(1) = \{z \in \mathbb{C}^* \mid |z| = 1\}$. The group is abelian, so every irreducible representation has dimension 1, and every function is a class function.

The irreducible representations are $\psi_n : z \mapsto z^n$ for each $n \in \mathbb{Z}$. A continuous function $U(1) = S^1 \to \mathbb{C}$ is the same as a 2π-periodic function $\mathbb{R} \to \mathbb{C}$. All such functions have a *Fourier series* $f(t) = \sum a_n \exp(2\pi i n t)$, where we have

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} f(t)\overline{\exp(2\pi i n t)}\, dt.$$

This is analogous to $f = \sum a_i \chi_i$, where

$$a_i = \frac{1}{|G|} \sum f(g)\overline{\chi_i(g)},$$

in the case where G is finite.

§3 Rings

§3.1 Definitions

Informally, a ring is a set of elements which can be added and multiplied, so that the natural properties we'd expect of addition and multiplication all hold.

Example 3.1

$\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all rings.

There are various rings between $\mathbb{Z}$ and $\mathbb{Q}$: for example, $\mathbb{Z}[\frac{1}{2}] = \{\frac{a}{2^k} \mid a, k \in \mathbb{Z}\}$ is a ring. Similarly, $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ and $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ are rings. So is

$$\mathbb{Z}[\zeta] = \{a_0 + a_1\zeta + \cdots + a_{n-1}\zeta^{n-1} \mid a_i \in \mathbb{Z}\}$$

(where ζ is a nth root of unity).

Definition 3.2. A *ring* R is a set with two binary operations $R \times R \to R$, written as $+$ and $\cdot$, which satisfy the following axioms:

- $(R, +)$ is an abelian group: addition is associative and commutative, there is an additive identity 0_R, and every element has an additive inverse.
- Multiplication is also associative and commutative, and there is a multiplicative identity 1_R. (So under multiplication, R is a *semigroup* — a group without the condition that every element has an inverse.)
- Addition and multiplication satisfy *distributivity*: we have

$$a(b + c) = ab + ac.$$

In this class, we'll use "ring" in this sense; but usually this is called a *commutative unital ring*. In defining a ring, one can drop the commutativity of multiplication, or the existence of 1_R. If everything holds except for $ab = ba$ — and we add distributivity in the other direction as well, meaning $(b + c)a = ba + ca$ — then R is a *noncommutative ring*.

Example 3.3

The set of $n \times n$ matrices, or the set of endomorphisms of any vector space, are noncommutative rings.

Example 3.4

If G is a group, then take the vector space with basis v_g for $g \in G$. If multiplication is defined as $v_g v_h = v_{gh}$, then this forms the *group ring*, which is noncommutative if G is nonabelian.

From now, all rings we will look at will be commutative, unless otherwise specified.

Lemma 3.5

In any ring R, for all $a \in R$, we have

$$0_R \cdot a = 0_R.$$

Proof. Pick any $x \in R$. Then

$$(0_R + x) \cdot a = 0_R \cdot a + x \cdot a.$$

But $0_R + x = x$, so

$$(0_R + x) \cdot a = x \cdot a.$$

Since $x \cdot a$ has an additive inverse, we can cancel it out to get $0_R \cdot a = 0_R$. □

Corollary 3.6

0_R cannot have a (multiplicative) inverse unless $0_R = 1_R$.

The axioms do not prohibit $0_R = 1_R$, but this only results in trivial examples: if $0_R = 1_R$, then for any $x \in R$ we get

$$x = x \cdot 1_R = x \cdot 0_R = 0_R,$$

so the ring only contains one element. This is the *zero ring*, which is a legitimate but trivial example. In all other rings, $0_R \neq 1_R$, as we would expect.

Definition 3.7. A ring where every nonzero element has a (multiplicative) inverse, and $0_R \neq 1_R$, is called a *field*.

Example 3.8

$\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields; $\mathbb{Z}$ and $\mathbb{Z}[i]$ are not fields.

$\mathbb{Z}/n$ is a field if and only if n is prime.

By definition, the zero ring is not a field.

Example 3.9

$\mathbb{C}[x]$, the set of polynomials in one variable with complex coefficients, is a ring. The set $C^\infty(\mathbb{R})$ — the set of real-valued functions which are continuous and infinitely differentiable — is also a ring.

§3.2 Homomorphisms

Definition 3.10. A *homomorphisms* of rings is a map $\varphi : R \to S$ such that:

- $\varphi(a+b) = \varphi(a) + \varphi(b)$;
- $\varphi(ab) = \varphi(a)\varphi(b)$;
- $\varphi(1_R) = \varphi(1_S)$.

Example 3.11

The map $\mathbb{Z} \to \mathbb{Z}/n$ with $a \mapsto \overline{a} = a \mod n$ is a ring homomorphism.

Remark 3.12. The definition does not require $\varphi(0_R) = 0_S$ ecause this is implied by the first property: since we have additive inverses

$$\varphi(0_R) + \varphi(a) = \varphi(0_R + a) = \varphi(a) \implies \varphi(0_R) = 0_S.$$

But this does not work with multiplication, since we do not necessarily have inverses so cannot cancel. There are maps satisfying the first two conditions but not the third: as a trivial example, take R to be the zero ring and S to be any nonzero ring, and map $0_R \mapsto 0_S$.

If φ is a bijection, it is easy to check that its inverse is a homomorphism as well.

Definition 3.13. φ is an *isomorphism* if it is a bijective homomorphism.

If φ is one-to-one (but not necessarily onto), then φ identifies R with a *subring* of S — it's bijective on its image, which must be closed under addition, multiplication, and subtraction, and must contain 1_S.

Definition 3.14. The *kernel* of φ is the set of $a \in R$ such that $\varphi(a) = 0_S$.

Then the kernel is an additive subgroup. But it's also compatible with multiplication in some ways: it turns out that $\ker \varphi$ is an *ideal*.

§3.3 Ideals

Definition 3.15. An *ideal* is a subset $I \subset R$ such that I is an additive subgroup of R (meaning that it is closed under addition and taking additive inverses), and for any $a \in R$ and $x \in I$, we have $ax \in I$.

Then $\ker \varphi$ is an ideal because if $\varphi(x) = 0$, then $\varphi(ax) = \varphi(a)\varphi(x) = 0$ as well.

Example 3.16

If $n \in \mathbb{Z}$, then $n\mathbb{Z}$ (the set of integers divisible by n) is an ideal.

In general, for any ring R and any element $a \in R$, the set $aR = \{ax \mid x \in R\}$ is an ideal.

Definition 3.17. The ideal $aR = \{ax \mid x \in R\}$ is called a *principal ideal*, and is denoted (a).

We saw in 18.701 that every additive subgroup of $\mathbb{Z}$ is cyclic, meaning of the form $n\mathbb{Z}$. So every ideal of $\mathbb{Z}$ must be principal. But this isn't true in more general rings — we will see many examples later.

Definition 3.18. The ideal generated by $a_1, \ldots, a_n$ is the smallest ideal containing each of $a_1, \ldots, a_n$:

$$(a_1, \ldots, a_n) = \left\{ \sum a_i x_i \mid x_i \in R \right\}.$$

Ideals arise from kernels, and they play a similar role as normal subgroups did in group theory.

Remark 3.19. The name *ideals* comes from *ideal divisors* — we will discuss this later.

Note that ideals I are usually not rings, as they do not usually contain 1_R. If $1_R \in I$, then we must have $I = R$, since for every $x \in R$, we have $1_R x = x \in I$.

Definition 3.20. Let R be a ring, and $I \subset R$ an ideal. Then the *quotient ring* R/I is the quotient of the additive groups $(R, +)$ and $(I, +)$, where if we denote the coset of $x \in R$ as $x + I = \overline{x}$, then we define multiplication as

$$\overline{x} \cdot \overline{y} = \overline{xy}.$$

This is well-defined because if $a \in I$, then

$$(x + a)y = xy + ay,$$

and $ay \in I$. We can check that all the ring axioms are satisfied.

Many properties here are similar to what happens when we quotient a group by a normal subgroup. For example, if φ is a homomorphism, then there is an isomorphism from $R/\ker\varphi$ to $\operatorname{im}\varphi$. The Correspondence Theorem holds as well: there is a bijection between ideals in R/I and ideals in R containing I.

Example 3.21

If $R = \mathbb{Z}$ and $I = n\mathbb{Z}$, then $R/I = \mathbb{Z}/n$.

§3.4 Building New Rings

§3.4.1 Product Rings

Definition 3.22. Let R and S be rings. Then we define their *Cartesian product* as

$$R \times S = \{(r, s) \mid r \in R, s \in S\}.$$

Addition and multiplication are defined componentwise.

The axioms clearly hold — in particular, $1_{R\times S} = (1_R, 1_S)$.

Given a product ring $R \times S$, we have the *projection homomorphism* $R \times S \to R$ given by $(r, s) \mapsto r$. The kernel of this homomorphism is the set $(0, s)$, which is the ideal generated by $(0, 1_S)$.

Question 3.23. Given a ring Q, how can we recognize whether it's isomorphic to $R \times S$ for nonzero rings R and S?

First, if $Q = R \times S$, then consider the two elements $e_1 = (1_R, 0)$ and $e_2 = (0, 1_S)$. These are not units, but they have interesting properties:

Definition 3.24. An element e is *idempotent* if $e^2 = e$.

Equivalently, e is idempotent if $e(1-e) = 0$ — in particular, if e is idempotent, then so is $1 - e$.

Then e_1 and e_2 are both idempotent.

Remark 3.25. In linear algebra, if a matrix E is idempotent, then its only eigenvalues are 0 and 1. So it splits as the direct sum of eigenspaces $E = V_1 \oplus V_0$. We can do something similar in a ring.

Remark 3.26. All rings have the idempotents 1 and 0. If R is a field, then there are no others — $e(1-e) = 0$ implies $e = 0$ or 1. But this is not true for general rings.

Conversely, given an idempotent $e \in Q$, we can set $R = eQ$ and $S = (1-e)Q$. Then we can check that R and S are rings — essentially, we can multiply because $e^2 = e$, and the unit in R is e. Then $Q \cong R \times S$, since any $x \in Q$ can be written as $ex + (1-e)x$ — the inverse map is $(r, s) \mapsto r + s$. (This is an isomorphism because $e(1-e) = 0$, so "mixed" terms disappear when we multiply.)

Remark 3.27. R is not a subring of Q in our terminology, because $1_Q \notin R$.

Similarly, the map $R \to R \times S$ with $r \mapsto (r, 0)$ is compatible with addition and multiplication, but is not a homomorphism because 1_R is not sent to $1_{R \times S}$.

Combining these answers our question about when we can split a ring:

Proposition 3.28

A ring Q is isomorphic to a product $R \times S$ of nonzero rings R and S if and only if Q contains a nontrivial idempotent.

We can define the product of a collection of rings (which may be finite or infinite) in the same way.

§3.4.2 Adjoining Elements

Definition 3.29. If R is a subring of S, and $\alpha \in S$, then the ring $R[\alpha]$ is defined as the smallest subring of S containing both R and α.

This means

$$R[\alpha] = \left\{ \sum_{i=0}^{n} r_i \alpha^i \mid r_i \in R \right\},$$

since clearly all such elements must be in the subring, and this set is closed.

Example 3.30

When $S = \mathbb{C}$, $R = \mathbb{Z}$, and $\alpha = i$, we get the *Gaussian integers* $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$.

When $R = \mathbb{Q}$ and $\alpha = \sqrt{2}$, we get $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$.

WHen $R = \mathbb{Z}$ and $\alpha = \frac{1}{2}$, we get that $\mathbb{Z}[\frac{1}{2}]$ is the set of fractions whose denominator is a power of 2.

Similarly, if we start with elements $\alpha_1, \ldots, \alpha_n$ in S, then $R[\alpha_1, \ldots, \alpha_n]$ is the smallest subring of S containing R and all the α_i. It consists of all sums of products of powers of the α_i, with coefficients in R.

There is another way to think of adjoining elements:

Definition 3.31. Let R be a ring, and x a formal variable. Then the *polynomial ring* $R[x]$ is the set

$$R[x] = \left\{ \sum_{i=0}^{n} r_i x^i \mid r_i \in R \right\}.$$

Remark 3.32. For $R = \mathbb{C}$, $\mathbb{R}$, or $\mathbb{Q}$, we can think of $R[x]$ as the ring of polynomial functions from R to itself as well — but this isn't true in general.

In general, given some $\alpha \in R$ and $P \in R[x]$, we can still plug in α for x and get an element of R — so a polynomial P does give a function $R \to R$. But it can't necessarily be recovered from the function — so it contains more information than just the function, and we should think of polynomials in terms of formal expressions rather than functions.

For example, $P(x) = x^p - x$ is 0 for all $x \in \mathbb{F}_p$, so is the zero function but a nonzero polynomial. (In fact, the set of functions on $\mathbb{F}_p$ is finite-dimensional, while the space of polynomials is infinite-dimensional.)

Definition 3.33. The *evaluation homomorphism* at α is the map $R[x] \to R$ which sends $x \mapsto \alpha$.

We can define the ring of polynomials in multiple variables — denoted $R[x_1, \ldots, x_n]$ — in the same way.

Proposition 3.34 (Mapping Property)

Suppose we have a ring R, and a ring homomorphism $\varphi : R \to S$. Given $\alpha_1, \ldots, \alpha_n \in S$, there exists a unique extension of φ to a homomorphism $\tilde{\varphi} : R[x_1, \ldots, x_n] \to S$, such that $\tilde{\varphi}(r) = \varphi(r)$ for $r \in R$, and $\tilde{\varphi}(x_i) = \alpha_i$ for all i.

This is much less complicated than it appears. The unique extension is just evaluation:

$$\tilde{\varphi}\left(\sum r_{i_1 \ldots i_n} x_1^{i_1} \cdots x_n^{i_n}\right) = \sum \varphi(r_{i_1 \ldots i_n}) \alpha_1^{i_1} \cdots \alpha_n^{i_n}.$$

This follows directly from the properties of a homomorphism.

But it gives us another way of looking at our original definition: if $R \subset S$ is a subring, then

$$R[\alpha_1, \ldots, \alpha_n] = R[x_1, \ldots, x_n] / \ker \tilde{\varphi}.$$

(Here φ is the inclusion map $r \mapsto r$.)

Example 3.35

We have $\mathbb{Q}[\sqrt{2}] = \mathbb{Q}[x]/(x^2 - 2)$.

Similarly, $\mathbb{C} = \mathbb{R}[x]/(x^2 + 1)$. This is essentially how $\mathbb{C}$ is *constructed* — we define i as an abstract variable postulated to satisfy $i^2 = -1$.

So we can construct new rings as the quotient of a polynomial ring $R[x_1, \ldots, x_n]$ by ideals. If we want an element with a certain property, then we can add a variable and say that it satisfies that property (with a bit more work).

§3.4.3 Ideals in Polynomial Rings

Proposition 3.36

A ring R is a field if and only if it has exactly two ideals.

Any ring has the ideals $(0) = \{0\}$ and $(1) = R$ — these coincide only in the zero ring, which is not a field. So the proposition states that the ring is a field if and only if it has no other ideals.

Proof. Suppose R is a field, so it has at least two ideals (0) and (1). But there are no other ideals, because every element is invertible — if I is an ideal containing some element $x \neq 0$, then $1 = x^{-1}x$ is in I as well, so $I = (1)$.

Conversely, if R is not a field, then either it is the zero ring and only has one ideal, or it contains a nonzero x which is not invertible. Then (x) cannot contain 1, so (0), (x), and (1) are distinct ideals. □

Proposition 3.37

Every ideal in $F[x]$ is principal. More precisely, if $I \subset F[x]$ is a nonzero ideal and P a (nonzero) element of I with minimal degree, then $I = (P)$ and the images of 1, x, x^2, …, x^{n-1} form a basis in $F[x]/I$ (where $n = \deg P$).

Proof. This follows from division of polynomials with remainder: if $Q \in I$, then by polynomial division we can write

$$Q = P \cdot S + R,$$

where S and R are in $F[x]$ and $\deg R < \deg P$. But R must be in I, so if $R \neq 0$ then this contradicts the choice of P as having minimal degree. So $R = 0$, which means $I = (P)$.

Then considering the quotient $F[x]/(P)$, the images $\overline{1}$, $\overline{x}$, …, $\overline{x^{n-1}}$ must be linearly independent: since F is a field (so has no zero divisors), we have $\deg PQ = \deg P + \deg Q$, so a polynomial of degree less than n cannot be divisible by P.

Meanwhile, they span the quotient because of division with remainder again: we can write $Q = P \cdot S + R$ with $\deg R < n$, which means $\overline{Q} = \overline{R}$. □

Remark 3.38. This doesn't generalize fully to polynomial rings over an arbitrary ring R, but some parts do. First, it's not always true that

$$\deg PQ = \deg P + \deg Q.$$

(For example, in $\mathbb{Z}/4[x]$, $(2x)(2x+1) = 2x$ does not have degree 2.) Division with remainder also does not necessarily work: for example, we can't divide x^2 by $2x+1$ with remainder in $\mathbb{Z}[x]$.

But both work for monic polynomials. So if P is monic, then it's still true that every element of $R[x]/(P)$ can be written uniquely as

$$a_0\overline{1} + a_1\overline{x} + \cdots + a_{n-1}\overline{x^{n-1}},$$

for $a_i \in R$.

Last time, we said that $F[x]/(P)$ is like adjoining a root of P to F — for example, $\mathbb{C} = \mathbb{R}[x]/(x^2+1)$. Using this lemma, we can prove this: if we let $\alpha = \overline{x}$, then $R[x]/P$ consists of elements $a_0 + \cdots + a_{n-1}\alpha^{n-1}$. It's enough to understand how to multiply by α, and we can do this by using the polynomial relation given by $P(\alpha) = 0$.

§3.5 Maximal Ideals

Definition 3.39. An ideal $I \subsetneq R$ is *maximal* if the only ideal J strictly containing I is R itself.

In other words, I is not the entire ring, and the only ideals containing it are the obvious ones (I and R).

Example 3.40

Find the maximal ideals of $\mathbb{Z}$.

Proof. We saw earlier that the ideals of $\mathbb{Z}$ are $n\mathbb{Z}$. But $n\mathbb{Z}$ is contained in $m\mathbb{Z}$ iff $m \mid n$, so the maximal ideals are (p) for primes $p \in \mathbb{Z}$. □

Example 3.41

Find the maximal ideals of $F[x]$ for a field F.

Proof. Again, all ideals are of the form (P) for a polynomial P. But $(P) \subset (Q)$ if and only if $Q \mid P$, so the maximal ideals are (P) where P is *irreducible*, meaning P cannot be written as QR for Q and R in $F[x]$ with positive degree. □

Remark 3.42. The irreducible polynomials in $\mathbb{C}[x]$ are exactly the polynomials of degree 1, since the Fundamental Theorem of Algebra states that every polynomial over $\mathbb{C}$ is a product of linear factors. So the maximal ideals of $\mathbb{C}[x]$ are exactly $(x - a)$ for $a \in \mathbb{C}$.

Proposition 3.43

An ideal $I \subset R$ is maximal if and only if R/I is a field.

Proof. We know R/I is a field if and only if it has two ideals. But the Correspondence Theorem gives a direct correspondence between ideals in R/I and ideals of R containing I, so this is equivalent to there being exactly two ideals containing I, which is equivalent to I being maximal. □

Example 3.44

Let $R = F[x_1, \ldots, x_n]$, where F is a field. Fixing scalars $\alpha = (\alpha_1, \ldots, \alpha_n)$ (with $\alpha_i \in F$), we have the evaluation map $F[x_1, \ldots, x_n] \to F$ defined as

$$\mathrm{ev}_\alpha : P \mapsto P(\alpha_1, \ldots, \alpha_n).$$

The kernel of this map

$$\mathfrak{m}_\alpha = (x_1 - \alpha_1, \ldots, x_n - \alpha_n)$$

is a maximal ideal of R.

Remark 3.45. Note that $R/\mathfrak{m}_\alpha \cong F$ is a field, since we can repeatedly use polynomial division in order to replace all the x_i with α_i.

Similarly, we can take $R = F[x_1, \dots, x_n]/J$, where $J = (P_1, \dots, P_m)$. Then the $\mathfrak{m}_\alpha$ which contain J give maximal ideals of R. In order for $\mathfrak{m}_\alpha$ to contain J, all the P_i should be in $\ker \mathrm{ev}_\alpha$, meaning that α is a common root of the polynomials $P_1, \dots, P_m$.

§3.5.1 Hilbert's Nullstelensatz

Theorem 3.46 (Hilbert's Nullstelensatz)

The maximal ideals of $\mathbb{C}[x_1, \dots, x_n]$ are exactly the kernels of evaluation homomorphisms, and are thus in bijection with $\mathbb{C}^n$.

Proof. We already saw that the kernels $\mathfrak{m}_\alpha$ are maximal ideals of $\mathbb{C}[x_1, \dots, x_n]$, so it suffices to show that they are the only maximal ideals.

Let $\mathfrak{m} \subset \mathbb{C}[x_1, \dots, x_n]$ be a maximal ideal; then $F = \mathbb{C}[x_1, \dots, x_n]/\mathfrak{m}$ is a field. This gives a homomorphism from $\mathbb{C}$ to F (since it gives a homomorphism from $\mathbb{C}[x_1, \dots, x_n]$ to F); we want to show that this homomorphism is an isomorphism.

But *any* homomorphism between fields is injective — the kernel of the homomorphism must be an ideal, but the only ideals of a field are $\{0\}$ and the entire field. Since the homomorphism cannot map 1 to 0, the kernel cannot be the entire field, so must be $\{0\}$.

So it suffices to show that the homomorphism is surjective. Assume not. Then F strictly contains $\mathbb{C}$, so we can pick $z \in F$ with $z \notin \mathbb{C}$. Then consider the elements

$$\left\{ \frac{1}{z - \lambda} \mid \lambda \in \mathbb{C} \right\}.$$

The point is that $\mathbb{C}[x_1, \dots, x_n]$ is a countable union of finite-dimensional vector spaces over $\mathbb{C}$. But $\mathbb{C}$ is not countable, so there are uncountably many terms $\dfrac{1}{z - \lambda}$, and therefore infinitely many must fall into the same finite-dimensional vector space.

But then they must satisfy some linear dependence

$$\sum \frac{a_i}{z - \lambda_i} = 0,$$

where the $a_i \in \mathbb{C}$ and there are finitely many terms. But this can be written as a polynomial $P(z) = 0$, with complex coefficients. Since all polynomials over $\mathbb{C}$ factor, we can write

$$P(x) = c \prod_i (x - r_i),$$

for $r_i \in \mathbb{C}$. But $z \notin \mathbb{C}$, so z cannot equal any of the r_i, contradiction (as in a field, the product of nonzero terms cannot be zero). □

Remark 3.47. The key point of this proof was that the statement reduces to showing that if F is a field containing $\mathbb{C}$, which is finitely generated as a ring over $\mathbb{C}$ — meaning there exists a surjective map $\mathbb{C}[x_1, \dots, x_n] \to F$ — then we must have $F = \mathbb{C}$.

The proof doesn't use the fact that F is the quotient of a polynomial ring that much — it's used just to show that F is a countable union of finite-dimensional $\mathbb{C}$-vector spaces.

Remark 3.48. This theorem is true for all algebraically closed fields.

Corollary 3.49

The maximal ideals of $\mathbb{C}[x_1, \ldots, x_n]/(P_1, \ldots, P_m)$ are in bijection with the common roots of the polynomials P_1, …, P_m (more precisely, they are the kernels of the evaluation homomorphisms at those roots).

This follows from the fact that ideals in R/I are in correspondence with ideals in R containing I.

Definition 3.50. For a ring R, we define $\mathrm{MSpec}(R)$ as the set of all maximal ideals of R.

In algebraic geometry and commutative algebra, people try to work with $\mathrm{MSpec}(R)$ as a geometric object. Each element $r \in \mathbb{R}$ defines a "function" f_r on $\mathrm{MSpec}(R)$, where we send a maximal ideal $\mathfrak{m}$ to the element $\overline{r}$ in $R/\mathfrak{m}$ (which is a field).

If $R = \mathbb{C}[x_1, \ldots, x_n]/I$ is the quotient of a polynomial ring, then $R/\mathfrak{m} = \mathbb{C}$, so f_r is a function $\mathrm{Mspec}(R) \to \mathbb{C}$ (given by evaluating the polynomial r at the point corresponding to $\mathfrak{m}$). But in general, there isn't even a guarantee that the fields $R/\mathfrak{m}$ are all isomorphic — they may be different for different $\mathfrak{m}$, which is why "function" is in quotation marks.

§3.6 Inverting Elements and Fraction Fields

We've previously looked at the structure of $R[x]/(P)$ for a monic polynomial P, and seen that this essentially adds a root of P into our ring.

Similarly, we can consider $R[x]/(ax-1)$. This essentially adds a variable x and declares it to be the inverse of a — so this ring is the result of formally inverting a. We call this ring the *localization* of R at a, denoted $R_{(a)}$.

Example 3.51

If we invert 2 in $\mathbb{Z}$, we get

$$\mathbb{Z}_{(2)} \cong \left\{ \frac{a}{2^n} \mid n \in \mathbb{Z} \right\}.$$

If we invert a composite number, such as 6, then we get

$$\mathbb{Z}_{(6)} \cong \left\{ \frac{a}{2^m 3^n} \mid m, n \in \mathbb{Z} \right\}.$$

But we have to be careful when we say that we're adding an inverse of an element — it's possible that this procedure collapses some of R as well.

Lemma 3.52

If $ab = 0$, then the image of b in $R_{(a)}$ is 0.

Proof. The image of ab in $R_{(a)}$ must be 0. But the image of a is invertible, so the inverse of b must be 0. □

Example 3.53

If we invert 2 in $\mathbb{Z}/6$, then 3 will die. In fact, $(\mathbb{Z}/6)_{(2)} \cong \mathbb{Z}/3$. Similarly, $(\mathbb{Z}/4)_{(2)}$ is the zero ring.

Definition 3.54. An element $a \in R$ is a *zero divisor* if $a \neq 0$, but $ab = 0$ for some $b \neq 0$.

For example, 2 and 3 are zero divisors in $\mathbb{Z}/6$.

If a is not a zero divisor, then we have $R \subset R_{(a)}$, so inverting a doesn't collapse the ring.

Definition 3.55. A ring R is an *integral domain* if it has no zero divisors.

This concept is useful because in an integral domain, if $ax = ay$ with $a \neq 0$, then we must have $x = y$ — so cancellation works.

So we can invert elements of an integral domain, using this construction $R_{(a)}$. We can invert finitely many elements in a similar way. But we can also invert *every* element:

Definition 3.56. Let R be an integral domain. Then the *fraction field* of R is

$$\mathrm{Frac}(R) = \{(a, b) \mid a, b \in R, b \neq 0\}$$

modulo the equivalence relation $(a, b) \sim (c, d)$ if $ad = bc$.

This is the formal definition, but we usually write elements as $\frac{a}{b}$ instead of (a, b), and operations work in the same way as the usual operations on fractions.

Note that $\mathrm{Frac}(R)$ is a field containing R, where every nonzero element $\frac{a}{b}$ has an inverse $\frac{b}{a}$.

Example 3.57
If $R = \mathbb{Z}$, then $\mathrm{Frac}(R) = \mathbb{Q}$.

Example 3.58
If $\mathbb{R} = \mathbb{C}[x]$, then $\mathrm{Frac}(R)$ is the field of *rational functions* in one variable.

We can define the field of rational functions on n variables in a similar way. Such functions are defined almost everywhere, and have poles at a few points.

Example 3.59
If R is a field F, then $\mathrm{Frac}(R)$ is the field F as well.

Giving a homomorphism $\mathrm{Frac}(R) \to S$ is the same as giving a homomorphism $R \to S$ sending every nonzero element of R to a unit (an invertible element) in S.

§4 Factorization

§4.1 An Example

Proposition 4.1
Every polynomial $P \in F[x]$ factors as a product of irreducible polynomials in an essentially unique way.

In order to prove this, we'll first prove the following lemma:

Lemma 4.2

If an irreducible polynomial $P \in F[x]$ divides QS (for polynomials $Q, S \in F[x]$), then P must divide Q or S.

Proof. We know that $F[x]/(P)$ is a field, and a polynomial is divisible by P iff its image in $F[x]/(P)$ is 0. So $\overline{QS} = 0$ in $F[x]/(P)$, and since a field has no zero divisors, we must have $\overline{Q} = 0$ or $\overline{S} = 0$. □

Using this, we can prove unique factorization in the usual way:

Proof of 4.1. To show existence of a factorization into irreducibles, we can just keep factoring until stuck (more formally, we can induct on degree). To prove that this factorization is unique, suppose

$$P = P_1 \cdots P_n = Q_1 \cdots Q_m.$$

Then P_1 must divide Q_i for some i, WLOG Q_1. But since Q_1 is irreducible, then P_1 and Q_1 must be *equal*, up to a scalar. So we can cancel them out, and repeat with the remaining factors (again, we can make this more precise with induction). □

§4.2 Principal Ideal Domains

The key point in the proof that $F[x]$ has unique factorization was that $F[x]/(P)$ is a field. We can consider more generally when this is true:

Definition 4.3. An integral domain is a *principal ideal domain* (PID) if every ideal is principal.

Definition 4.4. An integral domain is a *unique factorization domain* (UFD) if every element factors as a product of irreducibles in an essentially unique way.

We can adapt this argument to show that every PID is a UFD. But the converse is not true — for example, $\mathbb{Z}[x]$ and $\mathbb{C}[x_1, \ldots, x_n]$ (polynomials in several variables) have unique factorization as well, but are not a PID.

For the discussion of factorization, we'll assume R is an integral domain (if $ab = 0$, then a or b is 0), so that we can cancel.

Recall that an element $a \in R$ is *irreducible* if a is not a unit, and if $a = bc$ then either b or c is a unit. We say two irreducible elements a and b are *associate* if they differ by a unit, meaning that $a = bu$ for a unit $u \in R$.

So in the definition of a UFD, "essentially unique" factorization means unique up to reordering the factors and association. (With integers, we can fix this ambiguity by working with positive primes only, but in general this doesn't work.)

Theorem 4.5

Every PID is a UFD.

Proof. To prove uniqueness of factorization, we can use a very similar proof to the one used for polynomials. If R is a PID and $p \in R$ is irreducible, then (p) is a maximal ideal. So $R/(p)$ is a field, which means that if $p \mid ab$, then $p \mid a$ or $p \mid b$.

Once we have this property, now if we take two factorizations, we can take some irreducible in one factorization and find an element associate to it in the other, and cancel them out.

So this proves uniqueness. We haven't shown that a factorization always *exists*, and we won't prove it in the general case. But in all the examples we'll look at, it's clear why the factorization process must terminate — we have an invariant (the size of an integer, or the degree of a polynomial) that strictly decreases when we factor nontrivially. (This argument doesn't work in the general case.) □

§4.3 Euclidean Domains

We proved earlier that the polynomial ring $F[x]$ (for a field F) is a PID. We can use a similar argument to show that a somewhat more general class of rings are PIDs.

Definition 4.6. A *Euclidean domain* is a domain R together with a *size function* $\sigma : R/\{0\} \to \mathbb{Z}_{\geq 0}$, such that for every a and b in R, with $b \neq 0$, we can find q and r in R such that $a = bq + r$, and either $\sigma(r) < \sigma(b)$ or $r = 0$.

This essentially is division with remainder — it states that either a is divisible by b, or we can perform division with remainder and end up with a strictly smaller remainder.

Example 4.7

The ring $R = \mathbb{Z}$ is a Euclidean domain, with size function $\sigma(n) = |n|$.

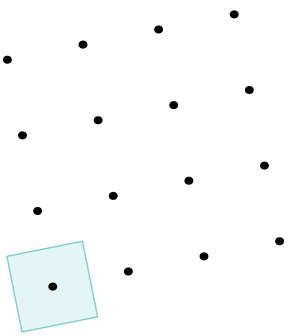
Example 4.8

The ring $R = F[x]$ (for a field F) is a Euclidean domain, with size function $\sigma(P) = \deg P$.

Example 4.9

The *Gaussian integers* $R = \mathbb{Z}[i]$ are a Euclidean domain, with size function $\sigma(a + bi) = a^2 + b^2$.

Proof. We can prove that the division with remainder property holds by geometry. Given b, the multiples of b form a square lattice (generated as a lattice by b and bi).



So by subtracting multiples of b, we can guarantee that a lands in the small square centered at the origin — more precisely, we can guarantee that $r = \alpha b + \beta i b$ where $-\frac{1}{2} \leq \alpha, \beta \leq \frac{1}{2}$. Then we have $\sigma(r) \leq \frac{1}{2}\sigma(b) < \sigma(b)$, as desired. □

Proposition 4.10

Any Euclidean domain is a PID, and therefore a UFD.

Proof. If $I \subset R$ is a nonzero ideal, then take an element $b \in I$ with minimal $\sigma(b)$. We know that for any $a \in I$, we can write $a = bq + r$, with $r = 0$ or $\sigma(r) < \sigma(b)$. The second case is impossible (since $r \in I$, but we chose b to have minimal size), so we must have $r = 0$. So b divides all elements of I, which means $I = (b)$. □

This is the exact same proof we gave for polynomials some time ago — the point of the definition is that the trick works a bit more generally. But it's not very general — in most cases, it doesn't work.

Example 4.11

The ring $R = \mathbb{Z}[\sqrt{-5}]$ is not a UFD, and therefore not a PID or Euclidean domain.

Proof. We have

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

It's possible to show that all of these elements — 2, 3, and $1 \pm \sqrt{-5}$ — are irreducible, so R does not have unique factorization.

Note that it *is* still possible to bound $\sigma(r)$ in terms of $\sigma(b)$ by the same geometric argument as before; but this bound will not be strong enough to imply $\sigma(r) < \sigma(b)$. □

§4.4 Factorization in Polynomial Rings

We've seen that PIDs have unique factorization, but the converse isn't true — there are cases where unique factorization is true, but there exist nonprincipal ideals.

For example, we'll see that $\mathbb{Z}[x]$ and $\mathbb{C}[x_1, \dots, x_n]$ are UFDs, but the ideals $(2, x) \subset \mathbb{Z}[x]$ and $(x, y) \subset \mathbb{C}[x, y]$ are not principal. In fact, the following theorem is true:

Theorem 4.12

If R is a UFD, then $R[x]$ is also a UFD.

Corollary 4.13

The rings $\mathbb{Z}[x]$ and $\mathbb{C}[x_1, \dots, x_n]$ are UFDs.

Proof of Corollary. For $\mathbb{Z}[x]$, this follows directly from the theorem (since $\mathbb{Z}$ is a PID). Meanwhile, for $\mathbb{C}[x_1, \dots, x_n]$, we can use induction: we have

$$\mathbb{C}[x_1, \dots, x_n] = \mathbb{C}[x_1, \dots, x_{n-1}][x_n]$$

(by thinking of n-variable polynomials as polynomials in the last variable x_n, whose coefficients are polynomials in the other $n - 1$ variables), so this follows immediately from the theorem as well. □

§4.4.1 Greatest Common Divisor

In order to prove Theorem 4.12, we'll first look at the concept of a gcd.

Definition 4.14. A gcd of two elements $a, b \in R$ is an element $d \in R$ such that $d \mid a$ or $d \mid b$, and for any other δ with this property, we have $\delta \mid d$.

It's clear that the gcd for integers has the same properties.

If $\gcd(a, b)$ exists, then it's unique up to association — if d and d' are both gcd's of a and b, then we must have $d \mid d'$ and $d' \mid d$, and in an integral domain, this means they must differ by a unit. (We can write $d = ud'$ and $d' = vd$, so then $uvd = d$. This implies $uv = 1$, since we're in an integral domain.) However, the gcd doesn't necessarily exist:

Example 4.15

In $\mathbb{Z}[\sqrt{-5}]$, there is no gcd of $2 + 2\sqrt{-5}$ and 6.

Proof. Both 2 and $1 + \sqrt{-5}$ are common divisors, and they are maximal in the sense that it's impossible to multiply them by non-units and still get a common divisor. This property for 2 implies that if the gcd existed, it would have to be 2, but $1 + \sqrt{-5}$ does not divide 2, contradiction. □

Lemma 4.16

In a UFD, the gcd always exists.

Proof. The usual way of calculating the gcd (over integers) using prime factorization still works — to find $\gcd(a, b)$, we can write down their prime factorizations, and take the smaller power of each irreducible. □

Remark 4.17. In a PID, if $\gcd(a, b) = d$, then we have $(a, b) = (d)$, which means d can be written in the form $ap + bq$. But this is not true in general — for example, in $\mathbb{C}[x, y]$, $\gcd(x, y) = 1$, but $1 \notin (x, y)$.

§4.4.2 Gauss's Lemma

Let R be a UFD. To prove unique factorization in $R[x]$, we'll use Gauss's Lemma.

Definition 4.18. A polynomial $P \in R[x]$ is *primitive* if the gcd of its coefficients is a unit in R.

Lemma 4.19 (Gauss's Lemma)

If P and Q are primitive, then so is PQ.

Proof. It's enough to show that for any irreducible $p \in R$, we can find a coefficient of PQ not divisible by p (as then by unique factorization, no element other than units can divide the gcd of its coefficients).

Let $P = \sum a_i x^i$ and $Q = \sum b_j x_j$, and let n be the maximal integer with $n \nmid a_n$ and m the maximal integer with $m \nmid b_m$. Then in PQ, the coefficient of x^{m+n} comes from $a_n b_m$, and other terms $a_i b_j$ where at least one of a_i and b_j is divisible by p; so this coefficient cannot be divisible by p. □

Remark 4.20. When $R = \mathbb{Z}$, we can think of this as reducing mod p — then P and Q are nonzero polynomials mod p, and since $\mathbb{F}_p$ is a field, this means PQ is nonzero mod p as well.

Using this, we can get a good sense of which polynomials are irreducible in R — as we'll see later, these are the irreducible elements of R, and primitive polynomials in $R[x]$ which are irreducible in $F[x]$, where $F = \mathrm{Frac}(R)$. So using unique factorization in R and in $F[x]$ can let us understand factorization in $R[x]$.

§4.4.3 Factorization of Integer Polynomials

We'll now prove Theorem 4.12 (that if R is a UFD, then $R[x]$ is a UFD as well) for the case of $R = \mathbb{Z}$. The proof in the general case is very similar (using the abstract construction of the gcd, instead of the more familiar gcd over integers).

Question 4.21. Given a polynomial $P \in \mathbb{Z}[x]$, there are two natural questions about its factorization:

1. Can P be factored in $\mathbb{Q}[x]$?
2. What about in $\mathbb{Z}[x]$?

The second question seems easier in some sense — we have more tools for answering it, which can be used to reduce potential factorizations to a finite number of possibilities. For example, we can reduce mod p, use the fact that the constant terms of the factors multiply to the constant term of P (which has finitely many factorizations), and so on.

Example 4.22

The polynomial $P(x) = 3x^3 + 2x + 2$ is irreducible in $\mathbb{Z}[x]$.

Proof. Assume P factors as P_1P_2 where P_1 and P_2 are nonconstant. Then reduce mod 2, so we have

$$P_1P_2 \equiv x^3 \pmod 2.$$

Since the leading terms of P_1 and P_2 are odd, the factorization in $\mathbb{F}_2$ cannot be $1 \cdot x^3$ (as then one polynomial would be constant), so it must be $x \cdot x^2$. But then the constant terms of P_1 and P_2 are both even, so the constant term of P would have to be 0 mod 4, contradiction. □

Meanwhile, trying to factor in $\mathbb{Q}[x]$ seems harder — we can no longer reduce to finitely many possibilities using the same methods (for example, the constant term of P can be factored in infinitely many ways as a product of *rationals*).

But it turns out that Gauss's Lemma implies that the two questions are actually essentially the same. First, it's clear that for any polynomial P, we can factor out the gcd of its coefficients, to write $P = nQ$ where Q is primitive. Meanwhile, Gauss's Lemma has the following corollary:

Corollary 4.23

If $P, Q \in \mathbb{Z}[x]$ such that $P \mid Q$ in $\mathbb{Q}[x]$ and P is primitive, then $P \mid Q$ in $\mathbb{Z}[x]$ as well.

Proof. We know that $Q = P \cdot S$ for some $S \in \mathbb{Q}[x]$. By clearing denominators and factoring out the gcd of the coefficients, we can then write $S = \frac{a}{b}T$, where $T \in \mathbb{Z}[x]$ is primitive. Then we have

$$bQ = aPT.$$

By Gauss's Lemma PT is primitive, so the gcd of the coefficients on the right-hand side is a. Meanwhile, the gcd of the coefficients on the left-hand side is clearly divisible by b. So $b \mid a$, which means $S \in \mathbb{Z}[x]$. □

Remark 4.24. Another way to view this proof is to define the *content* of a polynomial in $\mathbb{Z}[x]$ as the gcd of its coefficients, and extend the definition to $\mathbb{Q}[x]$ in the obvious way (such that the content of qP is q times the content of P for any $q \in \mathbb{Q}$). Then Gauss's Lemma implies that the content is multiplicative (the content of PQ is the content of P times the content of Q); in this case, if $Q = PS$, then since P has content 1 and Q has integer content, S must have integer content as well.

This implies that for example, $3x^3 + 2x + 2$ is irreducible in $\mathbb{Q}[x]$ as well. More generally:

Corollary 4.25

The irreducible elements of $\mathbb{Z}[x]$ are exactly $\pm p$ for integer primes p, and primitive polynomials which are irreducible in $\mathbb{Q}[x]$.

Proof. It's straightforward to check that all such elements are irreducible — first, it's clear that $\pm p$ are irreducible in $\mathbb{Z}[x]$ (since the factors would have to be integers). Meanwhile, if a polynomial is irreducible in $\mathbb{Q}[x]$, then the only way to factor it is to pull out constants, which is impossible for a primitive polynomial.

On the other hand, if $P \in \mathbb{Z}[x]$ is not of this form, then we can check that it factors:

- If P is an integer, then it factors as a product of integers.
- If P is not primitive, then we can factor out the gcd of its coefficients.
- If P is primitive but factors as $P = Q_1Q_2$ in $\mathbb{Q}[x]$, then we can rescale the factors so that $Q_1 \in \mathbb{Z}[x]$ is primitive; then by Corollary 4.23, we must have $Q_2 \in \mathbb{Z}[x]$ as well. □

Corollary 4.26

$\mathbb{Z}[x]$ is a UFD.

Proof. The existence of factorization is clear — start by writing $P = p_1 \cdots p_d Q$, where the p_i are integer primes and Q is primitive. Then if Q isn't irreducible in $\mathbb{Q}[x]$, we can factor it as a product of primitive polynomials in $\mathbb{Z}[x]$. We continue doing this until all factors are irreducible; this process must terminate since the degrees of the polynomials decrease at each step.

Meanwhile, to prove uniqueness, it suffices to show that if $P \in \mathbb{Z}[x]$ is irreducible, then if $P \mid Q_1Q_2$, we have $P \mid Q_1$ or $P \mid Q_2$. (Then we can finish by the same argument as in Lemma 4.2.)

If P is an integer prime p, then this follows from Gauss's Lemma. Otherwise, P is primitive and irreducible in $\mathbb{Q}[x]$. By unique factorization in $\mathbb{Q}[x]$, we know that if $P \mid Q_1Q_2$, then $P \mid Q_1$ or $P \mid Q_2$ in $\mathbb{Q}[x]$. But since P is primitive, by Corollary 4.23, the divisibility must hold in $\mathbb{Z}[x]$ as well. □

§4.5 Gaussian Integers

Unique factorization may be an abstract property, but unique factorization in $\mathbb{Z}[i]$ can be used to solve a concrete problem in number theory:

Question 4.27. Which integers n can be written in the form $a^2 + b^2$?

Example 4.28

We have $5 = 2^2 + 1^2$, while 6 and 21 cannot be written as the sum of two squares.

The question relates to Gaussian integers because

$$n = a^2 + b^2 \iff n = (a + bi)(a - bi).$$

In particular, this makes the following result clear:

Corollary 4.29

If n and m can be written as a sum of squares, so can nm.

Proof. If we have $n = \alpha\overline{\alpha}$ and $m = \beta\overline{\beta}$, then $nm = \alpha\beta\overline{\alpha\beta}$. □

This motivates considering the case where n is prime — it turns out that this will be enough to solve the general question.

Lemma 4.30

If p is an integer prime, then:

- p can be written as $a^2 + b^2$ iff p is not a prime in $\mathbb{Z}[i]$.
- This occurs iff $p = 2$ or $p \equiv 1 \pmod 4$.

Proof. The first statement is almost a direct application of unique factorization. On one hand, if p is prime in $\mathbb{Z}[i]$ and $p = (a + bi)(a - bi)$, then we must have $p \mid a + bi$ or $p \mid a - bi$. Either way, then $p \mid a$ and $p \mid b$, which is impossible.

On the other hand, if p is not a prime in $\mathbb{Z}[i]$, then we can write $p = \alpha\beta$ for some $\alpha, \beta \in \mathbb{Z}[i]$ (which are not units). It's clear that α and β cannot be integers (since p is prime), so they must be complex conjugates; then if $\alpha = a + bi$, we have $p = \alpha\overline{\alpha} = a^2 + b^2$.

Now we'll show that p is prime in $\mathbb{Z}[i]$ iff $p = 2$ or $p \equiv 1 \pmod 4$. When $p = 2$, we have $2 = (1 + i)(1 - i)$, so now assume p is an odd (integer) prime.

Claim — p is *not* prime in $\mathbb{Z}[i]$ iff there exists $\alpha \in \mathbb{Z}[i]$ such that $p \nmid \alpha$, but $p \mid \alpha\overline{\alpha}$.

Proof. One direction is clear — if $p \mid \alpha\overline{\alpha}$ but $p \nmid \alpha$, then by definition p cannot be prime.

For the other direction, by definition, if p is not prime, there exist $\alpha, \beta \in \mathbb{Z}[i]$ such that $p \nmid \alpha, \beta$, but $p \mid \alpha\beta$. Then we have

$$p \mid \alpha\beta\overline{\alpha\beta} = (\alpha\overline{\alpha})(\beta\overline{\beta}).$$

But both $\alpha\overline{\alpha}$ and $\beta\overline{\beta}$ are integers, so since p is an integer prime, p must divide one of them. □

This reduces the problem to one in $\mathbb{F}_p$, of determining when there exist nonzero $a, b \in \mathbb{F}_p$ such that $a^2 + b^2 = 0$. But since $\mathbb{F}_p$ is a field, we can divide by b^2, so this occurs iff -1 is a square in $\mathbb{F}_p$.

Now consider the multiplicative group $\mathbb{F}_p^\times$, which is an abelian group of order $p - 1$. We know its only element of order 2 is -1 (since $x^2 - 1 = (x - 1)(x + 1)$ only has the roots ± 1).

Consider the homomorphism $\varphi : \mathbb{F}_p^\times \to \mathbb{F}_p^\times$ where $\alpha \mapsto \alpha^2$. Then $\ker(\varphi) = \{\pm 1\}$, so by the homomorphism theorem, $|\mathrm{im}(\varphi)| = \frac{p-1}{2}$. A group has an element of order 2 iff its order is even (one direction follows from the fact that the order of an element always divides the order of the group, while the other follows from the Sylow Theorems), so since -1 is the only possible element of order 2, then -1 is in $\mathrm{im}(\varphi)$ (and is therefore a square) iff $\frac{p-1}{2}$ is even, or equivalently iff $p \equiv 1 \pmod 4$. □

This gives a classification of which integer primes remain prime in $\mathbb{Z}[i]$, and we can use this to work out a classification of *all* primes in $\mathbb{Z}[i]$:

Theorem 4.31

The full list of primes in $\mathbb{Z}[i]$, up to association, is:

- Integer primes $p \equiv 3 \pmod 4$;
- $a \pm bi$, where $a^2 + b^2 = p$ for an integer prime $p \equiv 1 \pmod 4$;
- $1 + i$.

Note that we have $2 = (1+i)(1-i)$, but $1+i$ and $1-i$ are associate, so 2 only contributes *one* prime (up to association).

Proof. First we'll check that all such elements are primes. We've already shown that if $p \equiv 3 \pmod 4$ is an integer prime, then p remains prime in $\mathbb{Z}[i]$. For the other two cases:

Claim — If $a^2 + b^2 = p$ is an integer prime, then $a + bi$ is prime in $\mathbb{Z}[i]$.

Proof. Define the norm $\mathrm{N}(a+bi) = a^2 + b^2$, which is multiplicative. Assume for contradiction that there exist α and β (which are not units) with $a + bi = \alpha\beta$, so we have

$$p = \mathrm{N}(a+bi) = \mathrm{N}(\alpha)\,\mathrm{N}(\beta).$$

Then since p is an integer prime, $\mathrm{N}(\alpha)$ or $\mathrm{N}(\beta)$ must be 1, and α or β must be a unit, contradiction. □

So then the second and third types of elements are all prime in $\mathbb{Z}[i]$ as well.

Now we will check that there are no other primes — it's enough to check that every non-unit $\alpha \in \mathbb{Z}[i]$ is divisible by some element of this list. To do so, we again use the norm — if α is not a unit, then we have $\alpha\overline{\alpha} = n$ for some integer $n > 1$. Let p be a prime divisor of n. Then if $p \equiv 3 \pmod 4$, we must have $p \mid \alpha$ (or $p \mid \overline{\alpha}$, which also implies $p \mid \alpha$), since p is prime. Otherwise, we can write $p = (a+bi)(a-bi)$, so $a+bi$ must divide α or $\overline{\alpha}$, and therefore $a+bi$ or $a-bi$ must divide α. □

This classification of primes gives us a full answer to our initial question:

Corollary 4.32

If n has prime factorization $p_1^{d_1} \cdots p_r^{d_r}$, then n can be written as a sum of squares iff all primes $p_i \equiv 3 \pmod 4$ have an even exponent d_i.

Proof. First, in order to show that all such n work, we know that all $p \not\equiv 1 \pmod 4$ can be written as a sum of squares, and all squares can trivially be written as a sum of squares. So any n of the form described is the product of numbers which can be written as a sum of squares, which means n itself is a sum of squares.

Conversely, if $n = (a+bi)(a-bi)$, then we want to check that every $p \equiv 3 \pmod 4$ appears an even number of times in the prime factorization of n (over integers). But p remains prime in $\mathbb{Z}[i]$, and if p has exponent d in the factorization of $a+bi$ (in $\mathbb{Z}[i]$), then it has exponent d in the factorization of $a-bi$ as well. Then p has exponent $2d$ in the prime factorization of n in $\mathbb{Z}$, which is even. □

§4.6 Fermat's Last Theorem

In 1847, a French mathematician announced a proof of Fermat's Last Theorem; this proof was later found to be wrong in general, but valid in a few cases. The initial steps of the argument are somewhat similar to the ideas we saw using factorization in $\mathbb{Z}[i]$.

Theorem 4.33 (Fermat's Last Theorem)

If $n > 2$, then there are no solutions in nonzero integers to

$$a^n + b^n = c^n.$$

Assume n is odd, and let ζ be a primitive nth root of unity. Then in the ring of *cyclotomic integers* $\mathbb{Z}[\zeta]$, we can factor

$$a^n + b^n = (a+b)(a+\zeta b)(a+\zeta^2 b)\cdots(a+\zeta^{n-1}b).$$

In many cases, it's possible to check that the factors are pairwise coprime. If we were working over $\mathbb{Z}$ and an nth power factored as a product of coprime integers, then all factors would have to be nth powers as well (up to multiplication by ± 1). Here, we'd like to conclude similarly that all factors are nth powers in $\mathbb{Z}[i]$ (up to multiplication by units); this eventually leads to a contradiction.

But in order to make this conclusion in $\mathbb{Z}$, we used the fact that $\mathbb{Z}$ is a UFD. However, $\mathbb{Z}[\zeta]$ is usually *not* a UFD — in fact, there's finitely many ζ for which it is a UFD. So this proof does not work. But Kummer later discovered that it's possible to modify it so that it works, under a weaker condition — that p is a *regular prime*. We will discuss this more later.

§5 Factorization in Number Fields

We will now discuss factorization in rings similar to $\mathbb{Z}[i]$ and $\mathbb{Z}[\zeta]$. We can think of $\mathbb{Z}[i]$ as a subring of $\mathbb{Q}[i]$, and similarly of $\mathbb{Z}[\zeta]$ as a subring of $\mathbb{Q}[\zeta]$ — the advantage of this perspective is that $\mathbb{Q}[i]$ and $\mathbb{Q}[\zeta]$ are fields.

Definition 5.1. A *number field* is a subfield of $\mathbb{C}$ which is finite-dimensional as a $\mathbb{Q}$-vector space.

§5.1 Algebraic Numbers and Integers

In the case of $\mathbb{Q}[i]$, we performed factorization in the ring $\mathbb{Z}[i]$, not the field $\mathbb{Q}[i]$. To describe the analog of $\mathbb{Z}[i]$ in a general number field, we need the concept of an *algebraic integer*.

Definition 5.2. A number $\alpha \in \mathbb{C}$ is *algebraic* if it is the root of a polynomial $P \in \mathbb{Q}[x]$.

If F is a number field with $\dim_{\mathbb{Q}}(F) = n$, and $\alpha \in F$, then we can write down the powers 1, α, …, α^n. For linear algebra reasons, these elements must be linearly dependent. This linear dependence gives a polynomial with rational coefficients that α is a root of, which means α must be algebraic.

Conversely, if $\alpha_1, \ldots, \alpha_n$ are algebraic, then $\mathbb{Q}[\alpha_1, \ldots, \alpha_n]$ is always a number field (since if we have a polynomial $P(\alpha) = 0$ of degree d, then we can express any α^k with $k \geq d$ in terms of lower powers; so there's finitely many terms involving the α_i that we can have).

If α is algebraic, then the set of polynomials $P \in \mathbb{Q}[x]$ for which $P(\alpha) = 0$ forms an ideal in $\mathbb{Q}[x]$. Since $\mathbb{Q}[x]$ is a PID, this means the ideal is generated by any of its minimal-degree polynomials.

Definition 5.3. If α is algebraic, the *minimal polynomial* of α is the monic polynomial that generates the ideal

$$\{P \mid P(\alpha) = 0\} \subset \mathbb{Q}[x].$$

Definition 5.4. α is an *algebraic integer* if its minimal polynomial lies in $\mathbb{Z}[x]$.

Lemma 5.5

α is an algebraic integer iff α is the root of *any* monic polynomial $P \in \mathbb{Z}[x]$.

Proof. First, clearly if α is an algebraic integer, then it's a root of its minimal polynomial, which is in $\mathbb{Z}[x]$.

For the other direction, we can use Gauss's Lemma — let P be a monic polynomial with $P(\alpha) = 0$, and let P_m be the minimal polynomial of α. Then we can rescale the minimal polynomial to a polynomial $Q = \frac{a}{b} P_m$ which is in $\mathbb{Z}[x]$ and primitive. We know $Q \mid P$ in $\mathbb{Z}[x]$, and Q is primitive; so by Gauss's Lemma, $Q \mid P$ in $\mathbb{Z}[x]$ as well.

But this requires that the leading coefficient of Q divides the leading coefficient of P. This means Q must have leading coefficient ± 1. So then $Q = \pm P_m$, and $P_m \in \mathbb{Z}[x]$ as well. □

Corollary 5.6

If $\alpha \in \mathbb{Q}$, then α is an algebraic integer iff $\alpha \in \mathbb{Z}$.

Proof. This is immediate, as the minimal polynomial of α is $x - \alpha$. □

In this class, the number fields we will work with are mostly of the form $\mathbb{Q}[\sqrt{d}]$, called *quadratic number fields*. We can assume without loss of generality that d is a squarefree integer. (Note that d may be positive or negative.)

Example 5.7

If $d \in \mathbb{Z}$ is squarefree, the algebraic integers in the quadratic number field $\mathbb{Q}[\sqrt{d}]$ are:

- $a + b\sqrt{d}$ where $a, b \in \mathbb{Z}$, if $d \not\equiv 1 \pmod 4$;
- $a + b\sqrt{d}$ where either $a, b \in \mathbb{Z}$, or $a + \frac{1}{2}, b + \frac{1}{2} \in \mathbb{Z}$, if $d \equiv 1 \pmod 4$.

Proof. For any $\alpha \in \mathbb{Q}[\sqrt{d}]$, we can write $\alpha = a + b\sqrt{d}$ where $a, b \in \mathbb{Q}$. Then its minimal polynomial is

$$P(x) = (x - a - b\sqrt{d})(x - a + b\sqrt{d}) = x^2 - 2a + (a^2 - b^2 d).$$

(This is because α is a root of P, while if it were a root of a polynomial of lower degree, it would have to be rational.) This means α is an algebraic integer iff $2a \in \mathbb{Z}$ and $a^2 - b^2 d \in \mathbb{Z}$.

Case 1 ($a \in \mathbb{Z}$). Then we must have $b^2 d \in \mathbb{Z}$. If b were not an integer, then its denominator would be divisible by p for some prime p, so the denominator of b^2 would be divisible by p^2. Since d is squarefree, then $b^2 d$ would have denominator divisible by p, contradiction. So b must be an integer, and all such cases clearly work.

Case 2 ($a \in \mathbb{Z} + \frac{1}{2}$). Then $2b$ must be an integer for the same reason as in the previous case, so we must have $(a, b) = (k + \frac{1}{2}, m + \frac{1}{2})$. Then we have

$$a^2 - b^2 d = \frac{(2k+1)^2 - (2m+1)^2 d}{4},$$

which is an integer iff $d \equiv 1 \pmod 4$. □

Example 5.8

The rings of algebraic integers in $\mathbb{Q}[i]$ and $\mathbb{Q}[\sqrt{-5}]$ are $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-5}]$, respectively, while the ring of algebraic integers in $\mathbb{Q}[\sqrt{-3}]$ is $\mathbb{Z}[\omega]$ (where ω is a primitive 3rd root of unity).

In the case of $\mathbb{Q}[i]$, we saw that $\mathbb{Z}[i]$ was a ring.

Theorem 5.9

For a number field F, the set of algebraic integers in F is a subring of F. Furthermore, it is the largest subring that is finitely generated as an abelian group under addition.

We won't prove this in general — but we'll see later that the algebraic integers are closed under addition and multiplication, which will show that they do form a subring of F. But in our specific example of $\mathbb{Q}[\sqrt{d}]$, this fact is clear.

§5.2 Ideal Factorization

We saw that the existence of unique factorization in $\mathbb{Z}[i]$ has concrete uses — we used it to solve the problem of which integers can be written as a sum of squares. We'd *like* this trick to work more generally, but unfortunately it doesn't — we've seen that $\mathbb{Z}[\sqrt{-5}]$ does not have unique factorization.

But we can instead prove a weaker statement. Instead of proving unique factorization as a product of prime *elements*, we'll prove unique factorization as a product of prime *ideals*.

§5.2.1 Motivation

To motivate the concept of factorization into ideals — we've seen that unique factorization into *elements* fails in general, so instead of trying to factor as a product of elements, we can try to factor as a product of *ideal divisors*.

We've named the concept, but we don't yet know what they are — but we know some ways we'd like them to behave. If we're using these ideal divisors for factorization, then they should appear in the factorization of *actual* ring elements; and in particular, they should arise as $\gcd(a_1, \ldots, a_n)$ for some elements $a_1, \ldots, a_n$ of the ring. (The gcd is not defined in a general ring; but if we've restored unique factorization into ideal divisors, then the gcd does again make sense.)

So we can introduce the formal gcd of elements in the ring. But in good situations where we *already* have unique factorization, such as $\mathbb{Z}$ and $\mathbb{Z}[i]$, we know that $\gcd(a_1, \ldots, a_n)$ is the generator of the ideal $(a_1, \ldots, a_n)$. So we can declare our formal gcd to actually *be* that ideal — we'll define $\gcd(a_1, \ldots, a_n)$ to be the ideal $(a_1, \ldots, a_n)$.

Remark 5.10. In fact, this is where the term *ideal* comes from. We introduced it by thinking about the kernel of a ring homomorphism, but it originally arose from thinking about these ideal divisors.

This is where the shorthand notation $(a_1, \dots, a_n)$ for $\gcd(a_1, \dots, a_n)$ comes from as well.

§5.2.2 Prime Ideals

To factor into ideals, we need a concept of *prime* ideals. The definition is very similar to the case of prime *elements*:

Definition 5.11. An ideal $I \subset R$ is *prime* if $I \neq R$, and for all $a, b \in R$, if $ab = I$ then $a \in I$ or $b \in I$.

Note that if $I = (a)$ is principal, then I is prime if and only if a is prime (this follows directly from the definition).

Proposition 5.12

An ideal $I \subset R$ is prime if and only if R/I is an integral domain.

Proof. This is clear from the definition as well — if we use $\overline{a}$ to denote a mod I (so $\overline{a} = 0$ if and only if $a \in I$), then the definition can be restated as that $\overline{ab} = 0$ implies $\overline{a}$ or $\overline{b}$ is 0. But this is exactly the definition of an integral domain (that there are no zero divisors). □

Corollary 5.13

All maximal ideals are prime.

Proof. An ideal is maximal if and only if R/I is a field; but all fields are integral domains, so all maximal ideals are prime as well. □

§5.2.3 Ideal Multiplication

Definition 5.14. Given two ideals $I, J \subset R$, their product is

$$IJ = \left\{ \sum a_i b_i \mid a_i \in I, b_i \in J \right\}.$$

It's clear that ideal multiplication is both commutative and associative.

Note that if I is generated by $a_1, \dots, a_m$ and J is generated by $b_1, \dots, b_n$, then IJ is generated by their pairwise products $a_i b_j$ (for $1 \leq i \leq m$ and $1 \leq j \leq n$). In particular, if $I = (a)$ and $J = (b)$, then $IJ = (ab)$ — so ideal multiplication is compatible with the usual multiplication of ring elements.

Proposition 5.15

We have $IJ \subset I \cap J$.

Proof. All terms in IJ are linear combinations of elements of I with coefficients in J, and therefore in R; so since I is an ideal, they must all be in I. Similarly, all terms in IJ must also be in J. □

On the other hand, it's not necessarily true that $IJ = I \cap J$.

Example 5.16

In the ring $\mathbb{Z}$, if $I = (m)$ and $J = (n)$, then $IJ = (mn)$, while $I \cap J = (\text{lcm}(m, n))$. In particular, $IJ = I \cap J$ if and only if m and n are relatively prime.

Now that we've defined the components of ideal factorization, we're ready to state our main theorem:

Theorem 5.17

In the ring of algebraic integers in a number field, every nonzero ideal factors uniquely (up to permutation of factors) as a product of prime ideals.

Unlike in the case of factorization into elements, we don't have to worry about association — ideals already capture information up to association (multiplication by a unit). For example, α and β are associate iff $(\alpha) = (\beta)$.

This theorem holds for *any* number field, but we'll only prove it in the case of imaginary quadratic fields.

§5.2.4 Ideals and Lattices

From now, we'll let $F = \mathbb{Q}[\sqrt{d}]$, where d is negative and squarefree, and we'll let R be the ring of algebraic integers in F. As we've seen previously, in Example 5.7, we have

$$R = \begin{cases} \mathbb{Z}[\sqrt{d}] & \text{if } d \not\equiv 1 \pmod 4 \\ \mathbb{Z}[\frac{1+\sqrt{d}}{2}] & \text{if } d \equiv 1 \pmod 4. \end{cases}$$

In either case, we can think of R as a *lattice* in $\mathbb{C}$ (an additive subgroup of $\mathbb{C}$ generated by two noncollinear vectors) — for example, $\mathbb{Z}[i]$ is a square lattice.

There are a few elementary properties of lattices that we'll make use of:

Proposition 5.18

If L and L' are lattices with $L' \subset L$, then:

1. L/L' is finite.
2. If L'' is a subgroup of L, and $L' \subseteq L'' \subseteq L$, then L'' must also be a lattice.

Just using these properties, we can quickly deduce a few important observations about ideals in R.

Corollary 5.19

Every nonzero ideal in R is a lattice.

Proof. First, this is clearly true for a *principal* ideal — if $I = (\alpha)$, then I can be obtained by multiplying the elements of R by α. But multiplying a lattice by a complex number produces another lattice similar to the original one (since we just scale and rotate it), so I is also a lattice.

But now if I is an arbitrary nonzero ideal, we can pick some nonzero $\alpha \in I$; then $\alpha R \subset I \subset R$. So by the above proposition, since αR and R are both lattices, I must be a lattice as well. □

Lemma 5.20

A nonzero ideal in R is prime if and only if it is maximal.

Proof. We've already shown that all maximal ideals are prime, so it suffices to show all prime ideals are maximal.

Let I be a prime ideal. Then by Proposition 5.12, R/I is an integral domain. But since R and I are both lattices, R/I must be finite.

Meanwhile, in any finite ring S, every element a which is not a zero divisor must be invertible — consider the list of elements ab, over all $b \in S$. All such elements must be distinct — if $ab = ac$, then we would have $a(b-c) = 0$, and therefore $b = c$. But there are exactly $|S|$ such elements, and only $|S|$ *possible* elements — so then the elements ab must cover every element of S, and in particular, there must exist b with $ab = 1$.

In our situation, since R/I is finite and there are *no* zero divisors, then *every* nonzero element must be invertible. So R/I is a field, which means I is maximal. □

§5.2.5 Proof of Unique Factorization

To prove uniqueness of ideal factorization in R, the key point is the following.

Proposition 5.21

Ideal multiplication has the following two properties:

1. The *Cancellation Property* — if we have ideals such that $IJ = I'J$ and $J \neq 0$, then we must have $I = I'$.
2. Divisibility is the same as inclusion — if $I \subset J$, then there exists an ideal J' such that $I = JJ'$.

Note that in *any* ring, the converse of the second point is true — as we've seen before, if $I = JJ'$ for some ideal J', then $I \subset J$. However, the direction written here is true for rings of algebraic integers as well as some other examples, but it isn't true in general.

To prove this, we'll use the following lemma.

Lemma 5.22

For any ideal $I \subset R$, the ideal $I\overline{I}$ is principal, and is generated by an integer.

Here $\overline{I}$ is the ideal formed by taking the complex conjugate of all elements in I.

Proof. We've seen that I is a lattice, so we can pick generators α and β which generate I as a lattice. Then they must also generate I as an *ideal*, so $I = (\alpha, \beta)$ and $\overline{I} = (\overline{\alpha}, \overline{\beta})$. This means

$$I\overline{I} = (\alpha\overline{\alpha}, \beta\overline{\beta}, \alpha\overline{\beta}, \beta\overline{\alpha}).$$

Note that $\alpha\overline{\alpha}$, $\beta\overline{\beta}$, and $\alpha\overline{\beta} + \beta\overline{\alpha}$ are all integers. So we can define

$$n = \gcd(\alpha\overline{\alpha}, \beta\overline{\beta}, \alpha\overline{\beta} + \beta\overline{\alpha})$$

(where the gcd is taken in $\mathbb{Z}$).

We then claim that $I\overline{I} = (n)$. It's clear that $I\overline{I}$ contains n, so it suffices to show that (n) contains all generators of $I\overline{I}$. By definition, we know (n) contains $\alpha\overline{\alpha}$, $\beta\overline{\beta}$, and $\alpha\overline{\beta} + \beta\overline{\alpha}$, so it's enough to check that (n) contains $\alpha\overline{\beta}$, or equivalently, that $\frac{\alpha\overline{\beta}}{n}$ is in R.

But to see this, we can consider its minimal polynomial

$$P(x) = \left(x - \frac{\alpha\overline{\beta}}{n}\right)\left(x - \frac{\beta\overline{\alpha}}{n}\right) = x^2 - \frac{\alpha\overline{\beta} + \beta\overline{\alpha}}{n} + \frac{\alpha\overline{\alpha} \cdot \beta\overline{\beta}}{n^2}.$$

By the definition of n, each of these coefficients is an integer; so then $\frac{\alpha\overline{\beta}}{n}$ is an algebraic integer and is therefore in R, as desired. □

Remark 5.23. In the case of a general number field, it's possible to perform a similar argument, but multiplying by all Galois conjugates instead. (We'll later learn what this means.)

This lemma gives rise to the following definition, which will be quite useful later.

Definition 5.24. The *norm* of an ideal I, denoted $\mathrm{N}(I)$, is the positive integer n such that $I\overline{I} = (n)$.

Note that if $I = (\alpha)$, then $I\overline{I} = (\alpha\overline{\alpha})$, so $\mathrm{N}(I) = \alpha\overline{\alpha}$. So in this case, the norm of the ideal coincides with the usual norm of a complex number.

It's also clear from the definition that the norm is multiplicative, since

$$IJ \cdot \overline{IJ} = I\overline{I} \cdot J\overline{J} \implies \mathrm{N}(IJ) = \mathrm{N}(I)\,\mathrm{N}(J).$$

Now we can use this lemma to prove our proposition.

Proof of Proposition 5.21. The main idea is that if J is principal, the proposition is almost obvious; and we can use the above lemma to reduce to the case where J is principal.

First, suppose $J = (\alpha)$ for some nonzero $\alpha \in R$. Then for the first point, it's clear that if $\alpha I = \alpha I'$, we must have $I = I'$. Meanwhile, for the second, if $I \subset (\alpha)$, then we must have $\frac{x}{\alpha} \in R$ for all $x \in I$. So we can take $J' = \frac{I}{\alpha} = \{\frac{x}{\alpha} \mid x \in I\}$ (which is an ideal of R), and we have $(\alpha)J' = \alpha \cdot \frac{I}{\alpha} = I$, as desired.

Now consider the general case. Then to prove the first point, we have

$$IJ = I'J \implies IJ\overline{J} = I'J\overline{J}.$$

But since $J\overline{J}$ is principal, the conclusion immediately follows from the special case where J is principal, proved above.

Similarly, for the second point, if $I \subset J$, then $I\overline{J} \subset J\overline{J}$. Let $J\overline{J} = (n)$; then as before, we can set $J' = \frac{I\overline{J}}{n}$, which must be an ideal of R. Then we have

$$J'J\overline{J} = J'(n) = I\overline{J},$$

and using the Cancellation Property, we can cancel out $\overline{J}$ to get $J'J = I$. □

Finally, we'll need one more lemma.

Lemma 5.25

Every non-unit ideal $I \subset R$ is contained in a maximal ideal.

Proof. Assume that $I \neq 0$. Then R/I is finite, so there are finitely many ideals in R/I, and therefore one must be maximal. But ideals of R containing I are in bijection with ideals of R/I, so there must be a maximal ideal of R containing I as well. □

Remark 5.26. In fact, this lemma is true in *any* ring. Of course, the proof we used here doesn't work in general, since we used the fact that R/I is finite, so the poset of its non-unit ideals (ordered by inclusion) necessarily has a maximal element (one with no element larger than it). In general, an *arbitrary* poset which is infinite doesn't need to have a maximal element (for example, just take the set of positive integers, ordered by size).

But in this case, the poset of ideals has an *additional* property — every increasing chain of ideals $I_1 \subset I_2 \subset \cdots$ is majorated by the union $I_1 \cup I_2 \cup \cdots$ (which is an ideal as well). Then using Zorn's Lemma from set theory, it's possible to show that there *is* a maximal ideal in R/I.

In a future class, we'll see a condition on R which is much more general than R/I being finite for all I (in particular, it'll hold for polynomial rings such as $\mathbb{Z}[x]$ and $\mathbb{C}[x_1, \ldots, x_n]$), which can be used to prove this lemma without relying on Zorn's Lemma.

Now we can put these together to prove unique factorization.

Proof of Theorem 5.17. There are two claims we need to prove — that a factorization into prime ideals *exists*, and is *unique*.

We'll prove existence first. Let $I \subset R$ be a nonzero ideal, which is not the unit ideal. Then by Lemma 5.25, there is a maximal (and therefore prime) ideal $\mathfrak{m}$ with $I \subset \mathfrak{m}$. By Proposition 5.21, this means we can factor $I = \mathfrak{m} \cdot J$ for some ideal J. Then if J is the unit ideal, we're done; otherwise, we can factor J in the same way, and so on.

It suffices to show that this process terminates; to do so, we can look at the norm. We have

$$\mathrm{N}(I) = \mathrm{N}(\mathfrak{m})\,\mathrm{N}(J),$$

and we must have $\mathrm{N}(\mathfrak{m}) > 1$ (as $1 \notin \mathfrak{m}$). So the norm decreases at each step, which means the process *must* eventually terminate.

Now we'll prove uniqueness. Suppose an ideal I factors in two ways, as

$$I = \mathfrak{p}_1 \cdots \mathfrak{p}_m = \mathfrak{q}_1 \cdots \mathfrak{q}_n.$$

It's enough to show that we must have $\mathfrak{q}_i = \mathfrak{p}_1$ for some i, since then we can cancel out this common ideal using Proposition 5.21, and then repeatedly perform the same argument until we've matched up all factors.

Assume for contradiction that $\mathfrak{q}_i \neq \mathfrak{p}_1$ for all i. Then since all these ideals are prime and therefore maximal, $\mathfrak{q}_i$ cannot be *contained* in $\mathfrak{p}_1$ for any i. So for each i, we can pick an element $x_i \in \mathfrak{q}_i$ such that $x_i \notin \mathfrak{p}_i$.

But then consider the element $x_1 \cdots x_n$. By definition, $x_1 \cdots x_n$ must be in $\mathfrak{q}_1 \cdots \mathfrak{q}_n$. On the other hand, it cannot be in $\mathfrak{p}_1$, since none of the x_i are (and $\mathfrak{p}_1$ is prime). So $x_1 \cdots x_n$ cannot be in $\mathfrak{p}_1 \cdots \mathfrak{p}_m$ either, contradiction. □

Remark 5.27. It's possible to restructure the final step in the proof as the slightly more general claim that if $\mathfrak{p}$ is a prime ideal, then for any ideals I and J with $I \not\subset \mathfrak{p}$ and $J \not\subset \mathfrak{p}$, we must also have $IJ \not\subset \mathfrak{p}$. (This is quite similar to the *definition* of a prime ideal, but involving ideals instead of elements.)

This can be proved in the same way — take x and y, which are in I and J respectively but not in $\mathfrak{p}$. Then xy is in IJ but not $\mathfrak{p}$.

§5.3 List of Prime Ideals

We'll now analyze the structure of ideals in R more closely. First, similarly to the case of Gaussian integers, we can obtain a full list of prime ideals by attempting to factor the integer primes into ideals.

Theorem 5.28

For each integer prime q, there are three possibilities for its ideal factorization:

- (q) is a prime ideal.
- (q) factors as $\mathfrak{p}_1\mathfrak{p}_2$ where $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are distinct and conjugate to each other. Then $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are prime ideals.
- (q) facotrs as $\mathfrak{p}^2$; then $\mathfrak{p}$ is a prime ideal.

This gives a full list of the nonzero prime ideals of R.

In these three cases, q is called *inert*, *splitting*, and *ramified*, respectively. Ramified primes can be thought of as edge cases — in the Gaussian integers the only ramified prime was 2, and in general there's finitely many ramified primes (and they come from the prime factors of d).

The proof is very similar to the one used for the Gaussian integers; it follows directly from the following lemma.

Lemma 5.29

If $\mathfrak{p}$ is a prime ideal in R, then either $\mathfrak{p} = (q)$ for an integer prime q, or $\mathfrak{p}\overline{\mathfrak{p}} = (q)$ for an integer prime q.

Proof. Suppose $\mathfrak{p}\overline{\mathfrak{p}} = (n)$, and assume n is not prime. Then we can write $n = ab$ for integers $a, b > 1$, so

$$\mathfrak{p}\overline{\mathfrak{p}} = (a)(b).$$

Since $\mathfrak{p}$ and $\overline{\mathfrak{p}}$ are prime, by unique factorization we must then have $\mathfrak{p} = (a)$ and $\overline{\mathfrak{p}} = (b)$ (or the other way around). This means $\mathfrak{p} = \overline{\mathfrak{p}} = (a)$, and a must be an integer prime. □

In fact, it turns out that similarly to the case of Gaussian integers, we can describe exactly when a prime splits as a product of ideals.

Lemma 5.30

If q is an odd integer prime, then q remains prime (or in other words, (q) is a prime ideal) if and only if the equation

$$a^2 - db^2 = 0$$

has no solutions in $\mathbb{F}_q$ other than $(0, 0)$.

Proof. First, if (q) is prime and there exists a solution (a, b), then

$$a^2 - db^2 = (a - b\sqrt{d})(a + b\sqrt{d})$$

is divisible by q, so one of the factors $a \pm b\sqrt{d}$ must be divisible by q. This implies $q \mid a$ and $q \mid b$.

Conversely, suppose (q) is not prime; then there exist α and β such that $q \mid \alpha\beta$, but $q \nmid \alpha$ and $q \nmid \beta$. Then we have

$$q \mid \mathrm{N}(\alpha\beta) = \alpha\overline{\alpha} \cdot \beta\overline{\beta},$$

so since q is an integer prime, we must have $q \mid \alpha\overline{\alpha}$ or $\beta\overline{\beta}$. Without loss of generality, $q \mid \alpha\overline{\alpha}$; then if $\alpha = a + b\sqrt{d}$, we have $q \mid a^2 - db^2$. Since $q \nmid \alpha$, then q cannot divide both a and b; so this gives a nonzero solution to $a^2 - db^2 = 0$ in $\mathbb{F}_q$. □

Remark 5.31. The reason we required q to be odd in the theorem is that when $q = 2$, we must be more careful about the possibility that a and b are half-integers (when q is odd, this isn't a concern, since 2 is invertible in $\mathbb{F}_q$); but it is possible to obtain a description of the case $q = 2$ as well.

This gives a description of when a prime q is inert, splitting, or ramified:

- If $a^2 \equiv b^2 d$ has no nontrivial solutions mod q (equivalently, d is not a square mod q), then q is inert, and (q) is a prime ideal.
- If $a^2 = b^2 d$ has a nontrivial solution, then $(q) = \mathfrak{p}\overline{\mathfrak{p}}$. In most cases, $\mathfrak{p}$ and $\overline{\mathfrak{p}}$ are distinct, so q is splitting and both are prime ideals.
- In a few cases, we have $\mathfrak{p} = \overline{\mathfrak{p}}$. In fact, this occurs exactly when q is an odd prime divisor of d, and when $q = 2$ and $d \not\equiv 1 \pmod 4$. These q are ramified, and we get one prime ideal $\mathfrak{p}$.

§5.4 Ideal Classes

We can also consider what ideals "look like," which gives rise to the concept of similarity of ideals. In this section, we assume all ideals are nonzero.

Definition 5.32. Two ideals I and J of R are *similar* if there exists some $\lambda \in F$ such that $\lambda I = J$.

Note that two ideals are similar under this definition if and only if their corresponding lattices are similar in the geometric sense (meaning that one can be obtained from the other by scaling and rotating).

It's clear that similarity is an equivalence relation, so we will write $I \sim J$ to denote that I is similar to J. We can then think about the equivalence classes of ideals under similarity, which we call *ideal classes*. We'll use $[I]$ to denote the class of I.

Example 5.33

The ideal class $[(1)]$ is exactly the set of principal ideals.

Solution. The ideal class of (1) consists of ideals which can be written as $\lambda(1)$ for $\lambda \in F$. But we must then have $\lambda \in R$ (since $\lambda \cdot 1$ must be in the ideal, and therefore in R), so $\lambda(1) = (\lambda)$ is a principal ideal. □

In particular, if R is a PID, then all ideals are similar. In most cases, R is not a PID; then the ideal classes are, in some sense, a measure of the failure of R to be a PID.

It's clear that if $I \sim I'$, then $IJ \sim I'J$. So multiplying ideals gives an operation on the set of ideal classes, which is both commutative and associative. In fact, this makes the set of ideal classes an abelian group — $[(1)]$ is the unit, and since $I\overline{I}$ is principal for any ideal I, the inverse of $[I]$ is $[\overline{I}]$.

Definition 5.34. The *ideal class group*, denoted $\mathrm{Cl}(F)$, is the abelian group of ideal classes (whose operation is ideal multiplication).

Example 5.35

When $R = \mathbb{Z}[i]$ or $\mathbb{Z}[\omega]$ (where ω is a primitive 3rd root of unity, so this is the ring of algebraic integers in $\mathbb{Q}[\sqrt{-3}]$), then $\mathrm{Cl}(F)$ is trivial.

Example 5.36

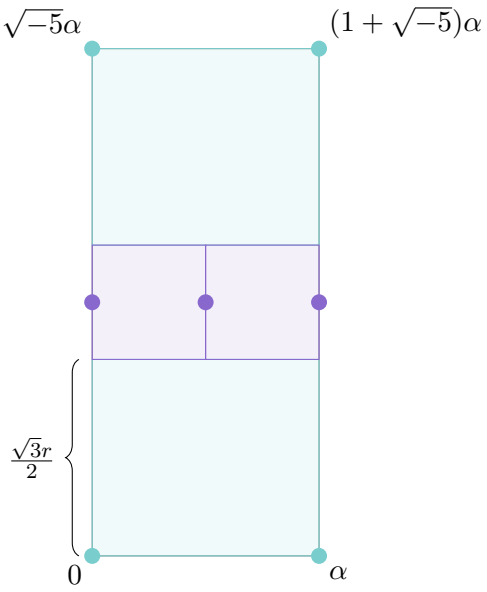
When $R = \mathbb{Z}[\sqrt{-5}]$, the ideal class group is $\mathbb{Z}/2\mathbb{Z}$ — the only two ideals up to similarity are (1) and $(2, 1+\sqrt{-5})$.

Proof. Let I be a nonzero ideal, and let α be an element of I with minimal nonzero norm. Let L be the lattice generated by α and $\sqrt{-5}\alpha$ (or equivalently, the lattice corresponding to the ideal (α)), so that $L \subset I$.

If $L = I$, then I is principal; so now assume $L \neq I$, so there is an element $\beta \in I$ with $\beta \notin L$. Without loss of generality, we can assume that β is in the rectangle spanned by α and $\sqrt{-5}\alpha$ (otherwise, we can subtract multiples of α and $\sqrt{-5}\alpha$ to translate β into this rectangle).

Claim — We must have $\beta = \alpha \cdot \frac{1+\sqrt{-5}}{2}$.

Proof. Let $r = |a|$. Then split the rectangle into smaller rectangles as shown:



If β is in a blue rectangle, then it has distance less than r from one of the blue dots; meanwhile, if β is in a purple rectangle, by straightforward computation we can check that β has distance less than $\frac{r}{2}$ from one of the purple dots. In either case, let this dot be γ.

In the first case, this is an immediate contradiction — we must have $\gamma \in I$, so then $\beta - \gamma \in I$, contradicting the assumption that α had minimal norm (and therefore minimal magnitude).

In the second case, we have $2\gamma \in I$, which means $2\beta - 2\gamma \in I$ as well. But $|2\beta - 2\gamma| < r = |\alpha|$, so this is again a contradiction unless $\beta = \gamma$.

But if β were either the leftmost or rightmost purple dot, then $\frac{\sqrt{-5}}{2}\alpha$ would be in I. Then, multiplying by $\sqrt{-5}$, we must have $\frac{5}{2}\alpha \in I$, and therefore $\frac{1}{2}\alpha \in I$; this is a contradiction (as α has minimal norm).

So β must equal the middle purple dot, which corresponds to $\frac{1+\sqrt{5}}{2}\alpha$. ■

So in this case, there is only one element of I inside this rectangle, which is $\beta = \alpha \cdot \frac{1+\sqrt{-5}}{2}$; then $I = (\alpha, \beta) = \frac{\alpha}{2}(2, 1+\sqrt{-5})$, as desired. □

§5.4.1 Finiteness of the Class Group

Our main theorem regarding the class group is the following.

Theorem 5.37

The class group $\mathrm{Cl}(F)$ is finite.

As before, this is true for *any* number field F, but we'll only prove it in the case of imaginary quadratic extensions.

We'll begin by proving a lemma. Let μ be $\sqrt{\frac{|d|}{3}}$ if $d \equiv 1 \pmod 4$, and $\sqrt{\frac{2|d|}{3}}$ if $d \not\equiv 1 \pmod 4$.

Lemma 5.38

Each ideal I contains a nonzero element α with

$$\alpha\overline{\alpha} \le \mu\,\mathrm{N}(I).$$

Proof. We'll use a few elementary geometric properties of lattices:

Fact — Let $L \subset \mathbb{R}^2$ be a lattice generated by v and w, and let Δ_L be the area of the parallelogram spanned by v and w. Then:

- Given L, the area Δ_L does not depend on the choice of basis vectors v and w.
- For any lattice $L' \subset L$, we have
$$\Delta_{L'} = [L : L']\Delta_L.$$

Recall that $[L : L']$ denotes the index of L' in L.

In our situation, let L be the lattice corresponding to I, and let $\Delta = \Delta_L$ for convenience.

Claim — If $v \in L$ has minimal length, then $|v|^2 \cdot \frac{\sqrt{3}}{2} \le \Delta$.

Proof. Let w be a vector of minimal length, of the vectors in L which are not multiples of v. Then $\{v, w\}$ must be a lattice basis for L (any point inside the parallelogram spanned by v and w must be at most $|w|$ away from one corner). So if the angle between v and w is θ, then

$$\Delta = |v| \cdot |w| \sin\theta.$$

Without loss of generality, we can assume $\theta \le \frac{\pi}{2}$ (otherwise, replace w with $-w$). Then we must have $\theta \ge \frac{\pi}{3}$ — we have $|w - v| \ge |w|, |v|$, so the angle opposite it must be the (weakly) largest among the three angles of the triangle. Then

$$\Delta \ge |v| \cdot |w| \cdot \frac{\sqrt{3}}{2} \ge \frac{\sqrt{3}}{2}|v|^2.$$ ■

Now it's enough to calculate Δ in terms of $\mathrm{N}(I)$. To do so, we'll relate $\mathrm{N}(I)$ to the *index* of I, so that we can use the second property of lattices.

Claim — If $\mathfrak{p}$ is a prime ideal, then $[J : \mathfrak{p}J] = \mathrm{N}(\mathfrak{p})$.

Proof. If $\mathfrak{p} = (q)$ for an integer prime q, then this is clear — the lattice of qJ is the lattice of J scaled by a factor of q in both directions, so $[J : qJ] = q^2$ (or in other words, $L/nL \cong (\mathbb{Z}/n\mathbb{Z})^2$ for any integer n).

So now assume that $\mathfrak{p}\overline{\mathfrak{p}} = (q)$ for an integer prime q. Then we have

$$[J : \mathfrak{p}J] \cdot [\mathfrak{p}J : \mathfrak{p}\overline{\mathfrak{p}}J] = [J : \mathfrak{p}\overline{\mathfrak{p}}J] = [J : qJ] = q^2.$$

But neither index can be 1, so both must be $q = \mathrm{N}(\mathfrak{p})$. ■

Claim — We have $\mathrm{N}(I) = [R : I]$.

Proof. Factor I into prime ideals, as $\mathfrak{p}_1\mathfrak{p}_2 \cdots \mathfrak{p}_k$. Then we have

$$[R : I] = [R : \mathfrak{p}_1] \cdot [\mathfrak{p}_1 : \mathfrak{p}_1\mathfrak{p}_2] \cdots [\mathfrak{p}_1 \cdots \mathfrak{p}_{k-1} : \mathfrak{p}_1 \cdots \mathfrak{p}_k].$$

Using the above claim and the multiplicativity of the norm, the right-hand side is

$$\mathrm{N}(\mathfrak{p}_1)\,\mathrm{N}(\mathfrak{p}_2) \cdots \mathrm{N}(\mathfrak{p}_k) = \mathrm{N}(I).$$ ■

Claim — We have that Δ is $\mathrm{N}(I)\sqrt{d}$ if $d \not\equiv 1 \pmod 4$, and $\frac{1}{2}\mathrm{N}(I)\sqrt{d}$ if $d \equiv 1 \pmod 4$.

Proof. Let Δ_R be the area corresponding to the lattice of R (the unit ideal); it's easy to check that Δ_R is $\sqrt{d}$ if $d \not\equiv 1 \pmod 4$ and $\frac{1}{2}\sqrt{d}$ if $d \equiv 1 \pmod 4$.

Then using the properties of lattices, we have

$$\Delta = [R : I]\Delta_R = \mathrm{N}(I)\Delta_R,$$

as desired. ■

Combining these claims, if we take α to be the element of I with minimal magnitude, then

$$\alpha\overline{\alpha} \cdot \frac{\sqrt{3}}{2} \le \begin{cases} \mathrm{N}(I)\sqrt{d} & \text{if } d \not\equiv 1 \pmod 4 \\ \frac{1}{2}\mathrm{N}(I)\sqrt{d} & \text{if } d \equiv 1 \pmod 4, \end{cases}$$

which means $\alpha\overline{\alpha} \le \mu\,\mathrm{N}(I)$. □

Using this lemma, we can prove the following proposition by a clever algebraic trick.

Proposition 5.39

Every ideal class has a representative I with $\mathrm{N}(I) \le \mu$.

Proof. Suppose we have a nonzero ideal I; we'll then find an ideal similar to it with norm at most μ.

First, take $\alpha \in \overline{I}$ with $|\alpha|^2 \leq \mu \,\mathrm{N}(I)$. Now we have $(\alpha) \subset \overline{I}$, so by Proposition 5.21 (since inclusion implies divisibility), we must have $(\alpha) = \overline{I} \cdot J$ for some ideal J. But then

$$1 = [(\alpha)] = [\overline{I}] \cdot [J] = [I]^{-1}[J],$$

which means $[J] = [I]$. Meanwhile, we have

$$\mathrm{N}(J) \cdot \mathrm{N}(I) = \mathrm{N}(\alpha) \leq \mu \,\mathrm{N}(I),$$

which means $\mathrm{N}(J) \leq \mu$. □

From this proposition, the finiteness of the class group is clear:

Proof of Theorem 5.37. By Proposition 5.39, every ideal class has some representative with bounded norm. But for any prime q, there's at most two ideals with norm q — so for any integer n, there's only finitely many ideals with norm n. So there's only finitely many ideal classes as well. □

In fact, Proposition 5.39 gives an effective way to *compute* the class group, as well — the ideal class group is generated by the classes of the *prime* ideals, so it's enough to find the prime ideals of norm at most μ, which can be done by analyzing the factorization of (q) for each prime $q \leq \mu$. (Proposition 5.39 doesn't guarantee that I *itself* is prime, but if it isn't, then we can factor it as a product of prime ideals with smaller norm.)

§5.5 Generalizations

So far, we've mostly worked with imaginary quadratic fields. Now we'll see how similar ideas may apply to a few other fields.

§5.5.1 Real Quadratic Fields

Now let $F = \mathbb{Q}[\sqrt{d}]$ where $d > 0$ (previously, we worked with the case $d < 0$). Define δ with $\delta^2 = d$, so that $F = \mathbb{Q}[\delta]$. In this case (called a *real* quadratic field), many of these concepts work the same way, but there are a few differences.

First, we can't use complex conjugation. However, we still have an operation of conjugation in F, where we send $a + b\delta \mapsto a - b\delta$.

Similarly, R is no longer a lattice in $\mathbb{C}$. But we can instead embed it as a lattice in $\mathbb{R}^2$ (instead of thinking of our plane as $\mathbb{C}$, we think of it as a ring with two coordinates and componentwise addition and multiplication) via the map $a + b\delta \mapsto (a + b\sqrt{d}, a - b\sqrt{d})$. (The reason to use this map is intuitively that we could have constructed F via the abstract construction $\mathbb{Q}[x]/(x^2 - d)$. In that sense, δ is an abstract square root of d, and can correspond to $\sqrt{d}$ or $-\sqrt{d}$.)

The norm of an element $\alpha = a + b\delta$ is now $\mathrm{N}(\alpha) = (a + b\delta)(a - b\delta) = a^2 - b^2 d$. This is still an integer, but it's not necessarily positive (and it's *not* equal to $|\alpha|$). In particular, unlike the case of an imaginary quadratic field (where there were very few units), there are infinitely many units (or in other words, there are infinitely many solutions to the *Pell equation* $a^2 - b^2 d = 1$).

In this framework, many of the arguments we used generalize to the case of real quadratic fields as well; in particular, the uniqueness of ideal factorization can be proved similarly. Our arguments can be generalized to rings of algebraic integers in *any* number field, but this requires significantly more work.

§5.5.2 Function Fields

Let K be a field. Suppose we replace $\mathbb{Q}$ with $K(t)$ (this notation denotes $\mathrm{Frac}(K[t])$, the field of rational functions in t). Then in this situation, F is a field containing $K(t)$, which is finite-dimensional over $K(t)$ (as a vector space); and R is the set of $\alpha \in F$ for which $P(\alpha) = 0$ for a monic polynomial $P \in K[t][x]$.

Two especially common examples are $K = \mathbb{C}$ and $K = \mathbb{F}_p$; here we'll focus on the case $K = \mathbb{C}$.

For concreteness, we'll again focus on the case where F is a quadratic extension, so

$$R = \mathbb{C}[t, z]/(z^2 - P(t))$$

where $P(t)$ is squarefree (and therefore does not have multiple roots).

Theorem 5.40

In R, every ideal can be uniquely factored as a product of prime ideals.

We won't prove this, but in the case of number fields, an important question we looked at was how (p) factors as a product of ideals in R (where p is an integer prime). We can ask the same question here:

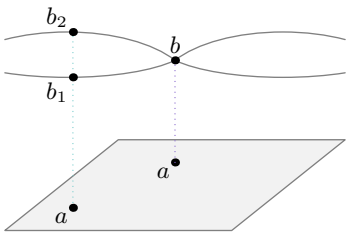
Question 5.41. How do the prime ideals of $\mathbb{C}[t]$ factor as a product of prime ideals in R?

First, the prime ideals of $\mathbb{C}[t]$ are exactly $(t - \lambda)$ for $\lambda \in \mathbb{C}$ — since $\mathbb{C}[t]$ is a PID, the prime ideals are exactly the ones generated by irreducible polynomials, and irreducible polynomials over $\mathbb{C}$ must be linear.

Meanwhile, it's still true that all nonzero prime ideals in R are maximal (it's again possible to use a finiteness argument, but using that the quotient R/I is a finite-dimensional vector space, instead of that it is finite). So then we can use Nullstelensatz to describe them — all maximal ideals are of the form

$$\mathfrak{m}_{a,b} = \ker(\mathrm{ev}_{a,b})$$

for some $(a, b) \in \mathbb{C}^2$ with $b^2 = P(a)$, where $\mathrm{ev}_{a,b}$ is the evaluation homomorphism $Q(t, z) \mapsto Q(a, b)$. In particular, each value $a \in \mathbb{C}$ corresponds to two values of b, namely $\pm\sqrt{P(a)}$ — except roots of P, which correspond to only one value of b. This gives a *ramified double cover*:



The analog of complex conjugation is $b \mapsto -b$.

Then $t - \lambda$ is in $\mathfrak{m}_{a,b}$ if and only if $\lambda = a$ (since $t - \lambda$ must evaluate to 0 when we plug in $(t, z) = (a, b)$). So $t - \lambda$ lies in two maximal ideals of R if λ is not a root of P, and one if λ is a root of P. In either case, it's possible to check that $t - \lambda$ is the product of these two ideals:

Proposition 5.42

Let b be a complex number with $b^2 = P(\lambda)$. Then

$$(t - \lambda) = \mathfrak{m}_{\lambda,b}\mathfrak{m}_{\lambda,-b}.$$

In the case where λ is a root of P (so the two prime ideals of R are the same), we again say that $t-\lambda$ is a *ramified* prime; and in the case where λ is not a root of P (so the two prime ideals are distinct), we say that $t-\lambda$ is a *splitting* prime. Note that there are no inert primes, because $\mathbb{C}$ is algebraically closed; but if we instead worked over $\mathbb{F}_p$, then we *would* have analogues of inert primes.

Example 5.43

If $P(t)=t$, then we have $R=\mathbb{C}[z,t]/(z^2-t)\cong\mathbb{C}[z]$. In this case, the only ramification point is 0; and factorization is the familiar equation

$$(t-\lambda)=(z-\sqrt{\lambda})(z+\sqrt{\lambda}).$$

To conclude, here is a table of the analogs between quadratic number fields and quadratic extensions of $\mathbb{C}(t)$. For convenience, we assume $d\not\equiv 1 \pmod 4$.

$\mathbb{Q}$	$\mathbb{C}(t)$
$\mathbb{Z}$	$\mathbb{C}[t]$
$\mathbb{Z}[\sqrt{d}]$	$\mathbb{C}[t,z]/(z^2-P(t))$
p	$t-\lambda$
Ramified primes	$t-\lambda$ where $P(\lambda)=0$
Splitting primes	$t-\lambda$ where $P(\lambda)\neq 0$
Inert primes	N/A

§6 Modules

We are trying to tell a story where rings are the protagonist, and for a protagonist to be interesting, it must act. This motivates the concept of modules.

Definition 6.1. Let R be a ring. A module M over R is an abelian group, together with an *action map* $R\times M\to M$, with $(r,m)\mapsto rm$, subject to the following axioms:

- $1m=m$ (where 1 is the multiplicative identity of R);
- $r_1(r_2m)=(r_1r_2)m$;
- Distributivity in both variables: $(r_1+r_2)m=r_1m+r_2m$, and $r(m_1+m_2)=rm_1+rm_2$.

The first two axioms are very similar to the definition of a group action on a set. So a ring to a module is like a group G to a G-set (a set with an action by G).

Example 6.2

If $R=F$ is a field, then a module is the same as a vector space.

Remark 6.3. The definition also applies to a noncommutative ring R, in the same way — this definition does not reference commutativity. We can have more examples of modules over non-commutative rings: for example, we can take $R=\mathrm{Mat}_n(F)$ and $M=F^n$ (since matrices act on column vectors by multiplication). As another example, if $R=\mathbb{C}[G]$ is the group ring, then a module is the same as a complex representation of G.

For any ring R, there is a uniquely defined homomorphism $\mathbb{Z}\to R$, where $1\mapsto 1_R$. A related fact is that every abelian group has a unique structure as a $\mathbb{Z}$-module: we know that 1 (in $\mathbb{Z}$) must map $m\to m$, so

then by distributivity, $n = 1 + 1 + \cdots + 1$ must map

$$v \mapsto \underbrace{v + v + \cdots + v}_{n}.$$

Similarly, $-n$ must map v to $-(v + v + \cdots + v)$. So a $\mathbb{Z}$-module is the same as an abelian group.

Example 6.4

Describe a module over $\mathbb{C}[x]$.

Solution. First, we get a $\mathbb{C}$-vector space V by looking at how the constants act. But then we need to see what x does. We know x must define a linear map $A : V \to V$, where $xv = Av$. Then every other element's action is defined: for a polynomial $P = a_n x^n + \cdots + a_0$, we get

$$Pv = a_0 v + a_1 Av + \cdots + a_n A^n v.$$

(The vector space may or may not be finite-dimensional.) □

Example 6.5

Describe a module over $\mathbb{Z}/n$.

Solution. The main point is that if R/I is a quotient of R, then every R/I-module is also a R-module, where we define $r(m)$ to be $\overline{r}(m)$ (where $\overline{r}$ denotes r mod I). Meanwhile, I must act by 0 in the R/I-module. So a R/I module is the same as a R-module where every element of I acts in a trivial way (meaning $rv = 0$ for all $r \in I$).

So a $\mathbb{Z}/n$-module is the same as an abelian group where the order of every element divides n — meaning $na = 0$ for all a in the group.

Then for every m, if $\overline{m}$ denotes m mod n, we can write

$$\overline{m}_v = \underbrace{v + v + \cdots + v}_{m}.$$

In order for this to be well-defined, the sum should not depend on our representative for the residue; this is guaranteed by the condition $na = 0$. □

For any ring R, there is a simple example of a module: the *free module* over R is $M = R$ itself, where the action is multiplication — we have $r(x) = rx$. This is parallel to the observation that a group G acts on itself by left multiplication.

§6.1 Submodules

Definition 6.6. Given a module M, a *submodule* $N \subset M$ is an abelian subgroup which is invariant under the R-action — meaning $rx \in N$ for all $x \in N$ and $r \in R$.

If $N \subset M$ is a submodule, we can define their quotient M/N, where we take the quotient as abelian groups. This carries a module structure as well, given by $r\overline{m} = \overline{rm}$ (where $\overline{m}$ denotes m mod N). This is well-defined because N is a submodule — we have $r(m + n) = rm + rn$, and $rn \in N$.

Then the Homomorphism Theorem and Correspondence Theorem work in the exact same way as in abelian groups.

Example 6.7

What are the submodules of the free module?

Solution. The answer is exactly the ideals of R — we're looking for abelian subgroups of R which are invariant under multiplication by all terms in R, and by definition these are ideals. □

Definition 6.8. Given two modules M and N, their *direct sum* is

$$M \oplus N = \{(m, n) \mid m \in M, n \in N\}$$

with the action

$$r(m, n) = (rm, rn).$$

Definition 6.9. The *free module* of rank n is

$$R^n = \underbrace{R \oplus R \oplus \cdots \oplus R}_{n}.$$

In the case where $R = F$ is a field, the free module of rank n is exactly F^n, the standard n-dimensional vector space.

§6.2 Homomorphisms

In linear algebra, we work with matrices. Matrices are also relevant here — they come up when we want to understand linear maps (the term in this case is different, but it's the same concept).

Definition 6.10. A *homomorphism* from a module M to a module N is a homomorphism of abelian groups $\varphi : M \to N$, which is compatible with the R-action — meaning $\varphi(rm) = r\varphi(m)$ for all $r \in R$ and $m \in M$.

In vector spaces, this is the same as a linear map.

We'll use $\mathrm{Hom}_R(M, N)$ to denote the set of all homomorphisms $M \to N$. Note that homomorphisms can be added and rescaled, in the same way as linear maps: $(\varphi_1 + \varphi_2)(m) = \varphi_1(m) + \varphi_2(m)$, and $(r\varphi)(m) = r\varphi(m)$. So then $\mathrm{Hom}_R(M, N)$ is a R-module.

It's easy to understand homomorphisms from the free module $\mathrm{Hom}_R(R, M)$. In order to give a homomorphism $R \to M$, we need to decide where 1_R is sent; suppose our homomorphism φ maps $1_R \mapsto m$, for some $m \in M$. But then this uniquely determines where every element is sent: for any $r \in R$, we have

$$\varphi(r) = \varphi(r \cdot 1_R) = r \cdot \varphi(1_R) = rm.$$

So a homomorphism is determined by $\varphi(1_R)$, and there are no restrictions on m — this is why R is called a free module. This means $\mathrm{Hom}_R(R, M)$ is isomorphic to M: the bijection is given by mapping $\varphi \in \mathrm{Hom}_R(R, M)$ to $\varphi(1)$, and $m \in M$ to the homomorphism $\varphi_m : r \mapsto rm$.

Similarly, $\mathrm{Hom}_R(R^n, M)$ is equally easy to understand. Now R^n is generated by the elements 1_i which have a 1 in their ith place, and 0's everywhere else. So $\mathrm{Hom}_R(R^n, M)$ is isomorphic to M^n, where the bijection sends $\varphi \in \mathrm{Hom}_R(R^n, M)$ to the element $(\varphi(1_1), \varphi(1_2), \ldots, \varphi(1_n))$, and $(m_1, \ldots, m_n) \in M$ to the homomorphism $\varphi(x_1, \ldots, x_n) = \sum x_i m_i$.

In particular, we have $\mathrm{Hom}_R(R^n, R^m) = (R^m)^n = \mathrm{Mat}_{m\times n}(R)$ — we can write homomorphisms in the way we're used to in linear algebra, where $A \in \mathrm{Mat}_{m\times n}(R)$ sends $(x_1, \dots, x_n)^t$ to $A(x_1, \dots, x_n)^t$. So as long as we work with free modules and homomorphisms, there's many parallels to linear algebra. (But in linear algebra, there's various characterizations of matrices that no longer hold here.)

Remark 6.11. The direct sum $M \oplus N$ is the same as $M \times N$. This is true for finite sums and products: we have

$$M_1 \oplus \cdots \oplus M_n = M_1 \times \cdots \times M_n.$$

But it's not true for infinite ones.

§6.3 Generators and Relations

Definition 6.12. A collection of elements $m_1, \dots, m_n \in M$ form a system of *generators* if every $x \in M$ can be expressed as $\sum r_i m_i$ for $r_i \in R$.

If such a finite set exists, we say M is *finitely generated*.

When R is a field, a set of generators is a system of vectors which span the space. In linear algebra, we could drop some of the vectors to get a basis, where the representation is unique; but that isn't true here in general.

§6.3.1 Presentation Matrices

Consider a system of generators $a_1, \dots, a_n$ of M. These define a homomorphism of modules $\varphi : R^n \to M$ (since every list of n elements defines a homomorphism), and since the a_i are generators, the homomorphism must be onto.

If the presentation as $x = \sum r_i a_I$ is unique, or equivalently $\ker \varphi = \{0\}$, then φ is an isomorphism and M is free. But this is usually not the case.

But we still have $M \cong R^n/\ker \varphi$ by the Homomorphism Theorem, where $\ker \varphi$ is another module. If $\ker \varphi$ is again finitely generated, then we say M is *finitely presented*; we'll need to assume that our modules are finitely presented for now, but we'll see later that for a large class of rings, every finitely generated module is finitely presented.

Now if M is finitely presented, we can choose a set of generators $b_1, \dots, b_m$ which generate $\ker \varphi$. We can then use this to completely describe M: each b_i is an element of R^n, so we can think of it as a column vector with n entries. Then we can write a $n \times m$ matrix

$$B = \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_m \\ | & | & & | \end{bmatrix}.$$

Now we have the onto map $\psi : R^m \to \ker \varphi$, which gives a map $\psi : R^m \to R^n$ with $\mathrm{im}\, \psi = \ker \varphi$. Similarly to in linear algebra, ψ is given by this matrix: we have $\psi(x) = Bx$ (for column vectors x).

Then $\ker \varphi = BR^m$ — another way to think of this is that the kernel is the span of the column vectors b_i. Since we saw that $M \cong R^n/\ker \varphi$, we then have $M \cong R^n/BR^m$. We call B the *presentation matrix* of M.

§6.3.2 Row and Column Operations

Note that the presentation of a given module is not at all unique:

Example 6.13

If $R = \mathbb{Z}$, the module $M = \mathbb{Z}/5\mathbb{Z}$ has the obvious presentation $\mathbb{Z}/\begin{bmatrix}5\end{bmatrix}\mathbb{Z}$, but it also has the presentation

$$\mathbb{Z}^2/\begin{bmatrix}2 & 1\\ 1 & 3\end{bmatrix}\mathbb{Z}^2.$$

Proof. Let B be the presentation matrix. We have $\det B = 5$. Then $B\mathbb{Z}^2$ is the lattice spanned by the column vectors $(2,1)^t$ and $(1,3)^t$. This lattice L has index $|\det B|$ as a subgroup of $\mathbb{Z}^2$ — we mentioned this earlier, when discussing quadratic number fields — so $\mathbb{Z}^2/L$ has five elements. The only abelian group with 5 elements is $\mathbb{Z}/5$, so then this presentation gives $\mathbb{Z}/5$. □

Our goal will be to generalize this example, and understand modules over $\mathbb{Z}$ more systematically from their presentations. The idea is to start with an arbitrary presentation matrix, and do some operations on the matrix which don't change the module. We can use these operations to reduce the matrix to a specific form, from where it's easy to understand the module.

We'll use $M = M_B$ to denote that M is the module with presentation matrix B.

Proposition 6.14

The elementary column operations on matrices over a ring R are the following:

1. Multiply a column by a *unit* in R.
2. Add an arbitrary multiple of one column to another.
3. Swap two columns.

These operations do not change the span of the columns.

This is true for the same reason as in a field. Then if B' is obtained from B by elementary column operations, we have $BR^m = B'R^m$.

Another useful way to think of elementary column operations is in terms of matrix multiplication. If B' is obtained from B by column operations, then $B' = BC$ where C is a $m \times n$ *invertible* matrix — this is because all the elementary column operations correspond to invertible matrices.

Remark 6.15. We still use $\mathrm{GL}_m(R)$ to denote the group of invertible matrices. Now $C \in \mathrm{GL}_m(R)$ means that $\det C$ is a unit in R — it's no longer enough to just require that it's nonzero. In fact, the converse is true as well.

Now we can write $B'R^m = BCR^m$, but since C is invertible, we have $CR^m = R^m$ (since the map defined by C is an isomorphism), which means $B'R^m = BR^m$. This is the same argument, but in the language of matrix multiplication instead.

We've discussed column operations, so we may wonder about row operations as well. We can define row operations in the exact same way.

Proposition 6.16

If B' is obtained from B by elementary *row* operations, then they are presentation matrices for *isomorphic* modules, meaning $M_B = M_{B'}$.

Proof. We can prove this by an argument exclusively in terms of matrix multiplication. If B' is obtained from B by row operations, then $B' = CB$ where $C \in \mathrm{GL}_n(R)$ is an invertible $n \times n$ matrix.

Now we want an isomorphism between R^n/BR^m and $R^n/B'R^m = R^n/CBR^m$. But that isomorphism is just given by multiplication by C:

$$\begin{array}{ccc} R^n/BR^m & & R^n/CBR^m \\ \uparrow & & \uparrow \\ R^n & \longrightarrow & R^n \end{array}$$

The map $x \mapsto Cx$ is an isomorphism from R^n to itself. But $x \in BR^m$ iff $Cx \in CBR^m$, so its restriction gives an isomorphism from BR^m to $B'R^m$, and therefore an isomorphism of quotients as well. □

§6.3.3 Smith Normal Form

In the case where R is a Euclidean domain, we can actually use these elementary row and column operations to reduce our matrix to a simple form, and get a classification of finitely presented modules.

Theorem 6.17

Evern $n \times m$ matrix over a Euclidean ring R can be reduced by elementary row and column operations to *Smith normal form*, where there are 0's everywhere except on the diagonal, and the diagonal entries satisfy $d_{11} \mid d_{22} \mid d_{33} \mid \cdots$.

Remark 6.18. In terms of matrix multiplication, the theorem states that for any B, we can write a matrix $D = ABC$ of this form, for $A \in \mathrm{GL}_n(R)$ and $C \in \mathrm{GL}_m(R)$. THen $M_D \cong M_B$. This version of the statement is also true when R is a PID, but it's possible that D is not the result of elementary row and column operations.

We can use this to concretely understand finitely presented modules:

Corollary 6.19

A finitely presented module over R is a direct sum of cyclic modules: we have

$$M \cong R^n \oplus R/d_1 \oplus R/d_2 \oplus \cdots \oplus R/d_n,$$

where $d_1 \mid d_2 \mid \cdots \mid d_n$.

This is clear from the theorem: it's easy to compute the image of a diagonal matrix of this form. We have to assume the module is finitely *presented* for now, but we'll later see that in this case, all finitely generated modules are finitely presented.

A *cyclic module* is a module generated by one element, which must be of the form R/I.

Before we prove the theorem, we'll look at a simple example.

Example 6.20

Describe the modules corresponding to the 2×2 presentation matrices

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}.$$

Solution. The easy case is when the matrix is diagonal. In the first case, we have $M_B \cong \mathbb{Z}/2 \times \mathbb{Z}/3$, since in order to find the coset of a vector $(x, y)^t$, we only care about x mod 2 and y mod 3. Similarly, in the second case $M_B \cong \mathbb{Z}/5$ — we can kill the second coordinate, since everything is divisible by 1. (In other words, the second coordinate gives the relation $x = 0$ on the second generator, which means it's useless.)

In the third case we have the two column vectors $(2, 1)^t$ and $(1, 3)^t$. Since $\det B = 5$, the matrix is degenerate mod 5, and we can make a column divisible by 5: add twice the first column to the second, to get

$$B' = B \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 5 \\ 1 & 5 \end{bmatrix}.$$

Now we keep $(2, 1)^t$, but replace $(1, 3)^t$ with $(5, 5)^t$.

Notice that $(2, 1)^t$ and $(1, 1)^t$ form a basis in $\mathbb{Z}^2$, since their determinant is 1. So in order to get the kernel, we took a basis, kept one of the vectors, and multiplied the other by 5. So this is actually isomorphic to our second example — by changing the basis of the lattice, we can get that the kernel is spanned by $(1, 0)^t$ and $5(0, 1)^t$. So we see $M_B \cong \mathbb{Z}/5$ again. □

Now we'll prove the theorem. To motivate the proof, note that the gcd of all matrix entries doesn't change under elementary row and column operations, since $\gcd(x, y) = \gcd(x+cy, y)$, for instance. So if the theorem is true, then we know what d_{11} is — it's the gcd of all the entries.

So even before doing any calculations, we know what d_{11} is. This suggests the main idea — we want to run the Euclidean Algorithm to decrease the numbers, and eventually get the gcd.

Proof of Theorem 6.17. Recall that in an Euclidean domain, we have a size function $\sigma : R \setminus \{0\} \to \mathbb{Z}_{\geq 0}$, where given any nonzero a and b, we can write $a = bq + r$, where $r = 0$ or $\sigma(r) < \sigma(b)$. We'll write $|a|$ instead of $\sigma(a)$.

If $B = 0$ then we're done; so assume $B \neq 0$. The key step is to put the gcd of all coefficients in the top-left corner.

Lemma 6.21

By row and column operations, we can arrive at a matrix B' such that $b'_{11} = \gcd(b_{ij}) = \gcd(b'_{ij})$.

Proof. First, by permuting rows and columns, we can guarantee that b_{11} is the nonzero element of minimal norm — so $|b_{11}| \leq |b_{ij}|$ for all i and j (with $b_{ij} \neq 0$).

Now if b_{11} divides all the b_{ij}, then we're done. If not, we want to modify the matrix to find a smaller entry.

If there exists b_{ij} not divisible by b_{11} in the first row or column, then we can directly apply a row or column operation to reduce it, using division with remainder: we can write $b_{ij} = qb_{11} + r$, and then subtract q times the first row or column from this element's row or column. This is the same as the Euclidean Algorithm.

If not, then all entries in the first row and column are divisible by b_{11}. So we can use row and column operations to kill the entire first row and column, except for b_{11} itself:

$$\left[\begin{array}{c|cccc} b_{11} & 0 & 0 & \cdots & 0 \\ \hline 0 & * & * & \cdots & * \\ 0 & * & b_{ij} & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{array}\right].$$

Now if there is an entry b_{ij} not divisible by b_{11}, we can bring it into the first row — since b_{i1} and b_{1j} are both 0, we can add row i to the first row, which brings b_{ij} to b_{1j} without changing b_{11}.

But now we can perform the previous step: add a multiple of the first column in order to make this entry smaller than b_{11}.

Now that we've created a smaller element, permute the rows and columns to put the smallest element in the top-left again; this means we've strictly decreased the size of the nonzero entry in the top-left.

Now we keep repeating this. We can't keep decreasing the size, since sizes are nonzero integers; so this must eventually terminate, at which point b_{11} divides all other entries. □

Now induct on $m+n$. By the lemma, we can go from B to B', where b'_{11} divides all b'_{ij}. Now since b_{11} divides all elements in the first row and column, we can again use it to kill them. So we've transformed the matrix into the form

$$\left[\begin{array}{c|cccc} b'_{11} & 0 & 0 & \cdots & 0 \\ \hline 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{array}\right].$$

Now we're done with the first row and column, and we can delete them and perform operations only on the remaining submatrix. By the inductive hypothesis, we can reduce that smaller matrix to the form we want, so we're done. □

Example 6.22

Convert the matrix

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

to Smith normal form.

Solution. First, permute the rows to get

$$\begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix}.$$

Now 1 already divides all other entries, so we can subtract to get

$$\begin{bmatrix} 1 & 3 \\ 0 & -5 \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 0 & -5 \end{bmatrix}.$$

□

Example 6.23

Convert the matrix

$$\begin{bmatrix} 4 & 2 & 6 \\ 1 & 2 & 3 \end{bmatrix}$$

to Smith normal form.

Solution. We can perform the series of operations

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 2 & 6 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 4 & -6 & -6 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 0 & -6 & -6 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 0 & -6 & 0 \end{bmatrix}.$$

□

Remark 6.24. In some sense, we're diagonalizing the matrix here. But this is a different kind of diagonalization than we saw in linear algebra with Jordan normal form — in that case we used conjugation, while here we use elementary operations.

§6.3.4 Classification of Abelian Groups

Taking R to be $\mathbb{Z}$, since a $\mathbb{Z}$-module is the same as an abelian group, we can get a classification of finitely presented abelian groups (we'll later see how to replace "finitely presented" with "finitely generated").

Corollary 6.25

A finitely presented abelian group is isomorphic to

$$\mathbb{Z}/d_1 \times \mathbb{Z}/d_2 \times \cdots \times \mathbb{Z}/d_n \times \mathbb{Z}^a,$$

where $d_1 \mid d_2 \mid \cdots \mid d_n$.

It's possible to rewrite this: the Chinese Remainder Theorem states that if $n = ab$ for relatively prime a and b, t hen $\mathbb{Z}/n \cong \mathbb{Z}/a \times \mathbb{Z}/b$. So factoring each d_i as a product of prime powers, we can decompose the terms $\mathbb{Z}/d_i$ as a product of cyclic groups with prime power order $\mathbb{Z}/p^m$.

We can look at some properties of this composition.

Question 6.26. Which parts of the decomposition are unique?

Note that $G \times H \supset G$ — this is because $G \times H$ consists of all elements (g, h), while G consists of the elements $(g, 1)$. So this means $\mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_n$ is a finite subgroup of our abelian group. So the product of the finite factors is then exactly the set of elements in the group with finite order — since if an element had a nonzero part in the free term, it wouldn't have finite order.

Definition 6.27. The set of elements with finite order is called the *torsion subgroup* A_f, so we have

$$A_f = \mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_n.$$

On the contrary, the free factor is not uniquely defined as a subgroup. For example, take $A = \mathbb{Z}/2 \times \mathbb{Z}$, so $A_f = \mathbb{Z}/2$. But A contains two free groups — one generated by $(0, 1)$, and the other generated by $(1, 1)$. Both are complementary to $\mathbb{Z}/2$, but they are not the same.

But it's easy to see that the rank a of the free factor *is* well-defined. This is because $\mathbb{Z}^a = A/A_f$, but $\mathbb{Z}^a \ncong \mathbb{Z}^b$ if $a \neq b$. (Otherwise we would have an $a \times b$ matrix B and $b \times a$ matrix C with $BC = 1_a$ and $CB = 1_b$. This is impossible even dropping the requirement that they have integer coefficients — if $a < b$ then $\operatorname{rank}(CB) \leq a < b = \operatorname{rank}(1_b)$, contradiction.)

We can further analyze the torsion subgroup A_f. By using the Chinese Remainder Theorem and grouping together factors corresponding to the same prime, we can write

$$A_f = A_{p_1} \times A_{p_2} \times \cdots \times A_{p_m},$$

where each factor is of the form $A_p = \prod \mathbb{Z}/p^j$.

Example 6.28

Write $\mathbb{Z}/36 \times \mathbb{Z}/6$ in this form.

Solution. We can split $36 = 4 \cdot 9$ and $6 = 2 \cdot 3$, to get $(\mathbb{Z}/4 \times \mathbb{Z}/2) \times (\mathbb{Z}/9 \times \mathbb{Z}/3)$. □

Note that A_p is a p-Sylow subgroup of A_f. Since the group is abelian, by the Sylow Theorems A_p is uniquely defined. In fact A_p is exactly the set of elements whose order is a power of p — this is called the *p-torsion subgroup*.

Lemma 6.29

The multiplicities of the powers of p in the decompositions $A_p = \prod \mathbb{Z}/p^j$ are uniquely defined.

For example, this means $\mathbb{Z}/4 \times \mathbb{Z}/4$ is not isomorphic to $\mathbb{Z}/2 \times \mathbb{Z}/8$.

Proof. Let $A = \mathbb{Z}/p^{a_1} \times \cdots \times \mathbb{Z}/p^{a_n}$. There are two main observations here.

First consider A/pA. Each term contributes $\mathbb{Z}/p$, so $A/pA = (\mathbb{Z}/p)^n$. This means $|A/pA| = p^n$, where n is the number of factors — so any two decompositions must have the same number of factors.

Meanwhile, we can also look at pA. We have $p\mathbb{Z}/p^a \cong \mathbb{Z}/p^{a-1}$. So replacing A with pA reduces each of the exponents by 1, and

$$pA = \prod \mathbb{Z}/p^{a_i - 1}.$$

(It's possible that some of these factors are trivial.)

Now use induction on $|A|$. If

$$A = \mathbb{Z}/p^{a_1} \times \cdots \times \mathbb{Z}/p^{a_n} = \mathbb{Z}/p^{a_1'} \times \cdots \times \mathbb{Z}/p^{a_m},$$

then we must have $n = m$ by the first observation, and

$$pA \cong \prod \mathbb{Z}/p^{a_i - 1} = \prod \mathbb{Z}/p^{a_i' - 1}$$

by the second. But by the induction hypothesis, we can match all $a_i > 1$ with $a_i' > 1$. Meanwhile, since $m = n$, we can also match the $a_i = 1$ with $a_i' = 1$. □

§6.3.5 Polynomial Rings

Now consider the decomposition in the case of $R = F[t]$, where F is a field. The theorem says that a finitely presented module is of the form

$$M \cong R/(P_1) \times \cdots \times R/(P_n) \times R^a,$$

where $P_1 \mid P_2 \mid \cdots \mid P_n$. Similarly to before, we can rewrite the decomposition as

$$M \cong \prod R/Q_i^{a_i} \times R^a,$$

where the Q_i are irreducible.

In particular, consider M which are finite-dimensional as a vector space over F. Then there should be no free factor R^a.

When we started discussing modules, we saw that a $F[t]$ module is the same as a F-vector space V, together with a linear operator $V \to V$ (the action of t). So understanding isomorphism classes of modules is the same as understanding this situation, which was studied in linear algebra.

If $P(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_0$, recall that $R/(P)$ has a basis consisting of $\overline{1}$, $\overline{t}$, …, $\overline{t^{n-1}}$. Let $e_i = \overline{t^{i-1}}$. Then we have $te_i = e^{i+1}$ for $1 \le i \le n-1$, while

$$te_n = -a_0e_1 - a_1e^2 - \cdots - a_{n-1}e_n.$$

So then the matrix corresponding to t is

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ 0 & 0 & 1 & \cdots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -a_n \end{bmatrix}.$$

Now consider $F = \mathbb{C}$, where the only irreducible polynomials are linear — so we have $Q_i(t) = t - \lambda_i$. If $\lambda_i = 0$, then using the above basis, we get a matrix with 0's on the diagonal, 1's directly below the diagonal, and 0's everywhere else — this is exactly Jordan normal form where the diagonal is 0.

Meanwhile, in general, we can use the basis $\overline{(t-\lambda_i)^j}$ instead of $\overline{t^j}$. Then the situation is the same, except that we add a scalar to the matrix (because the action of $t - \lambda_i$ corresponds to this matrix, so we add the scalar matrix λ_i to get the action of t), so we get the same matrix with λ_i on the diagonal instead of 0.

So this actually proves Jordan normal form.

§6.4 Noetherian Rings

Definition 6.30. A ring R is *Noetherian* if every ideal in R is finitely generated.

For example, if R is a field or a PID, it's clearly Noetherian.

The notion of Noetherian rings is useful for the following reason:

Proposition 6.31

A ring R is Noetherian iff every submodule in a finitely generated R-module is itself finitely generated.

Corollary 6.32

If R is Noetherian, then every finitely generated module is finitely presented.

So if R is Noetherian, then we can actually replace the "finitely presented" condition above with "finitely generated." Meanwhile, the Hilbert Basis Theorem states that if R is Noetherian, so is $R[x]$. This gives us a powerful tool for proving that many rings are Noetherian.

Before we prove this, we'll look at a few observations.

Lemma 6.33

If we have a surjective homomorphism $\varphi : M \to N$ of R-modules, then:

1. If M is finitely generated, then N is also finitely generated.
2. If N is finitely generated, and $K = \ker(\varphi)$ is also finitely generated, then M is also finitely generated.

Remark 6.34. This is one place where the intuition from linear algebra is useful: it's helpful to think about the case where R is a field, so being finitely generated is equivalent to being finite-dimensional. In this case, we know more precisely that

$$\dim(M) = \dim(N) + \dim(K).$$

We won't get information this precise in the general case, but proving finiteness is fairly similar.

Proof of Lemma 6.33. The first part is obvious — take any set of generators $m_1, \dots, m_n$ for M. Then their images $\varphi(m_1), \dots, \varphi(m_n)$ must generate N.

For the second part, let $k_1, \dots, k_a$ be a set of generators for K, and $n_1, \dots, n_b$ a set of generators for N. Now pick $\tilde{n}_1, \dots, \tilde{n}_b$ such that $\varphi(\tilde{n}_i) = n_i$ (which we can do by surjectivity).

Now we claim that the k_i and $\tilde{n}_i$ generate M: given any $x \in M$, we can find $r_1, \dots, r_b$ in R such that

$$\varphi(x) = r_1 n_1 + \cdots + r_b n_b.$$

So then we have

$$\varphi(x - r_1\tilde{n}_1 - \cdots - r_b\tilde{n}_b) = \varphi(x) - r_1 n_1 - \cdots - r_b n_b = 0,$$

which means $x - \sum r_i\tilde{n}_i$ is in K. So we can express $x - \sum r_i\tilde{n}_i = \sum s_j k_j$, which gives an expression for x as a linear combination of the $\tilde{n}_i$ and k_j. □

Proof of Proposition 6.31. One direction is clear: an ideal of R, by definition, is the same as a submodule of the free module (which is generated by one element). So if every submodule of a finitely generated module is finitely generated, then R must be Noetherian.

For the other direction, we want to show that if every ideal is finitely generated, then so is every submodule of a finitely generated module.

The strategy is to reduce to the case of a free module R^n, and use induction on n — we know this is true for $n = 1$, so we want to reduce to this case.

Let M be a finitely generated module. Then by picking a set of generators, we can find a surjective homomorphism $\varphi : R^n \to M$ (fixing such a homomorphism is equivalent to fixing a set of generators).

Then by the Correspondence Theorem and the above lemma, it's enough to check that every submodule of R^n is finitely generated (since the submodules of M are exactly the images of the submodules of R^n containing $\ker\varphi$).

Now we can argue by induction on n. First, the base case $n = 1$ follows directly from the definition of a Noetherian ring, since submodules of R are exactly ideals.

For the inductive step, consider a submodule $N \subset R^n$, with $n > 1$. Now split $R^n = R \times R^{n-1}$, and take the projection homomorphism $\pi : R^n \to R^{n-1}$, which sends $(r_1, \dots, r_n) \mapsto (r_2, \dots, r_n)$.

Then $\pi(N)$ is a submodule of R^{n-1}, so by the induction assumption, it's finitely generated. Meanwhile, the kernel K of π is the set of elements of the form $(r, 0, 0, \dots)$ which are in N. But this is a submodule of the free rank-1 module R, so K is also finitely generated.

So by the lemma, since K and $\pi(N)$ are both finitely generated, N is finitely generated as well. □

Remark 6.35. For example, when considering submodules of $\mathbb{Z}^2$, we'd take the points on the x-axis as K, and the projections onto the y-axis as $\pi(N)$. Note that N is not necessarily $K \times \pi(N)$.

So now we have that if R is Noetherian, any finitely generated module is also finitely presented:

Proof of Corollary 6.32. If the module is finitely generated, then there is a surjective map $\varphi : R^n \to M$. Then $\ker\varphi$ is a submodule of R^n, so it must be finitely generated as well. □

This means the classification we saw earlier is actually a classification of all finitely *generated* abelian groups.

So we've seen why this notion is useful. But in order to use it, we want to see how to produce more examples of Noetherian rings, beyond just fields and PIDs.

First, there is a simple observation we can make:

Lemma 6.36

A quotient of a Noetherian ring is again Noetherian — if R is a Noetherian ring and I an ideal of R, then the ring $S = R/I$ is also Noetherian.

Proof. This is immediate from the Correspondence Theorem: an ideal in R/I is of the form $\overline{J}$, where $J \subset R$ is an ideal containing I (here the bar denotes taking mod I). Then we can just take the images of the generators — if $J = (x_1, \dots, x_n)$, then $\overline{J} = (\overline{x}_1, \dots, \overline{x}_n)$. So all ideals of R/I are finitely generated. □

Remark 6.37. A subring in a Noetherian ring is not necessarily Noetherian — so this is more subtle than the dimension of a vector space.

For example, $\mathbb{C}[x, y]$ is Noetherian. But the subring $\mathbb{C} + x\mathbb{C}[x, y]$ consisting of polynomials which are constant mod x is *not* Noetherian — the ideal $x\mathbb{C}[x, y]$ is not finitely generated.

§6.4.1 Hilbert Basis Theorem

There's actually a powerful tool that shows many rings are Noetherian:

Theorem 6.38 (Hilbert Basis Theorem)

If R is Noetherian, then $R[x]$ is also Noetherian.

Remark 6.39. When Hilbert proved this theorem in 1890, there's a legend that a famous mathematician Paul Gordan (referred to as the king of invariant theory) said this is not mathematics, it's theology. Previously, people had to work with rings case by case, and concretely produce finitely many elements generating an ideal. In contrast, this theorem has a very abstract proof that doesn't give much information about how to actually write down the generators.

This theorem has useful implications:

Corollary 6.40

If R is Noetherian, then $R[x_1, \dots, x_n]/I$ is also Noetherian, for any ideal I.

Corollary 6.41

Any algebraic subset in $\mathbb{C}^n$ (a subset given by a collection of polynomial equations) is always given by a *finite* set of polynomial equations.

Proof of Theorem 6.38. Let $I \subset R[x]$ be an ideal, so we want to check that I is finitely generated. It's enough to find a finite collection of polynomials $P_1, \dots, P_n$ in I and a bound d, such that every element in I can be reduced to a polynomial of degree d — meaning that

$$I \subset (P_1, \dots, P_n) + R[x]_{\le d}.$$

In other words, once the polynomials and d are fixed, then for every $P \in I$, we need to be able to find $Q_1, \dots, Q_n$ in $R[x]$ such that

$$\deg(P - \sum Q_i P_i) \le d.$$

If we know this, then

$$I \subset (P_1, \dots, P_n) + (I \cap R[x]_{\le d}).$$

But the second term is finitely generated over R, since $R[x]_{\le d}$ is a free module of rank $d+1$ over R, and R is Noetherian. So if it's generated by $S_1, \dots, S_m$, then I is generated by the P_i and S_i.

So now we want to figure out how to do this — in some sense, this is generalized division with remainder.

Consider the ideal $\overline{I}$ in R consisting of the *leading coefficients* of polynomials in I (along with 0). Then $\overline{I}$ is finitely generated. Let $P_1, \dots, P_n$ be polynomials whose leading coefficients generate $\overline{I}$, and let $d = \max(\deg P_i)$.

Now if we have a polynomial P of degree greater than d, we can cancel its leading coefficient — we can find Q_i such that $\sum Q_i P_i$ has the same degree and same leading coefficient as P, and then subtract them to decrease the degree of P by at least 1. We can then repeat this until P has degree at most d. □

Remark 6.42. This proof is not very constructive — it's not an effective way of finding generators.

§6.4.2 Chain Conditions

Proposition 6.43

A ring is Noetherian iff every increasing chain of ideals stabilizes. In other words, if

$$I_1 \subseteq I_2 \subseteq \cdots,$$

then from some point on, we have $I_n = I_{n+1} = \cdots$.

Example 6.44

Consider the case $R = \mathbb{Z}$.

Solution. Here a chain of ideals amounts to a list of integers d_1, d_2, ... where $d_i \mid d_{i-1}$ for all i. Clearly the sequence must stabilize, since it's a decreasing sequence of positive integers. □

Proof of Proposition 6.43. First suppose R is Noetherian, and we have a chain $I_1 \subseteq I_2 \subseteq \cdots$. Then their union $I = I_1 \cup I_2 \cup \cdots$ is an ideal. Since R is Noetherian, then I is finitely generated, so we can write $I = (a_1, \ldots, a_n)$. Then for each i, a_i must be contained in some ideal (since it's in their union). But there's finitely many a_i, so some I_m must contain all of them (since $I_k \subseteq I_{k+1}$, once a_i appears in one ideal, it appears in all following ones). So then $I_m = I$, and the sequence must stabilize at I_m — we must have $I_m = I_{m+1} = \cdots = I$.

For the other direction, suppose R is *not* finitely generated, and let $I \subset R$ be an ideal that's not finitely generated. Pick $a_1 \in I$, and define a_n inductively such that $a_n \in I$ but $a_n \notin (a_1, \ldots, a_{n-1})$ — this is possible because otherwise I would be finitely generated. Now we can take $I_n = (a_1, \ldots, a_n)$, which gives an infinite non-stabilizing chain of ideals. □

Remark 6.45. Sometimes the chain condition is given as the *definition* of Noetherian rings.

This has an application to unique factorization in PIDs. Previously, we proved uniqueness but not existence — it's nontrivial to prove existence in an abstract PID with no concept of a norm. But now if we can't factor an element into irreducibles, then we have an infinite factorization process, which produces an infinite chain of ideals. So this chain condition implies factoring terminates.

Proposition 6.46

In a Noetherian ring, every (non-unit) ideal is contained in a maximal ideal.

This is actually true in *any* ring, but the proof requires set theory. The rings which we are interested in are all Noetherian, so are covered by this.

Proof. Let $I \subset R$ be an ideal, and assume I is not contained in a maximal ideal. Set $I_1 = I$. Now find $I_2 \supset I$ (not equal to I or R), which is possible since I is not maximal. But I_2 is not maximal either, so we can similarly construct $I_3 \supset I_2$, and inductively build a chain $I_1 \subsetneq I_2 \subsetneq I_3 \subsetneq \cdots \subsetneq R$ — this is possible at every step because if not, I_n would be a maximal ideal containing I. This contradicts chain termination. □

Remark 6.47. As a final remark on modules: an important tool in studying modules is an upgrade on the presentation with generators and relations. The idea is that once we have generators and relations, we can also look at the relations between relations, and so on. First construct a surjective map $R^n \to M$ by fixing generators. This map has a kernel K_1, and we can construct a surjective map $R^m \to K_1$ by fixing its generators. This map has a kernel K_2, and we can construct a surjective map $R^\ell \to K_2$, and so on. This is called the *syzergy* or *free resolution*.

§7 Fields

Definition 7.1. A *field extension* is a pair of fields $L \supset K$. The extension is written as L/K.

The theory of field extensions developed from trying to understand how to systematically solve polynomial equations. We know a formula for solving quadratics; people were interested in seeing whether more general formulas existed.

An important example of a field extension is $L = K(\alpha)$, where α is the root of an irreducible polynomial. If P is a generic polynomial in $\mathbb{Q}[x]$ of degree at least 5, and $K = \mathbb{Q}(\alpha)$ for a root α of P, then we'll see that K is not contained in a field of the form $\mathbb{Q}(\beta_1, \dots, \beta_n)$ where $\beta_i^{d_i} \in \mathbb{Q}(\beta_1, \dots, \beta_{i-1})$ for each i (and some positive integers d_i). So it follows that there is no formula for the roots of P involving just radicals.

Another application is to compass and straightedge construction: let $\zeta_n \in \mathbb{C}$ be a primitive nth root of unity.

Question 7.2. When is $\mathbb{Q}(\zeta_n)$ contained in a field of the form $\mathbb{Q}(\alpha_1, \dots, \alpha_n)$ where $\alpha_i^2 \in \mathbb{Q}(\alpha_1, \dots, \alpha_{i-1})$ for all i?

In other words, we want to find out whether we can obtain ζ_n by the regular operations along with extracting square roots. This is interesting because it happens iff a regular n-gon can be constructed by a compass and straightedge. We'll see how to get a complete answer — this characterization was proved by Gauss. For example, the answer is yes for $n = 17$ and no for $n = 19$.

§7.1 Field Extensions

Definition 7.3. An extension L/K is *finite* if L is finite-dimensional as a vector space over K.

Essentially, what this means is that if we forget that we can multiply by elements of L, but remember that we can add them and multiply by elements of K, then we should get a finite-dimensional vector space.

We've already seen how to construct examples of finite extensions: if $P \in K[x]$ is an irreducible polynomial, then we saw that $K[x]/(P)$ is a field. (This is because $K[x]$ is a PID, so (P) is a maximal ideal.) In this case, the dimension of the vector space is $d = \deg P$, since we saw the monomials 1, x, $\dots$, x^{d-1} form a basis.

Definition 7.4. The *degree* of the extension L/K, denoted $[L : K]$, is the dimension of L as a vector space over K.

On the other hand, suppose we start with an extension L/K, and pick an element $\alpha \in L$. We say α is *algebraic* over K if it satisfies a polynomial equation — meaning that $P(\alpha) = 0$ for some nonzero $P \in K[x]$.

If α is algebraic, then we can take its minimal polynomial P (the monic polynomial of smallest degree with $P(\alpha) = 0$ — this is unique because all polynomials with $P(\alpha) = 0$ form an ideal, and all ideals are principal and generated by their minimal-degree element). The minimal polynomial has to be irreducible — if $P = P_1P_2$, then $P(\alpha) = P_1(\alpha)P_2(\alpha) = 0$, but since we're in a field, this would imply one of the factors is 0, contradicting minimality.

Then we have a homomorphism $K[x]/(P) \to L$ sending $x \mapsto \alpha$. But any homomorphism of fields is injective (alternatively, the kernel of the map $f : K[x] \to L$ sending $x \mapsto \alpha$ is exactly the set of polynomials P with $P(\alpha) = 0$, which is generated by the minimal polynomial of α). So we have an isomorphism $K[x]/(P) \cong K(\alpha)$, where $K(\alpha)$ is the subfield of L generated by α.

Remark 7.5. This isomorphism is important: later we'll look at automorphisms of fields, and this is the trick that will let us build such maps. For example, start with $K = \mathbb{Q}$ and $L = \mathbb{C}$, and take $\alpha = \sqrt[3]{2}$ and $\beta = \sqrt[3]{2}\omega$ for a primitive third root of unity ω. Then α and β are both roots of the irreducible polynomial $x^3 - 2 = 0$. So $\mathbb{Q}(\alpha) \cong \mathbb{Q}(\beta)$, since both extensions are isomorphic to the abstract construction $\mathbb{Q}[x]/(x^3 - 2)$.

Lemma 7.6

If we have a field extension L/K, then α is algebraic iff $K(\alpha)$ is finite-dimensional over K.

Proof. We've already seen that if α is algebraic, then $K(\alpha)$ is finite-dimensional. Meanwhile, a polynomial relation can be thought of as a linear relation between the powers of α. If $m > \dim_K K(\alpha)$, then 1, α, $\dots$, α^m must be linearly dependent; that linear dependence corresponds to a polynomial over K. □

Corollary 7.7

If L/K is finite, then every $\alpha \in L$ is algebraic over K.

§7.2 Towers of Extensions

Proposition 7.8

Suppose that we have a tower of field extensions $K \supset E \supset F$, where K/E and E/F are finite. Then K/F is finite, and

$$[K : F] = [K : E] \cdot [E : F].$$

Proof. Let $\alpha_1, \dots, \alpha_n$ be a basis for E as a vector space over F, and $\beta_1, \dots, \beta_m$ a basis for K as a vector space over E. Then we'll show that the terms $\alpha_i\beta_j$ form a basis for K/F.

But this is clear if we substitute notation. First, in order to see that this is a generating set, every $x \in K$ can be written as $x = \sum \lambda_i\beta_i$, while each $\lambda_i \in E$ can be written as $\lambda_i = \sum a_{ij}\alpha_j$, which gives

$$x = \sum a_{ij}\alpha_j\beta_i.$$

Similarly, to prove independence, if we have $\sum a_{ij}\alpha_j\beta_i = 0$, we can collect terms into

$$\sum \left(\sum a_{ij}\alpha_j\right)\beta_i = 0.$$

If the a_{ij} are not all 0, then one of the terms $\sum a_{ij}\alpha_j$ must be nonzero (since the α_j are linearly independent), and therefore the entire sum must be nonzero (since the β_i are linearly independent). □

Example 7.9

Find $[\mathbb{Q}(\alpha, \beta) : \mathbb{Q}]$, where $\alpha = \sqrt[3]{2}$ and $\beta = \sqrt[3]{2}\omega$.

Solution. We saw α and β both have minimal polynomial $x^3 - 2$. Then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$. Meanwhile, $x^3 - 2$ factors in $\mathbb{Q}(\alpha)$ as $(x - \alpha)(x^2 + \alpha x + \alpha^2)$. So $[\mathbb{Q}(\alpha, \beta) : \mathbb{Q}(\alpha)] = 2$, which means $[\mathbb{Q}(\alpha, \beta) : \mathbb{Q}] = 6$. □

We have the following useful fact about fields:

Fact 7.10 — Every field is a (possibly infinite) extension of either $\mathbb{Q}$, or $\mathbb{F}_p$ for a prime p.

Proof. Recall that for any ring R, there is a unique ring homomorphism $\mathbb{Z} \to R$, where $1 \mapsto 1_R$ and $n \mapsto \underbrace{1_R + \cdots + 1_R}_{n} = n_R$.

The image is a quotient of $\mathbb{Z}$. Meanwhile, we can look at the kernel. If R is a domain, then the kernel either is trivial (meaning the homomorphism is injective) or is (p) for a prime p — otherwise the image would be $\mathbb{Z}/n$ for n not prime, which is not a domain (as it has zero divisors).

Now if $R = F$ is a field, the generator of the kernel is called the *characteristic* of the field.

If $\operatorname{char}(F) = 0$, then $\mathbb{Z}$ is a subring of F. But then we must have a copy of $\mathbb{Q} = \operatorname{Frac}(\mathbb{Z})$ inside F, where this copy of $\mathbb{Q}$ is the set of fractions $\frac{n_R}{m_R}$.

On the other hand, if $\operatorname{char}(F) = p$, then we have $\mathbb{Z}/p \subset F$. □

Now we'll return to towers of extensions $E/F/K$ — here E/K is called the *composite* extension, while E/F and F/K are called *intermediate* extensions. We saw that $[E : K] = [E : F] \cdot [F : K]$; in particular, E/K is finite iff E/F and F/K are finite.

Corollary 7.11

If $\alpha, \beta \in L$ are algebraic over K, then $\alpha + \beta$, $\alpha\beta$, and $\frac{\alpha}{\beta}$ are also algebraic.

Proof. If α and β are algebraic, then $K(\alpha)/K$ and $K(\alpha, \beta)/K(\alpha)$ are both finite — since β satisfies a polynomial relation with coefficients in K, it satisfies the same polynomial relation with coefficients in $K(\alpha)$. So we can conclude that $K(\alpha, \beta)$ is finite, and therefore any element in it is algebraic. □

Corollary 7.12

Given an arbitrary extension, the set of elements in L which are algebraic over K form a subfield of L, called the *algebraic closure* of K in L.

This is an abstract argument that doesn't exactly tell us how to construct the polynomial; but it's possible to come up with a procedure to write down an equation as well.

Example 7.13

Let $\alpha = \sqrt{2}$ and $\beta = \sqrt{3}$, and $\gamma = \alpha + \beta$. Write down a polynomial equation for γ.

Solution. One possible way is to consider 1, γ, γ^2, $\ldots$. These are all linear combinations of 1, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{6}$, with coefficients in $\mathbb{Q}$. So they lie in a vector space of dimension at most 4, and we can use linear algebra to find the linear relation between 1, γ, $\ldots$, γ^4.

Alternatively, we'd like to find a polynomial with γ as a root, so we can try to think about what the other roots should be. We want $\sqrt{2}+\sqrt{3}$ to be a root, so we can guess that $\sqrt{2}-\sqrt{3}$ should also be a root — from an algebraic perspective, we should be able to switch the sign of the square root. Similarly, $-\sqrt{2}+\sqrt{3}$ and $-\sqrt{2}-\sqrt{3}$ should be roots. We can expand the polynomial with these four roots, and see that this does in fact work.

Essentially, the idea we used here is that there's a group of symmetries acting on the roots (by multiplying them by ± 1); we'll discuss this more in Galois theory. □

§7.2.1 Compass and Straightedge Construction

We can consider another corollary of Proposition 7.8:

Corollary 7.14
If $E/F/K$ is a tower of finite extensions, then $[F:K] \mid [E:K]$.

It's a fact (mentioned earlier) that a regular n-gon is constructible with compass and straightedge iff ζ_n lies in an extension $\mathbb{Q}(\alpha_1, \alpha_n)$ such that $\alpha_i^2 \in \mathbb{Q}(\alpha_1, \dots, \alpha_{n-1})$ for all i — we won't discuss the details here. This means we have a tower of extensions $F_n/F_{n-1}/\cdots/F_1/F_0$ where $F_0 = \mathbb{Q}$, and $[F_i : F_{i-1}] = 2$ for all i (the degree could be 1, but this is not useful).

Theorem 7.15
Let $n = p$ be prime. Then a regular p-gon can be constructed iff $p = 2^k + 1$.

Primes $p = 2^k + 1$ are called *Fermat primes*. There's only 5 known Fermat primes (3, 17, 257, and 65537); it's conjectured that there are no others, but we don't even know whether there's finitely or infinitely many.

We'll show one direction: that if ζ_p is constructible, then p is a Fermat prime.

Proof of necessity. First we want to find $\deg(\zeta_p)$, or the degree of the extension $[\mathbb{Q}(\zeta_p) : \mathbb{Q}]$ (which is called a *cyclotomic* extension).

We know ζ_p is a root of the polynomial $x^p - 1$, which we can immediately factor as

$$x^p - 1 = (x-1)(x^{p-1} + x^{p-2} + \cdots + 1).$$

We claim that the factor on the right is irreducible — we'll prove this via Eisenstein's criterion. First, it's primitive, so in order to show that it's irreducible over $\mathbb{Q}$, it suffices to show that it's irreducible over $\mathbb{Z}$. But we can shift to $t = x - 1$, so then

$$Q(t) = \frac{(t+1)^p - 1}{t} \equiv t^{p-1} \pmod{p}.$$

All non-leading coefficients of $Q(t)$ are divisible by p, and the free term p is not divisible by p^2. Eisenstein gives that such polynomials are always irreducible — to prove this, if $Q = Q_1 Q_2$ where Q_1 and Q_2 have positive degree, then we can reduce mod p to get $t^{p-1} = \overline{Q_1}\,\overline{Q_2}$, where both factors have nonzero degree. So both factors must have free term divisible by p, and the free term of their product must be divisible by p^2, contradiction.

So we have $\deg(\zeta_p) = p - 1$. On the other hand, if $\zeta_p \in F_n$ for a field extension of the form described, then $\deg(\zeta_p)$ must divide $[F_n : \mathbb{Q}]$, which is a power of 2. So $p - 1$ must be a power of 2. □

With our current tools, we can only show one direction — to show the other direction, we need a better extension of which fields can be obtained as the top floor of such a tower (it's necessary that the degree is a power of 2, but this may not be sufficient). In the case of $\mathbb{Q}(\zeta_p)$, the condition turns out to be sufficient as well.

§7.3 Splitting Fields

We've seen the construction where we start with an irreducible polynomial $P \in F[x]$, and construct $E = F[x]/(P)$. This is an extension of F of degree $n = \deg(P)$, and we can think of it as adjoining a root of the polynomial.

But there's another construction which produces a finite extension from a polynomial, which is harder to control.

Definition 7.16. Given a (not necessarily irreducible) polynomial $P \in F[x]$, a *splitting field* of P is an extension E/F such that:

1. P splits as a product of linear factors in $E[x]$;
2. $E = F(\alpha_1, \ldots, \alpha_n)$, where the α_i are the roots of P (in the above splitting).

Example 7.17

Find the splitting field of $P(x) = x^3 - 2$ over $F = \mathbb{Q}$ and $F = \mathbb{Q}(\omega)$, where ω is a primitive 3rd root of unity.

Solution. The splitting field of P over $F = \mathbb{Q}$ is $E = \mathbb{Q}(\sqrt[3]{2}, \omega\sqrt[3]{2})$, and we saw earlier that $[E : F] = 6$.

On the other hand, if $F = \mathbb{Q}(\omega)$, then P is still irreducible. But if we add one root then it becomes reducible — if α is a root, we have $x^3 - 2 = (x - \alpha)(x - \omega\alpha)(x - \omega^2\alpha)$. So $E = F(\sqrt[3]{2})$, and we have $[E : F] = 3$. □

Example 7.18

Find the splitting field of $P(x) = x^{p-1} + \cdots + 1$ over $F = \mathbb{Q}$.

Solution. All roots of P are of the form ζ_p^i for some i. So the splitting field is $\mathbb{Q}(\zeta_p)$, and $[E : F] = p - 1$. □

The splitting field is actually unique:

Proposition 7.19

If F is a field, and P a (not necessarily irreducible) polynomial in F, then there exists a unique extension E/F up to isomorphism, such that P splits as a product of linear factors in $E[x]$ as $P(x) = \prod(x - \alpha_i)$, and $E = F(\alpha_1, \ldots, \alpha_n)$.

Proof. The idea of the proof is fairly easy: we essentially add in roots one by one, and uniqueness follows from uniqueness in adjoining a root of an irreducible polynomial.

Use induction on the degree of P. Let P_1 be an irreducible factor of P, and let $F_1 = F[x]/(P_1)$. Then in $F_1[x]$, we have a factorization $P(x) = (x - \alpha)Q(x)$, where α is a root of P_1 (since this construction essentially adjoins a root of P_1 to F).

Now let E be the splitting field for Q over F_1. Then we claim E is also a splitting field for P over F. This follows directly from the definition: P splits completely in $E[x]$, and we know that $E = F_1(\alpha_2, \ldots, \alpha_n)$ where $\alpha_2, \ldots, \alpha_n$ are the roots of Q. But $F_1 = F(\alpha_1)$, so then $E = F(\alpha_1, \ldots, \alpha_n)$.

So we've proved existence of the splitting field. To prove uniqueness, suppose that E' is another splitting field; we'll construct an isomorphism between E' and E.

First, we can find a root α' of P_1 in E'. Then if we set $F_1' = F(\alpha') \subset E'$, we know that $F(\alpha') \cong F(\alpha)$, since both are isomorphic to $F[x]/(P)$. This isomorphism sends $Q \in F_1[x]$ to $Q' \in F_1'[x]$, where $P = (x - \alpha')Q'$. Now E' is a splitting field for Q' over F_1'. So uniqueness of the splitting field of Q implies that the isomorphism between F_1 and F_1' extends to an isomorphism between E and E'. □

§7.4 Finite Fields

Previously, we discussed the primary fields — every field contains either $\mathbb{Q}$ or $\mathbb{F}_p$ for some p. If F is now a finite field, then it can't contain $\mathbb{Q}$; so it contains $\mathbb{F}_p$ for some p.

Meanwhile, $F/\mathbb{F}_p$ must be finite as well — i.e. F is finite-dimensional as a $\mathbb{F}_p$-vector space. If this dimension is $n = [F : \mathbb{F}_p]$, then we have $|F| = p^n$. (We can forget about multiplication and choose a basis for F as a $\mathbb{F}_p$-vector space; this identifies F with $\mathbb{F}_p^n$, which has size p^n.)

So all finite fields have order $q = p^n$ for some n.

Theorem 7.20

For every prime p and every $n \geq 1$, there exists a field of $q = p^n$ elements; and any two such fields are isomorphic.

So there exists a unique field of q elements, denoted by $\mathbb{F}_q$.

Remark 7.21. This field is *not* $\mathbb{Z}/q\mathbb{Z}$ (if $n > 1$) — they are not isomorphic even as additive groups.

We could construct a field of q elements if we could find an irreducible polynomial of degree n in $\mathbb{F}_p[x]$. For example, if $p = 4k + 3$, then $\mathbb{F}_p[x]/(x^2 + 1) = \mathbb{F}_{p^2}$; similarly for $p = 2$, we have $\mathbb{F}_2[x]/(x^3 + x^2 + 1) = \mathbb{F}_8$. It's possible to prove that an irreducible polynomial of degree n always exists, by a counting argument; but this is not the approach we will use here.

Instead, we'll use a sort of magic trick:

Definition 7.22. The *Artin-Schreier polynomial* is the polynomial $A(x) = x^q - x$.

Lemma 7.23

If F is a field containing $\mathbb{F}_p$, then

$$\{x \in F \mid x^q - x = 0\}$$

is a subfield in F.

Proof. It suffices to check two properties: for α and β in F with $A(\alpha) = A(\beta) = 0$, we have $A(\alpha\beta) = A(\alpha + \beta) = 0$. (We also need to check that $A(\alpha^{-1}) = 0$.)

The multiplicative properties are obvious, and work even with an arbitrary exponent: we have $\alpha^q = \alpha$ and $\beta^q = \beta$, so

$$(\alpha\beta)^q = \alpha^q \beta^q = \alpha\beta.$$

Meanwhile, $A(\alpha + \beta) = 0$ follows from the identity that in a ring containing $\mathbb{F}_p$, we have

$$(x + y)^p = x^p + y^p.$$

This follows from the Binomial Theorem, since $\binom{p}{i}$ is divisible by p for all $1 \leq i \leq p - 1$. Now by induction we get $(x + y)^q = x^q + y^q$, so

$$(\alpha + \beta)^q = \alpha^q + \beta^q = \alpha + \beta,$$

as desired. □

We can use this to prove the theorem:

Proof of Theorem 7.20. First we'll show uniqueness: suppose F is a field of $q = p^n$ elements. Now consider the multiplicative group F^*, which is an abelian group of $q-1$ elements. The order of any element divides $|F^*| = q-1$, so $\alpha^{q-1} = 1$ for all nonzero α. This means

$$\alpha^q = \alpha \implies A(\alpha) = 0$$

for *all* $\alpha \in F$.

Now $A(x)$ is a polynomial of degree q, which has q roots in a field. It follows that $x - \alpha \mid A(x)$ for all $\alpha \in F$, and by unique factorization, we then have

$$A(x) = \prod_{\alpha \in F} (x - \alpha).$$

So F is the splitting field of A over $\mathbb{F}_p$, and uniqueness of F follows from uniqueness of the splitting field.

Meanwhile, for existence, let F be the splitting field of A. So we need to check that F has exactly q elements.

From the lemma, we see that $A(\alpha) = 0$ for all $\alpha \in F$ — this is because F is generated by the roots of A, but the roots form a subfield. In particular, this means $|F| \leq \deg A = q$.

In order to check that we actually have $|F| = q$, we need to see that A has no multiple roots in F (since counting elements of F corresponds to counting roots of A without multiplicity). We can do this by using derivatives: in analysis, we've seen that a root of higher order means the polynomial is tangent to the x-axis. In this setting we can't use limits, but we can use formal derivatives: the formal derivative of the polynomial $P(x) = a_n x^n + \cdots + a_0$ is

$$P'(x) = n a_n x^{n-1} + \cdots + a_1.$$

Then it's easy to prove that $(P+Q)' = P' + Q'$, and $(PQ)' = P'Q + PQ'$. But then if P has a multiple root α, we can write $P(x) = (x-\alpha)^2 Q(x)$, which means

$$P'(x) = 2(x-\alpha)Q(x) + (x-\alpha)^2 Q'(x),$$

which is divisible by $x - \alpha$. So it's still true that a multiple root is a root of the derivative.

In particular, if $\gcd(P, P') = 1$, then there are no multiple roots. But in this case, we have $A'(x) = qx^{q-1} - 1 = -1$, since F has characteristic $p \mid q$. (Note that we had a nonlinear polynomial whose derivative is constant.) So clearly $\gcd(A, A') = 1$, and A has no multiple roots. This means $|F| = q = p^n$. □

There is more that we can say about the structure of finite fields.

Lemma 7.24

If F is any field, and G is a finite subgroup of F^*, then G is cyclic.

Example 7.25

If $F = \mathbb{C}$, then finite subgroups of F^* are the nth roots of unity

$$\left\{ \exp \frac{2\pi i}{n} \right\} = \langle \zeta_n \rangle \cong \mathbb{Z}/n.$$

Proof of Lemma 7.24. By the classification of finite abelian groups, we know $G \cong \prod \mathbb{Z}/p_i^{n_i}$. So it's enough to check that no prime appears twice — then we can use the Chinese Remainder Theorem.

But suppose p appears twice. Then G contains a subgroup $\mathbb{Z}/p^a \times \mathbb{Z}/p^b$. But then G has at least p^2 elements of order dividing p (since there's p choices for the coordinate in each). This would mean the polynomial $x^p - 1$ has at least p^2 roots; but it has degree p, so this is impossible. □

Corollary 7.26

For any finite field $\mathbb{F}_q$, its multiplicative group $\mathbb{F}_q^*$ is cyclic, meaning $\mathbb{F}_q^* \cong \mathbb{Z}/(q-1)$.

Remark 7.27. Although we know in theory that $\mathbb{F}_q^* \cong \mathbb{Z}/(q-1)$, in practice it is difficult to find a generator, or to figure out what power to raise the generator to in order to get a given element. Many cryptography and encryption protocols are based on this.

Corollary 7.28

We have $\mathbb{F}_q \cong \mathbb{F}_p(\alpha)$, and therefore, there exists an irreducible polynomial of any degree over $\mathbb{F}_p$.

Proof. There exists $\alpha \in \mathbb{F}_q$ which generates the multiplicative group; then α must generate $\mathbb{F}_q$ as an extension of $\mathbb{F}_p$, since every element of $\mathbb{F}_q$ is a *power* of α. (The converse is false — it is possible to find α which generate the extension but not the multiplicative group.)

Then $\mathbb{F}_q = \mathbb{F}_p[x]/(Q)$ where Q is the minimal polynomial of α. So Q is an irreducible polynomial of degree n, where $q = p^n$. □

§7.4.1 Application to Number Theory

Finite fields arise in many areas of math and computer science; in particular, in number theory. If R is the ring of algebraic integers in a finite extension of $\mathbb{Q}$, then we can consider $R/(p)$. For example, if $p \equiv 3 \pmod 4$ and $R = \mathbb{Z}[i]$, then $R/(p) \cong \mathbb{F}_{p^2}$.

In particular, we'll look at the extension $\mathbb{Q}(\zeta_\ell)$, where ℓ is a prime. This extension is $\mathbb{Q}[x]/(x^{\ell-1} + \cdots + 1)$, and we have

$$R = \mathbb{Z}[x]/(x^{\ell-1} + \cdots + 1).$$

So then

$$R/(p) = \mathbb{F}_p[x]/(x^{\ell-1} + \cdots + 1),$$

so its dimension over $\mathbb{F}_p$ is ℓ. So if it is a field, it must be $\mathbb{F}_{p^\ell}$.

Assume $p \neq \ell$.

Proposition 7.29

In this case, $R/(p)$ is a field iff $\operatorname{ord}_{\mathbb{F}_\ell^*} p = \ell - 1$.

Proof. We can find a maximal ideal $\mathfrak{m}$ containing (p) in R. Then we have $R/\mathfrak{m} \cong \mathbb{F}_{p^a}$ for some a. Let the image of ζ_ℓ in $R/\mathfrak{m}$ be $\overline{\zeta}_\ell$. Then we know $\overline{\zeta}_\ell^\ell = 1$; so since it lies in $\mathbb{F}_{p^a}$ (whose multiplicative group has size $p^a - 1$), we get that $\ell \mid p^a - 1$.

Now if the order of p in $\mathbb{F}_\ell^*$ is $\ell - 1$, then we must have $a \geq \ell - 1$. So then $a = \ell - 1$, and $R/\mathfrak{m} \cong R/(p)$, which means $R/(p)$ is a field. (We can prove the converse similarly.) □

Example 7.30

Consider the case where $p = 3$ and $\ell = 5$.

Solution. Then $\operatorname{ord}_5 3 = 4$, so $R/(3)$ is a field. □

§7.5 Multiple Roots

Previously, we used the fact that a multiple root of P is also a root of P' in order to show that the Artin-Schreier polynomial doesn't have multiple roots.

> **Question 7.31.** Let $P \in F[x]$ be an irreducible polynomial. Can P have multiple roots in its splitting field (or equivalently, in any extension)?

If α is such a root, then α is also a root of P', and therefore a root of $\gcd(P, P')$ as well (where $\gcd(P, P')$ is the polynomial Q which generates (P, P') as an ideal).

But P is irreducible, and $\deg P' < \deg P$. So if $P' \neq 0$, then this means $\gcd(P, P') = 1$, and no such α can exist. However, it's possible that $P' = 0$.

> **Question 7.32.** When can we have $\deg P \geq 1$ and $P' = 0$?

We have $(x^n)' = nx^{n-1}$, and if the field has characteristic 0, then this is always nonzero. Meanwhile, if the field has characteristic p, then this is zero iff $p \mid n$. So if $P' = 0$, then we must have

$$P(x) = Q(x^p) = a_n x^{pn} + a_{n-1} x^{p(n-1)} + \cdots + a_0,$$

where $p = \text{char}(F)$. So we want to see when such a polynomial is irreducible.

If $F = \mathbb{F}_q$ is finite, then we know $a^q = a$ for all $a \in F$. This means we can extract pth roots of the coefficients, since $(a^{p^{n-1}})^p = a$ — so we can write $a_i = b_i^p$ for some $b_i \in F$. Then we have

$$P = b_n^p x^{pn} + b_{n-1}^p x^{p(n-1)} + \cdots + b_0^p.$$

But this allows us to extract a pth root of the *polynomial*: we then have

$$P = (b_n x^n + b_{n-1} x^{n-1} + \cdots + b_0)^p.$$

On the other hand, there exist examples of such irreducible P in infinite fields. Take $F = \mathbb{F}_q(t)$ to be the field of rational functions over t, and $P(x) = x^p - t$. This is irreducible, but its derivative is identically 0.

> **Definition 7.33.** An extension E/F is *separable* if the minimal polynomial (over F) of every algebraic element $\alpha \in E$ has no multiple roots.

So if F has characteristic 0 or is finite, then every extension is separable. We'll only look at these instances, so we will generally assume all our extensions are separable.

§7.6 Primitive Element Theorem

> **Theorem 7.34**
> If E/F is a finite separable extension, then $E = F(\alpha)$ for some α.

This means even if we obtained E by adjoining *several* elements, we can actually obtain it by adjoining just one element.

Proof. If F is finite, then E is finite as well; so $E = F(\alpha)$ for any generator α of the multiplicative group E^*.

The more interesting case is when F is infinite. Then it is enough to prove the statement in the case where E is generated by two elements α and β — then we can induct on the number of generators to finish (since $E = F(\alpha_1, \dots, \alpha_n)$ for some n).

Suppose $E = F(\alpha, \beta)$. Now set $\gamma_t = \alpha t + \beta$. We claim that for all but finitely many values of t, we have $E = F(\gamma_t)$.

Let P be the minimal polynomial of α and Q the minimal polynomial of β, and let $K \supset E$ be a field where both P and Q both split completely. Then we can write

$$P(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n),$$
$$Q(x) = (x - \beta_1)(x - \beta_2) \cdots (x - \beta_m),$$

where the α_i are all distinct, and the β_i are all distinct (by separability); assume $\alpha_1 = \alpha$ and $\beta_1 = \beta$.

Choose t such that the mn elements $t\alpha_i + \beta_j$ are pairwise distinct (clearly there are finitely many t for which this fails — every possible equality corresponds to at most one value of t).

Now let $\gamma = \gamma_t$; it suffices to check that α and β are both in $E' = F(\gamma)$. In order to do this, we will look at polynomial equations for α over E'.

First, by definition α is a root of $P(x)$. But α is also a root of the polynomial $Q_1(x) = Q(\gamma - tx)$, which also has coefficients in E' — since plugging in α gives $Q(t\alpha + \beta - t\alpha) = Q(\beta) = 0$.

Let $S = \gcd(P, Q_1)$ (so S is the generator of the ideal (P, Q_1)), working in the ring $E'[x]$; then α is a root of $S(x)$. But S must *also* be a generator of (P, Q_1) in the larger field $K[x]$. And since our polynomials split completely, we can look at this factorization to determine S — by the condition on t, P and Q_1 have the unique common linear factor $x - \alpha$. This means $S(x) = c(x - \alpha)$ for some c. So since $S \in E'[x]$, this means $\alpha \in E'$; then $\beta = \gamma - t\alpha \in E'$ as well. So we must have $E' = E$. □

Example 7.35

Let $F = \mathbb{Q}$ and $E = \mathbb{Q}(\sqrt{3}, \sqrt[3]{7})$. Then all but finitely many linear combinations $\alpha = r\sqrt{3} + \sqrt[3]{7}$ generate E. The exceptions correspond to ratios $\frac{\beta_i - \beta_j}{\alpha_k - \alpha_\ell}$ obtained by taking different square roots of 3 and cube roots of 7; in this case, these ratios are not even real, so there are no exceptions.

§7.7 Geometry of Function Fields

Another important example of field extensions is extensions of $\mathbb{C}(t)$, and we can think of such extensions via geometry. Let $F = \mathbb{C}(t)$, and suppose $E = F[x]/(P)$ is a finite extension of F. We can assume that P is a primitive polynomial with coefficients in $\mathbb{C}[t]$ (by scaling appropriately).

We can instead think of P as a polynomial in $\mathbb{C}[t, x]$. So another way to think of these extensions is that F is the fraction field of $\mathbb{C}[t]$, and E is the fraction field of the ring $R = \mathbb{C}[t, x]/(P)$. To connect these rings to geometry, we consider the maximal spectrums

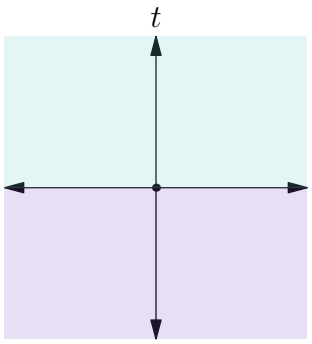
$$X = \mathrm{MSpec}(R) = \{(a, b) \in \mathbb{C}^2 \mid P(a, b) = 0\},$$
$$Y = \mathrm{MSpec}(\mathbb{C}[t]) = \mathbb{C}.$$

Then we can construct the map $X \to Y$ which sends $(a, b) \mapsto a$.

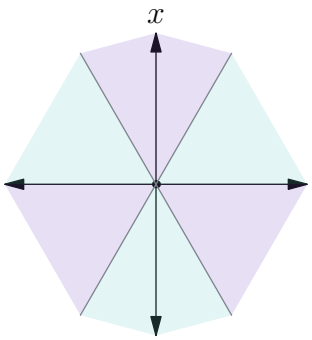
Example 7.36

Consider $P(t, x) = x^n - t$.

Solution. Then $R \cong \mathbb{C}[x]$, and the map $X \to Y$ raises a complex number to its nth power (since points in X are of the form (t, x) where $t = x^n$, and we're mapping $(t, x) \mapsto t$, so we can think of this as mapping $x \mapsto x^n$). This gives a *ramified covering* — every point in $\mathbb{C}$ except for 0 has n complex nth roots. One way to represent this geometrically is to draw the t-plane (which corresponds to Y) and the x-plane (which in this case corresponds to X). In the t-plane, we make a cut along the x-axis, turning it into two half-planes glued together.



For a point on the x-plane, raising it to the nth power multiplies the angle by n. So we cut the x-plane into $2n$ pieces (colored by which half-plane their points are mapped to):

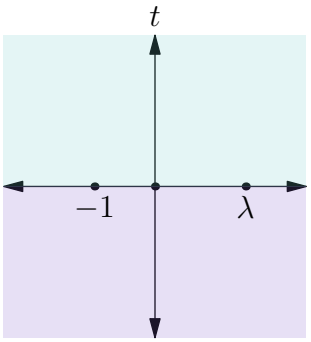


This describes the geometry of the map of raising a complex number to the nth power. □

Example 7.37

Consider $P(x, t) = x^2 - t(t + 1)(t - \lambda)$. For simplicity, assume $\lambda \in \mathbb{R}$.

Solution. We can again draw the $\mathbb{C}$-plane corresponding to t, and consider which values of x each t corresponds to. We again have a ramified double covering, with three ramification points — $t = 0$, -1, and λ (for every other point, there are two square roots). So in the t-plane (corresponding to Y), we can again make a cut and create two half-planes.



We'd now like to draw X, which consists of pairs (t, x) for which $x^2 = t(t+1)(t-\lambda)$. This is no longer a plane — in fact, it can be drawn in the shape of a bagel. To color this bagel, for each half-plane in Y, its pre-image splits into two pieces — there's two values of x corresponding to each t (depending on which branch of the square root we take). We can then glue these pieces together on the bagel, by thinking more precisely about these maps. □

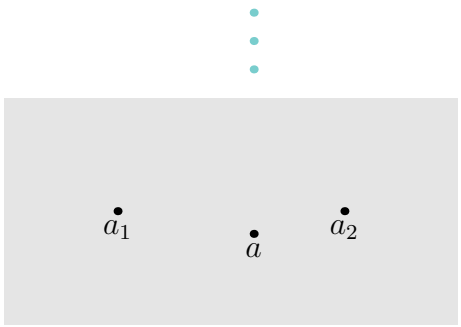
Here, we can also think of X as a *Riemann surface* (in the first example, X was simply a plane, and in the second, X was a bagel). We can then think of E as the field of rational functions on this Riemann surface.

§7.7.1 Ramified Coverings and Permutations

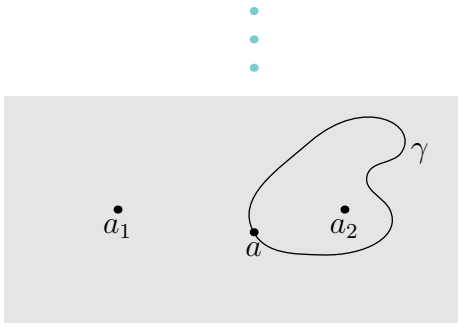
For each point of Y, we can look at the points of X which are mapped to it. In general, if we let $n = \deg(P)$, then there are exactly n points of X mapped to each point in Y — but there are finitely many exceptions. This gives a *ramified covering* of Y.

> **Remark 7.38.** It isn't a coincidence that the word *ramified* here is the same as the one used to describe the behavior of primes which factor as $\mathfrak{p}^2$ — it's the same phenomenon.

Let $a_1, \ldots, a_k$ be the ramification points (in Y), meaning that $|f^{-1}(a_i)| < n$ for all i, and $|f^{-1}(a)| = n$ for all a not equal to any of the a_i. Now take a generic point a (not equal to any a_i). Then there are n points in the pre-image of a (these points are in X):

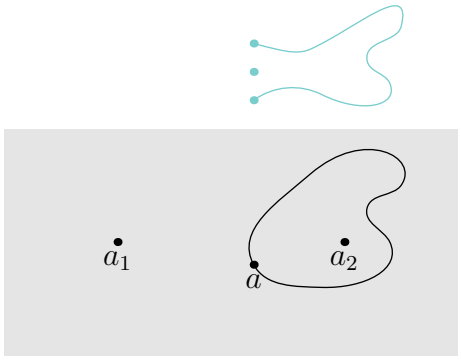


Now suppose we have a closed loop $\gamma : [0,1] \to Y$, such that $\gamma(0) = \gamma(1) = a$, and γ avoids all the ramification points of Y.



Now let's try to lift γ up to X — starting at a point x in the pre-image of a, we can *uniquely* define a continuous map $\widetilde{\gamma} : [0,1] \to X$ where $\widetilde{\gamma}(0) = x$, and $f(\widetilde{\gamma}(t)) = \gamma(t)$ for all t. (We're essentially drawing the path in X corresponding to the loop γ in Y — if we're given the starting point, there's a unique way to lift the loop continuously, since it doesn't pass through any of the ramification points.)

At the end of this lifted path, we'll still end up in the pre-image of a. But we may end up at a *different* point in this pre-image.

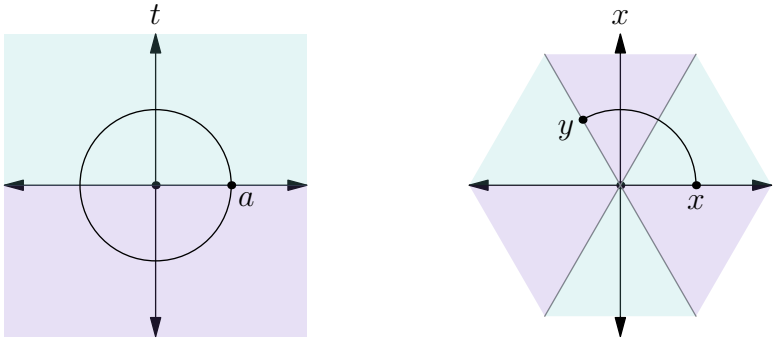


So then the loop defines a permutation σ_γ on the n points in the pre-image of a, defined by where the path starting at x ends up — if setting $\widetilde{\gamma}(0) = x$ gives $\widetilde{\gamma}(1) = y$, then we let $\sigma_\gamma(x) = y$.

Example 7.39

Consider the example $P(x,t) = x^n - t$, and let γ be the unit circle (the standard loop).

Solution. Here, both X and Y are $\mathbb{C}$, and we have $f(x) = x^n$. Suppose that $a = 1$, so the points in the pre-image are the nth roots of unity ζ_n^k, where $\zeta_n = \exp(2\pi i/n)$.



If we start at ζ_n^k in X, then we need to walk n times slower in X than in Y (since we're raising x to the nth power), so when we complete a full circle on Y, we complete $\frac{1}{n}$ of a circle on X. This means we end up at $\zeta_n^k \cdot \exp(2\pi i/n) = \zeta_n^{k+1}$.

So then σ_γ is the permutation $\zeta_n^k \mapsto \zeta_n^{k+1}$. □

So given γ, we can obtain a permutation of the pre-images of a *specific* point a. But it turns out that in certain cases, even more is true:

Theorem 7.40

If E/F is a *splitting field*, then the permutations σ_γ can be extended to an automorphism of X which fixes Y, coming from a field automorphism of E which fixes F.

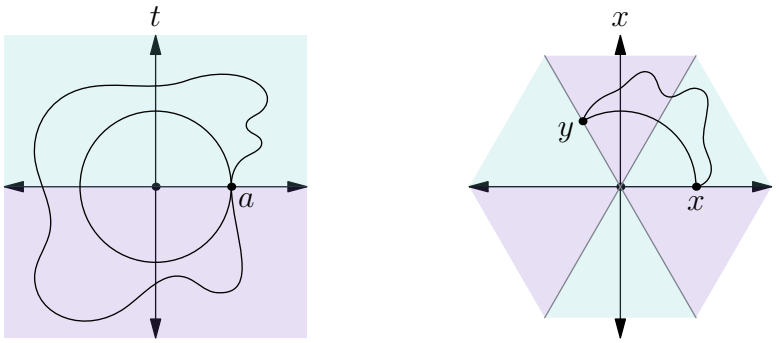
Example 7.41

In the example $P(x,t) = x^n - t$, where $E = \mathbb{C}(t)/(x^n - t)$, the automorphism corresponding to γ (the unit circle) is the map $x \mapsto \zeta_n x$. This automorphism fixes t, as $x^n = (\zeta_n x)^n$.

Example 7.42

In the example $P(x,t) = x^2 - t(t+1)(t-\lambda)$, we have a ramified double cover (with two points in the pre-image of each point, other than 0, -1, and λ). The automorphism on X we get from a loop can either be the identity, or $(t,x) \mapsto (t,-x)$.

Note that if γ is deformed continuously (with a still fixed), while avoiding the ramification points, then the permutation described doesn't change. In our example $P(x,t) = x^n - t$, any closed loop which goes around 0 exactly once will give the same permutation:



§7.8 Main Theorem of Algebra

There is a proof of the Main Theorem of Algebra using ideas similar to the ones we've seen here.

Theorem 7.43

The field $\mathbb{C}$ is algebraically closed — in other words, every nonconstant polynomial $P \in \mathbb{C}[x]$ has a root.

The proof uses the concept of a *winding number* — informally, the winding number of a loop γ (which avoids 0) is the number of times it goes around 0.

We can define winding numbers more precisely, by looking at permutations — let $f\colon \mathbb{C} \to \mathbb{C}$ be the exponential map $x \mapsto e^x$. Then for any nonzero $z \in \mathbb{C}$, its pre-image consists of the points $\{\log z + 2\pi i n \mid n \in \mathbb{Z}\}$.

Definition 7.44. Given a loop γ, the permutation σ_γ it defines (for the exponential map) sends $x \mapsto x + 2\pi i n$ for some integer n; we define n to be the *winding number* of γ, denoted as $w(\gamma)$.

Lemma 7.45

If $\gamma = \gamma_1\gamma_2$, then

$$w(\gamma) = w(\gamma_1) + w(\gamma_2).$$

Proof. This is clear, as we must have $\widetilde{\gamma}(t) = \widetilde{\gamma_1}(t) + \widetilde{\gamma_2}(t)$. □

With this definition, we can now prove the theorem.

Proof of Theorem 7.43. Let $P \in \mathbb{C}[x]$, with $P(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$, and assume for contradiction that P has no complex roots. Now for each $r > 0$, consider the loop

$$\gamma_r : t \mapsto P(re^{2\pi it})$$

(which is the image of the circle of radius r under P).

Then $w(\gamma_r)$ must be the same for all r, since we can deform all the loops to each other by gradually increasing r (we'll never pass through 0 because P has no complex roots).

But it's clear that $w(\gamma_r) = 0$ for sufficiently small r (since γ_r is a tiny loop near $P(0) \neq 0$). On the other hand, we can show that large loops have nonzero winding number:

Claim — If r is sufficiently large, then $w(\gamma_r) = n$.

Proof. We can write

$$P(x) = x^n \left(1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n}\right).$$

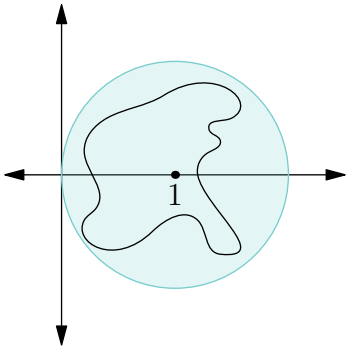
This means we can split $\gamma_r = \gamma_1\gamma_2$, where

$$\gamma_1 = r^n \cdot e^{2\pi int},$$
$$\gamma_2 = 1 + \frac{a_{n-1}}{re^{2\pi it}} + \cdots + \frac{a_0}{r^n e^{2\pi int}}.$$

It's clear that $w(\gamma_1) = n$. On the other hand, γ_2 is trapped in a small circle around 1: we have

$$|\gamma_2(t) - 1| \leq \frac{a_{n-1}}{r} + \cdots + \frac{a_0}{r^n},$$

and if r is sufficiently large, this is much less than 1.



But this means γ_2 can't go around 0 at all, so $w(\gamma_2) = 0$. Then $w(\gamma_r) = w(\gamma_1) + w(\gamma_2) = n$, as desired. ■

So then $w(\gamma_r)$ is *not* constant over all r, contradiction. □

Remark 7.46. One intuitive way to visualize the proof that $w(\gamma_r) = n$ for large r is that since $P(x)$ is approximately x^n for large $|x|$, then γ_r is approximately the circle of radius r (winding around the origin n times). We can imagine a person walking along the circle of radius r, with a dog on a leash (the dog's movement corresponds to the remaining terms $a_{n-1}x^{n-1} + \cdots + a_0$). As long as the leash is short enough, no matter how the dog moves, it must walk around the center of the circle the same number of times as the person.

§8 Galois Theory

§8.1 The Galois Group

Our main object of study is the Galois group:

Definition 8.1. The *Galois group* of an extension E/F, denoted $\mathrm{Gal}(E/F)$, is the group of automorphisms of E which are the identity on F.

Example 8.2

The Galois group $\mathrm{Gal}(\mathbb{C}/\mathbb{R})$ consists of two elements — the identity and complex conjugation.

The Galois group can store a lot of information about the structure of the field extension. But it only works well for *some* classes of extensions — more precisely, for splitting fields.

Theorem 8.3

Suppose that E/F is a splitting field of some polynomial. Then for *any* $\alpha \in E$, the minimal polynomial of α must split completely in E.

Example 8.4

Take $F = \mathbb{Q}$, and E to be the splitting field of $x^5 - 2$; then $E = \mathbb{Q}(\sqrt[5]{2}, \zeta_5)$. Let $\alpha = \sqrt[5]{2} + \zeta_5$, which generates E by the Primitive Element Theorem; then the minimal polynomial of α over $\mathbb{Q}$ has degree 20 (we have $[E : Q] = 20$, since $\sqrt[5]{2}$ has degree 5 over $\mathbb{Q}$ and ζ_5 has degree 4).

The theorem then states that all 20 complex roots of this minimal polynomial are inside E. We can actually explicitly describe these roots — they're of the form $\sqrt[5]{2}\zeta_5^i + \zeta_5^j$ for some integers $0 \le i \le 4$ and $1 \le j \le 4$, which are indeed in E.

Proof of Theorem 8.3. The idea is to use the fact that any given polynomial has a unique splitting field.

Suppose E is the splitting field of some polynomial Q, and fix $\alpha \in E$ with minimal polynomial P; then we want to show that P splits completely in E.

Let $K \supset E$ be the splitting field for P over E. Then in K, we have

$$P(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n),$$

where $\alpha_i \in K$ for all i and $\alpha_1 = \alpha$. It suffices to check that $\alpha_i \in E$ for all i (and we already know $\alpha \in E$).

But we have that $F(\alpha_i) \cong F(\alpha)$, since α_i and α have the same minimal polynomial. Now consider the perspective of these two intermediate fields. We know that E is the splitting field for Q over $F(\alpha)$. Meanwhile, $E(\alpha_i)$ is the splitting field for Q over $F(\alpha_i)$ — this is clear from the definition of the splitting field (in general, if $E = F(\beta_1, \dots, \beta_m)$ is the splitting field of Q (where the β_i are the roots of Q), then $E(\alpha_i) = F(\alpha_i, \beta_1, \dots, \beta_m)$ is the splitting field of Q over $F(\alpha_i)$).

But by the uniqueness of the splitting field, we can then extend the isomorphism between $F(\alpha)$ and $F(\alpha_i)$ to an isomorhpism between E and $E(\alpha_i)$ (since the isomorphism between $F(\alpha)$ and $F(\alpha_i)$ is the identity on F, which means Q is the same polynomial in both fields). This means

$$[E : F] = [E(\alpha_i) : F],$$

and since $E(\alpha_i) \supset E$, this means we must have $E(\alpha_i) = E$, and therefore $\alpha_i \in E$. □

Proposition 8.5

For any finite (separable) extension E/F, we have

$$|\mathrm{Gal}(E/F)| \leq [E : F],$$

with equality iff E is the splitting field of some polynomial.

Proof. Using the Primitive Element Theorem, we can let $E = F(\alpha)$ for some α. Then

$$[E : F] = \deg(\alpha) = \deg P,$$

where P is the minimal polynomial of α.

Meanwhile, an automorphism $\sigma : E \to E$ which fixes F is uniquely determined by $\sigma(\alpha)$ (since α generates the extension), so it suffices to find the number of possible choices for $\sigma(\alpha)$. But $\sigma(\alpha)$ can be any root of the minimal polynomial of α; so $|\mathrm{Gal}(E/F)|$ is equal to the number of roots of P in E.

The number of roots of P in E is at most $\deg P$, which immediately proves

$$|\mathrm{Gal}(E/F)| \leq [E : F].$$

If equality holds, then P must split completely in E; this immediately implies that E is the splitting field of P (since E is also generated by a root α of P).

On the other hand, if E is a splitting field, then by Theorem 8.3, P must split completely in E. Since P cannot have multiple roots, this means it has exactly $\deg P$ roots, and therefore there are exactly $\deg P$ automorphisms. □

Remark 8.6. It's possible to show that $|\mathrm{Gal}(E/F)| \leq [E : F]$ without using the Primitive Element Theorem (by induction on the number of generators), so it's true even without separability. But we do need separability to show that if E is a splitting field, then equality holds.

Definition 8.7. A finite extension E/F is *Galois* if $[E : F] = |\mathrm{Gal}(E/F)|$.

The main theorem we will discuss is the following:

Theorem 8.8

If E/F is a Galois extension with Galois group $\mathrm{Gal}(E/F)$, then there is a bijection between subgroups of G, and intermediate subfields $F \subseteq K \subseteq E$ — where a subgroup $H \subset G$ is mapped to its *fixed field*

$$K = E^H = \{x \in E \mid \sigma(x) = x \text{ for all } \sigma \in H\},$$

and a subfield K is mapped to the subgroup of $\sigma \in G$ which fix all elements of K (which is $\mathrm{Gal}(E/K)$).

In this bijection, note that if $H \mapsto K_H$, then $|H| = [E : K_H]$ — this follows immediately from the fact that $H = \mathrm{Gal}(E/K_H)$, and Proposition 8.5.

Remark 8.9. Any finite group is the Galois group of some extension — we'll show later that S_n is the Galois group of some extension, and any finite group is a subgroup of S_n for some n. However, whether any group can be a Galois group of an extension over $\mathbb{Q}$ is still open.

We will discuss the proof and some applications later; first, we will discuss how to compute the Galois group in a few examples.

§8.2 Examples of Galois Groups

For a polynomial P with splitting field E (over F), we use $\mathrm{Gal}(P)$ to refer to $\mathrm{Gal}(E/F)$.

Proposition 8.10

The Galois group $\mathrm{Gal}(P)$ acts on the roots of P; and if P is irreducible, this action is transitive.

Proof. We've already seen that elements of the Galois group must permute the roots of P (since F is fixed, $\sigma(\alpha)$ must satisfy any polynomial equation over F that α satisfies). Now if P is irreducible, we can write $P(x) = (x - \alpha_1) \cdots (x - \alpha_n)$; then to show transitivity, it suffices to show that for any i and j, there exists $\sigma \in \mathrm{Gal}(P)$ such that σ sends $\alpha_i \mapsto \alpha_j$.

But we know $F(\alpha_i) \cong F(\alpha_j)$ (since α_i and α_j have the same minimal polynomial). Further, E is the splitting field of P over both $F(\alpha_i)$ and $F(\alpha_j)$. So by the uniqueness of the splitting field, the isomorphism between $F(\alpha_i)$ and $F(\alpha_j)$ extends to an isomorphism $\sigma : E \to E$ whih sends $\alpha_i \mapsto \alpha_j$. □

Remark 8.11. This argument is fairly similar to the one used in the proof of Theorem 8.3.

However, we can't say much more than this in general — even knowing that $\mathrm{Gal}(P)$ acts transitively on a set of n elements, its size can be anywhere from n to $n!$.

Example 8.12

Compute the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$, where $\zeta_n = \exp(\frac{2\pi i}{n})$. For simplicity assume $n = p$ is prime (although a similar argument is still true in the general case).

Solution. Let $\zeta = \zeta_p$. We proved earlier that $[\mathbb{Q}(\zeta) : \mathbb{Q}] = p - 1$, and $\mathbb{Q}(\zeta)$ is the splitting field of the irreducible polynomial $x^{p-1} + x^{p-2} + \cdots + 1$.

Any automorphism $\sigma \in \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ must send $\zeta \mapsto \zeta^i$ for some $1 \le i \le p - 1$; this uniquely determines the automorphism, which we denote by σ_i.

In order to compute the group, it suffices to understand how these automorphisms compose — we have

$$\sigma_i\sigma_j(\zeta) = \zeta^{ij},$$

which means $\sigma_i\sigma_j = \sigma_{ij}$. So in this case, we have

$$\mathrm{Gal}(\mathbb{Q}(\zeta)/\mathbb{Q}) = (\mathbb{Z}/p)^\times \cong \mathbb{Z}/(p-1). \qquad \square$$

In this case, we were lucky because all roots of the polynomial were powers of one root ζ. In general, after fixing one root and sending it to another root, we still need to figure out what we can do with the remaining roots (which depends on the algebraic relations between the roots).

> **Remark 8.13.** This fact can be used to solve the problem of which regular polygons can be constructed by a ruler and compass; we will see this proof later.

So this is an example of a case where the Galois group is as small as possible. There are also examples where the Galois group is as *big* as possible:

Example 8.14

Let P be an irreducible polynomial (over $\mathbb{Q}$) of degree p, with exactly $p-2$ real roots and 2 roots which are complex conjugates (where p is prime). Then if E is the splitting field of P,

$$\mathrm{Gal}(E/\mathbb{Q}) = S_p.$$

Proof. The proof uses the following group-theoretic lemma:

Lemma 8.15

Suppose p is prime, and G is a subgroup of S_p such that G acts transitively on $[1,\dots,p]$ and G contains a transposition. Then $G = S_p$.

Proof. Since G acts transitively on $[1,\dots,p]$, it follows that $p \mid |G|$ (this is because if G acts transitively on X, then $|G| = |X| \cdot |\mathrm{Stab}_G(x)|$ for any $x \in X$, by splitting the elements of G by where they send x). So by the Sylow Theorems, G has an element σ of order p. The only elements of S_p with order p are long cycles (meaning cycles of p elements); so σ is a long cycle.

WLOG the transposition G contains is (12). Then there exists $1 \le i < p$ such that σ^i sends $1 \mapsto 2$ (by starting at 1, and following the arrows in σ until we reach 2); and $\gamma = \sigma^i$ is still a long cycle (since i is relatively prime to p). So we can WLOG assume $\gamma = (123\cdots p)$.

But now we can conjugate to produce more transpositions — we have $\gamma(12)\gamma^{-1} = (23)$, $\gamma^2(12)\gamma^{-2} = (34)$, and so on. So G contains all the standard transpositions (12), (23), (34), ... These transpositions generate S_p, so $G = S_p$. $\square$

In our setting, we know $\mathrm{Gal}(E/\mathbb{Q})$ acts by permutations on the p roots. Since P is irreducible, we know this action is transitive (we can send any root to any other root, since the abstract procedure of adjoining a root is equivalent to adjoining any specific root).

Meanwhile, complex conjugation is a transposition — first, to show it's an element of the Galois group, it must permute the roots of P; therefore the splitting field $E = \mathbb{Q}(\alpha_1,\dots,\alpha_p)$ is invariant under complex conjugation. Then it's clearly a field automorphism which fixes $\mathbb{Q}$, so it's an element of the Galois group. Then it swaps the two complex roots and fixes all the real roots, so it corresponds to a transposition.

So $\mathrm{Gal}(E/\mathbb{Q})$ acts transitively and contains a transposition; by the lemma, this means it's S_p. $\square$

Remark 8.16. This particular trick of showing the Galois group is S_n relies on the specific conditions given. But given a random polynomial (even if it doesn't satisfy those conditions), its Galois group is *probably* S_n — in order for the Galois group to *not* be S_n, we'd need conditions on the roots.

Remark 8.17. A similar argument can be used to produce an extension of $\mathbb{C}(t)$ with Galois group S_p. The analog of the condition on real roots is that $P(t,x) \in \mathbb{C}(t)[x]$ should have exactly one ramification point, which should be *simple* (meaning that the point has $p-1$ pre-images, and the map f is isomorphic on all but one of them and looks locally like the square map on the last). Then if E is the splitting field of P, we have $\mathrm{Gal}(E/\mathbb{C}(t)) = S_p$ as well, by a similar argument (the condition on ramification is now used to produce a transposition).

§8.3 The Main Theorem

Earlier, we stated the main theorem:

Theorem 8.18

If E/F is a Galois extension with Galois group $\mathrm{Gal}(E/F)$, then there is a bijection between subgroups of G, and intermediate subfields $F \subseteq K \subseteq E$ — where a subgroup $H \subset G$ is mapped to its *fixed field*

$$K = E^H = \{x \in E \mid \sigma(x) = x \text{ for all } \sigma \in H\},$$

and a subfield K is mapped to the subgroup of $\sigma \in G$ which fix all elements of K (which is $\mathrm{Gal}(E/K)$).

In order to prove this map is a bijection, we'll first prove the following equality:

Lemma 8.19

In both correspondences, we have $[E:K] = |H|$.

Proof. For the second correspondence, this is clear — if $K \mapsto H$, then $H = \mathrm{Gal}(E/K)$ by definition. But E/K is a Galois extension (if E is the splitting field of some polynomial over F, it's also the splitting field of the same polynomial over K), so we have $|H| = |\mathrm{Gal}(E/K)| = [E:K]$.

Now we'll prove the equality for the first correspondence — suppose $H \mapsto E^H$. Then it's clear from the definition that $H \subset \mathrm{Gal}(E/E^H)$, so

$$|H| \leq \left|\mathrm{Gal}(E/E^H)\right| = [E:E^H]$$

(since E/E^H is a Galois extension for the same reason as E/K was earlier). So it suffices to show the other direction of this inequality.

By the Primitive Element Theorem, we know $E = E^H(\alpha)$ for some α; so it suffices to check that α is a root of some polynomial in $E^H[x]$ of degree $|H|$.

To construct such a polynomial, we use the averaging trick — let

$$P(x) = \prod_{g \in H} (x - g(\alpha)).$$

(For example, if $E = \mathbb{C}$ and H consists of the identity and complex conjugation, then $P(x) = (x-z)(x-\overline{z})$ is a familiar polynomial.) *A priori*, we only know $P(x) \in E[x]$. But it's actually in $E^H[x]$ — to show this, it

suffices to show that for every $h \in H$, the coefficients of P are preserved by h. Let $P(x) = x^n + a_{n-1}x^{n-1} + a_0$; then we have

$$x^n + h(a_{n-1})x^{n-1} + \cdots + h(a_0) = \prod_{g \in H} (x - h(g(\alpha))) = P(x),$$

as desired (where the first equality follows from the fact that h is an automorphism, and the second from the fact that multiplication by any $h \in H$ permutes the elements of H).

Clearly, α is a root of $P(x)$ (since $1 \in H$), so then α is the root of a polynomial in $E^H[x]$ of degree $|H|$, which means $[E : E^H] \leq |H|$, as desired. □

Once we have this equality, the remainder of the proof is mostly just formal:

Proof of Theorem 8.18. To show that both maps are bijections, it suffices to show that one is the inverse of the other.

First, if we have $H \mapsto E^H \mapsto H'$, then by definition $H \subseteq H'$ (since H' consists of all elements E^H, but we know by the definition of E^H that all elements of H fix it). But we must have $|H| = |H'|$, so then $H = H'$.

Similarly, if we have $K \mapsto H \mapsto K'$, then we know $K \subset K'$ (since K' consists of the field fixed by H, but H was defined so that K is fixed by it). But $[E : K] = [E : K']$, so then $K = K'$. □

Remark 8.20. Note that this correspondence *reverses* inclusion — a larger subgroup corresponds to a smaller fixed field.

§8.4 Intermediate Extensions

Consider a Galois extension E/F, and let K be an intermediate subfield. Then E/K is a Galois extension (if E is the splitting field of P over F, it is also the splitting field of P over K); we've seen that $[E : K] = |\mathrm{Gal}(E/K)|$. On the other hand, the extension K/F may or may not be Galois.

Proposition 8.21

Let E/F be a Galois extension, and let $G = \mathrm{Gal}(E/F)$. For a field $E \supset K \supset F$, the extension K/F is Galois iff K is invariant under all $g \in G$, which occurs iff the subgroup H corresponding to K is normal. In that case, we have $\mathrm{Gal}(K/F) = G/H$.

Note that invariance means that for all $g \in G$, we have $x \in K$ iff $g(x) \in K$; or in other words, $g(K) = K$ for all $g \in G$.

Proof. First we'll show that K/F is Galois iff K is invariant under $\mathrm{Gal}(E/F)$. On one hand, if K is Galois, then it's the splitting field of some polynomial; so we have $K = F(\alpha_1, \ldots, \alpha_n)$ where $\alpha_1, \ldots, \alpha_n$ are the roots of a polynomial. But any automorphism of E preserving F must permute the roots of any polynomial over F; and therefore any such automorphism must preserve K.

On the other hand, if K is invariant under all $g \in G$, then there is a homomorphism $\mathrm{Gal}(E/F) \to \mathrm{Gal}(K/F)$, obtained simply by restricting every automorphism over E in the Galois group to an automorphism over K. The kernel of this homomorphism is $\mathrm{Gal}(E/K)$, so by the homomorphism theorem, its image has cardinality

$$\frac{|\mathrm{Gal}(E/F)|}{|\mathrm{Gal}(E/K)|} = \frac{[E : F]}{[E : K]} = [K : F].$$

But the image can have cardinality *at most* $|\mathrm{Gal}(K/F)|$, which is necessarily at most $[K : F]$; therefore equality must hold, and K/F is Galois.

Now we'll show that K is invariant under all $g \in G$ iff H is normal. Note that an automorphism γ fixes an element $x \in K$ iff $g\gamma g^{-1}$ fixes $g(x)$; so for any $g \in G$, the subgroup corresponding to $g(K)$ is exactly gHg^{-1}. In particular, $g(K) = K$ for all $g \in G$ iff $gHg^{-1} = H$ for all $g \in G$, meaning H is normal.

Finally, to show that $\mathrm{Gal}(K/F) = G/H$ if H is normal, again consider the homomorphism $G \to \mathrm{Gal}(K/F)$ where we restrict each automorphism to K. We've shown that this homomorphism is surjective, and its kernel is $\mathrm{Gal}(E/K) = H$; so then $\mathrm{Gal}(K/F) = G/H$. □

Remark 8.22. In some sense, the proposition states that $\mathrm{Gal}(K/F)$ makes sense iff G/H does; and in that case, they're the same.

§8.5 Some Applications and Examples

We'll now consider two important examples of Galois extensions.

§8.5.1 Cyclotomic Extensions

It's possible to use Theorem 8.18 to finally answer the remaining direction of our question about compass and straightedge constructions:

Proposition 8.23

If $p = 2^k + 1$ is a Fermat prime, then a regular p-gon can be constructed by a compass and straightedge.

Proof. Let ζ be a primitive pth root of unity; then it suffices to show that $\mathbb{Q}(\zeta)$ can be obtained by iteratively taking quadratic extensions — in other words, that there exists a tower of field extensions

$$\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_n = \mathbb{Q}(\zeta)$$

such that $[F_i : F_{i-1}] = 2$ for each i. (This is because quadratic extensions can always be obtained by extracting a square root; so this condition means we can obtain all elements of $\mathbb{Q}(\zeta)$ through arithmetic operations and taking square roots.)

But this is fairly clear from the Galois correspondence — we know that

$$\mathrm{Gal}(\mathbb{Q}(\zeta)/\mathbb{Q}) = (\mathbb{Z}/p)^\times \cong \mathbb{Z}/(p-1) = \mathbb{Z}/2^k.$$

Take the chain of subgroups $G_0 = \mathbb{Z}/2^k\mathbb{Z}$, $G_1 = 2\mathbb{Z}/2^k\mathbb{Z}$, $G_2 = 4\mathbb{Z}/2^k\mathbb{Z}$, $\ldots$, $G_k = 2^k\mathbb{Z}/2^k\mathbb{Z}$ — then $G_i/G_{i-1} \cong \mathbb{Z}/2\mathbb{Z}$ for all i. Then let F_i be the fixed field of G_i for all i. We've seen that the correspondence reverses inclusion, while we have $[\mathbb{Q}(\zeta) : F_i] = 2^i$ for all i, which means $[F_i : F_{i-1}] = 2$ for all i. □

We can make this idea a bit more concrete:

Example 8.24

Describe how to compute F_1, the quadratic extension of $\mathbb{Q}$ contained in $\mathbb{Q}(\zeta)$.

Solution. In the above construction, we took F_1 to be the fixed field of G_1, which is the subgroup of $(\mathbb{Z}/p)^\times$ consisting of the quadratic residues.

Consider the element

$$\alpha = \sum_{a \in \mathrm{QR}} \zeta^a,$$

which is invariant under all automorphisms in G_1 (since such automorphisms multiply all exponents by a quadratic residue).

In order to find its Galois conjugate, we can apply an automorphism not in G_1; this gives

$$\beta = \sum_{b \in \mathrm{NQR}} \zeta^b.$$

(In the case of $p = 5$, we have $\alpha = \zeta + \zeta^4$ and $\beta = \zeta^2 + \zeta^3$.)

We now want to compute the quadratic equation satisfied by α and β. We know $\alpha + \beta = -1$. Meanwhile, we can write

$$\alpha\beta = \sum_c n_c \zeta^c,$$

where the coefficients n_c are the number of solutions (a, b) to $c = a + b$ such that a is a quadratic residue and b a quadratic nonresidue.

Claim — We have $n_0 = 0$, while $n_1 = n_2 = \cdots = n_{p-1}$.

Proof. First, since -1 is a quadratic residue mod p, there are no solutions to $0 = a + b$ where a is a quadratic residue and b is not, so $n_0 = 0$.

On the other hand, for any nonzero c and c', there is a bijection between solutions with $a + b = c$ and solutions with $a + b = c'$ — suppose $c' = tc$. If t is a quadratic residue, then for any solution (a, b) for c, the pair (ta, tb) is a solution for c'. Meanwhile, if t is a quadratic nonresidue, then for any solution (a, b) for c, the pair (tb, ta) is a solution for c'. □

Meanwhile, $n_0 + \cdots + n_{p-1}$ is the total number of possible pairs (a, b), which is $(\frac{p-1}{2})^2$. So then $n_c = \frac{p-1}{4}$ for all nonzero c, which means

$$\alpha + \beta = \frac{p-1}{4} \sum \zeta^c = -\frac{p-1}{4}.$$

This gives the quadratic equation

$$\alpha^2 + \alpha - \frac{p-1}{4} = 0.$$

Using the quadratic formula gives $F_1 = \mathbb{Q}(\sqrt{p})$. □

Remark 8.25. The same argument for computing the quadratic extension of $\mathbb{Q}$ contained in $\mathbb{Q}(\zeta_p)$ works for *any* $p \equiv 1 \pmod 4$, not just Fermat primes. When $p \equiv 3 \pmod 4$, the argument works very similarly (except that n_0 is now $\frac{p-1}{2}$ instead of 0, since -1 is not a square), and gives that the extension is $\mathbb{Q}(\sqrt{-p})$.

The description of $\mathrm{Gal}(\mathbb{Q}(\zeta)/\mathbb{Q})$ can be generalized to the case where ζ is a primitive nth root of unity for *any* positive integer n (not necessarily prime).

Definition 8.26. The nth *cyclotomic polynomial* Φ_n is the monic polynomial in $\mathbb{Z}[x]$ whose roots are exactly the primitive nth roots of unity.

We then have

$$x^n - 1 = \prod_{d|n} \Phi_d(x),$$

by doing casework on the order of each nth root of unity (in $\mathbb{C}^\times$), which must divide n. This formula makes it clear that $\Phi_n \in \mathbb{Z}[x]$ (by induction), and lets us compute Φ_n for a given n.

Example 8.27

A few examples of Φ_n are

$$\begin{aligned}\Phi_1(x) &= x - 1,\\ \Phi_p(x) &= x^{p-1} + x^{p-2} + \cdots + 1,\\ \Phi_{12}(x) &= x^4 - x^2 + 1.\end{aligned}$$

Remark 8.28. It isn't true that for every n, the coefficients of Φ_n are always 0 or ± 1; but the first counterexample is $n = 105$.

Fact 8.29 — For all n, Φ_n is irreducible in $\mathbb{Q}[x]$.

We've already proved this fact for n prime; the proof is somewhat longer in general.

Note that $\deg(\Phi_n)$ is the number of elements of order n in $\mathbb{Z}/n\mathbb{Z}$, which is

$$|(\mathbb{Z}/n\mathbb{Z})^\times| = \varphi(n) = \prod (p_i^{d_i} - p_i^{d_i - 1}),$$

where $n = \prod p_i^{d_i}$. Since Φ_n is irreducible, we then have

$$[\mathbb{Q}(\zeta) : \mathbb{Q}] = \varphi(n).$$

Meanwhile, we know that $\mathbb{Q}(\zeta)$ is a splitting field, for the same reason as in the case where n was prime (it's still the splitting field of Φ_n).

In this case (for the same reason as in the prime case), we have

$$\mathrm{Gal}(\mathbb{Q}(\zeta)/\mathbb{Q}) = (\mathbb{Z}/n\mathbb{Z})^\times.$$

Note that this group is abelian (and therefore a product of cyclic groups), but is not necessarily cyclic — in fact, it's not cyclic unless n is 2, 4, or p^k or $2p^k$ for an odd prime p.

§8.5.2 Kummer Extensions

Definition 8.30. If F is a field containing a primitive nth root of unity (equivalently, if F contains exactly n elements with $x^n = 1$), then a *Kummer extension* E is an extension $E = F(\alpha)$ for some α such that $\alpha^n = a$ is a nonzero element of F.

Intuitively, we start with a field that already contains all nth roots of unity, and then attach a nth root of some other element. If $\mathrm{char}(F) = p$, then we need the additional assumption that $p \nmid n$; but our main examples will be in characteristic 0.

Proposition 8.31

Any Kummer extension E/F is Galois, and

$$\mathrm{Gal}(E/F) \cong \mathbb{Z}/m\mathbb{Z}$$

for some $m \mid n$. In fact, if $x^n - a$ is irreducible in $F[x]$, then $m = n$.

Proof. First, we have

$$x^n - a = \prod_{i=0}^{n-1} (x - \zeta^i \alpha),$$

where ζ is a primitive nth root of unity. Since $\zeta \in F$, this means E is the splitting field of $x^n - a$.

Now any $\sigma \in \mathrm{Gal}(E/F)$ is determined by $\sigma(\alpha)$, and we must have $\sigma(\alpha) = \zeta^i \alpha$ for some i; let σ_i be the automorphism mapping $\alpha \mapsto \zeta^i \alpha$, if such an automorphism is in the Galois group.

Then we have

$$\sigma_i \sigma_j(\alpha) = \sigma_i(\zeta^j \alpha) = \zeta^{i+j} \alpha$$

(since $\zeta \in F$, all σ must fix ζ). This means $\mathrm{Gal}(E/F)$ is isomorphic to a subgroup of $\mathbb{Z}/n\mathbb{Z}$, and every such subgroup is of the form $\mathbb{Z}/m\mathbb{Z}$ for some $m \mid n$.

We then have $m = \deg(E/F)$; this means $m = n$ iff $x^n - a$ is irreducible. □

§8.6 Solutions to Polynomial Equations

Galois theory has applications to solving polynomials — in particular it was motivated by proving the impossibility of a solution in radicals to a general polynomial equation of degree at least 5.

§8.6.1 Impossibility of Solving Quintics

In order to prove that it's impossible to solve a general quintic (or polynomial of higher degree) in radicals, we'll first introduce the notion of a solvable group. The main idea is to show that any polynomial which can be solved in radicals must have a solvable Galois group; while S_n is not solvable.

Definition 8.32. A finite group G is *solvable* if there exists a sequence of subgroups

$$G = G_0 \supset G_1 \supset G_2 \supset \cdots \supset G_n = \{1\},$$

such that for each i, G_i is a normal subgroup of G_{i-1} and G_{i-1}/G_i is abelian.

Lemma 8.33

If $G/K \cong H$ (or equivalently, if there is a surjective map $G \twoheadrightarrow H$ with kernel K), then:

1. If K and H are solvable, then G is solvable.
2. If G is solvable, then H is solvable.

Proof. For the first direction, start with the filtration $H = H_0 \supset \cdots \supset H_n = \{1\}$ for H. Then take G_i to be the pre-image of H_i (under the homomorphism), so we have $G = G_0 \supset \cdots \supset G_n = K$. Finally, we can append the filtration $K = K_0 \supset \cdots \supset K_m = \{1\}$ to the end in order to get a filtration of G.

For the second direction, given a filtration $G = G_0 \supset G_1 \supset \cdots \supset G_n = \{1\}$ of G, let H_i be the image of G_i. Then H_{i-1}/H_i is a quotient of G_{i-1}/G_i for all i, and a quotient of any abelian group is abelian. □

Lemma 8.34

S_5 is not solvable; in fact, A_5 is simple.

This lemma is true for all $n \geq 5$ — it's not true for $n = 3$ since $A_3 \cong \mathbb{Z}/3\mathbb{Z}$ is abelian, and it's not true for $n = 4$ since the Klein 4-group K_4 is a normal subgroup of S_4 (consisting of the elements (12)(34), (13)(24), (14)(23), and the identity). It's possible to prove it in the general case by analyzing conjugacy classes, and proving that if we have all the elements of one conjugacy class, we can generate elements from any other conjugacy class. However, for the sake of time, we'll here present a simpler proof that only works for $n = 5$.

Proof. The class equation for A_5 is

$$60 = 1 + 15 + 20 + 12 + 12$$

(corresponding to the conjugacy classes of the identity, (12)(34), (123), (12345), and (13245), respectively).

Then if N is a normal subgroup, it must be a union of conjugacy classes; furthermore, for any 5-cycle in one of the conjugacy classes of size 12, its square is in the other; so N must contain both or neither of the two classes.

Now this means

$$|N| = 1 + \varepsilon_1 \cdot 15 + \varepsilon_2 \cdot 20 + \varepsilon_3 \cdot 24$$

(where the ε_i are all 0 or 1). But we also know $|N| \mid 60$. If N contains some conjugacy class other than the one of the identity, then we must have $\varepsilon_1 = 1$ (or else $|N|$ would be odd, and therefore would need to divide 30, which is impossible); then we must have $\varepsilon_2 = 1$ (or else $|N|$ would not be divisible by 3, and therefore would need to divide 20); then we must have $\varepsilon_3 = 1$ as well (or else $|N|$ would not be divisible by 5). This means $N = A_5$. □

Now we will apply the concept of solvability to the Galois group of our polynomial.

Definition 8.35. A finite extension E/F is a *radical extension* if $E = F(\alpha_1, \ldots, \alpha_n)$, where there exist positive integers n_i such that $\alpha_i^{n_i} \in F(\alpha_1, \ldots, \alpha_{i-1})$ for all i.

In other words, a finite extension is radical if each of its elements can be obtained by starting with F and performing the arithmetic operations (addition, subtraction, multiplication, and division) and extracting radicals.

Example 8.36

The field

$$\mathbb{Q}\left(\sqrt[3]{3 + \sqrt[5]{7 + \sqrt{2}}}\right)$$

is a radical extension of $\mathbb{Q}$.

Proposition 8.37

Every radical extension of F is contained in a Galois extension with a solvable Galois group.

In this proof, we'll assume that $\mathrm{char}(F) = 0$ (although it is possible to generalize this).

For the proof, it will be convenient to generalize Proposition 8.31 to allow us to simultaneously extract multiple nth roots:

Lemma 8.38

Let F be a field containing a primitive nth root of unity. Then if $E = F(\beta_1, \dots, \beta_k)$, where $\beta_i^n \in F$ for all i, then

$$\mathrm{Gal}(E/F) \subset (\mathbb{Z}/n\mathbb{Z})^k.$$

In particular, $\mathrm{Gal}(E/F)$ is abelian.

Proof. The proof is essentially the same as before — each automorphism must map $\beta_i \mapsto \zeta^{j_i}\beta_i$ for some j_1, $\dots$, j_k; and composing automorphisms corresponds to componentwise addition. □

Proof of Proposition 8.37. We use induction on n.

For the base case $n = 1$, we have $E \subset F(\zeta, \alpha)$ where ζ is a nth root of unity, and $\alpha^k \in F$ for some k. This extension is Galois, as it is the splitting field of $x^k - \alpha^k$.

Now consider the tower of field extensions $F(\zeta, \alpha)/F(\zeta)/F$. We know that both extensions are Galois, and $\mathrm{Gal}(F(\zeta, \alpha)/F(\zeta))$ is a subgroup of $\mathbb{Z}/k\mathbb{Z}$, while $\mathrm{Gal}(F(\zeta)/F)$ is a subgroup of $(\mathbb{Z}/k\mathbb{Z})^\times$. Both are abelian, so $\mathrm{Gal}(F(\zeta, \alpha)/F)$ is solvable, by taking the chain of subgroups

$$\mathrm{Gal}(F(\zeta, \alpha)/F) \supset \mathrm{Gal}(F(\zeta, \alpha)/F(\zeta)) \supset \{1\}$$

and using the fact that

$$\mathrm{Gal}(F(\zeta, \alpha)/F)/\mathrm{Gal}(F(\zeta, \alpha)/F(\zeta)) \cong \mathrm{Gal}(F(\zeta)/F).$$

Now for the inductive step, suppose we know that $F(\alpha_1, \dots, \alpha_{n-1}) \subset E'$, where E'/F is a Galois extension and $\mathrm{Gal}(E'/F)$ is solvable; and suppose that $\alpha^k \in F(\alpha_1, \dots, \alpha_{n-1})$ (where $\alpha = \alpha_n$).

Let ζ be a primitive kth root of unity; then we want to add ζ to E', as well as α. However, if we just added α, it's possible that the resulting extension wouldn't be Galois (as the polynomial $x^k - \alpha^k$ is necessarily in $F(\alpha_1, \dots, \alpha_{n-1})$, but is not necessarily in F). To avoid this issue, we add all the Galois conjugates of α as well — let $\beta_1, \dots, \beta_d$ be the Galois conjugates of α^k, meaning the images of α^k under all automorphisms in $\mathrm{Gal}(E'/F)$. Then we take

$$E = E'\left(\zeta, \sqrt[k]{\beta_1}, \sqrt[k]{\beta_2}, \dots, \sqrt[k]{\beta_d}\right).$$

Now if E' is the splitting field of a polynomial Q over F, then E is the splitting field of the polynomial

$$Q(x) \cdot (x^k - 1) \cdot \prod_{g \in \mathrm{Gal}(E'/F)} (x^k - g(a)),$$

where $a = \alpha^k$ (we've seen earlier, in the proof of Theorem 8.18, that this polynomial is in $F[x]$).

Now we can again consider the tower of extensions $E/E'(\zeta)/E'/F$. We know that $\mathrm{Gal}(E'/F)$ is solvable by the inductive assumption, $E'(\zeta)/E'$ is a subgroup of $(\mathbb{Z}/k\mathbb{Z})^\times$ and is therefore abelian, and $E/E'(\zeta)$ is a subgroup of $(\mathbb{Z}/k\mathbb{Z})^d$ (here d is the number of Galois conjugates of α^k; in particular $d = |\mathrm{Gal}(E/F)|$). □

Remark 8.39. This proof is essentially a direct application of Proposition 8.31 (on the Galois group of the radical extension $F(\alpha)/F$), except for the technicality of needing to add in all Galois conjugates in order to make E a splitting field.

Combining these observations gives us the conclusion:

Corollary 8.40

There are many polynomials whose splitting fields are not radical extensions of $\mathbb{Q}$, and therefore for which we cannot write their roots in terms of radicals.

Example 8.41

The polynomial $2x^5 - 5x - 10$ has Galois group S_5 (as seen earlier), which means its splitting field cannot be radical (as S_5 is not solvable).

§8.6.2 Symmetric Polynomials

We've seen that it's impossible to solve a general polynomial of degree at least 5 (in radicals). But Galois theory can give us more than a result about impossibility — it can also be used to *find* solutions when possible?

Question 8.42. Given a polynomial, how can we compute its Galois group and find its roots (if it is possible to do so)?

In order to answer this, we'll use symmetric polynomials.

Definition 8.43. Symmetric polynomials in n variables are polynomials which are invariant under permutations of the variables; we denote the ring of symmetric polynomials as

$$R_n = \mathbb{Z}[x_1, \dots, x_n]^{S_n} \subset \mathbb{Z}[x_1, \dots, x_n].$$

Example 8.44

If $n = 3$, then we have $x_1^3 + x_2^3 + x_3^3 \in R_3$, while $x_1^3 \notin R_3$.

It is easy to write down polynomials in R_n — we can start with any monomial, and average it over all permutations of variables. The most obvious example is the power sums $x_1^k + \cdots + x_n^k$, but it turns out that the most *useful* example is actually the elementary symmetric functions:

Definition 8.45. The *elementary symmetric functions*, denoted $\sigma_1^{(n)}, \dots, \sigma_n^{(n)}$ (where n is the number of variables), are defined as

$$\sigma_k^{(n)} = \sum_{i_1 < \cdots < i_k} x_{i_1} x_{i_2} \cdots x_{i_k}.$$

We will drop the superscript when the number of variables is unambiguous.

Example 8.46

We have $\sigma_2^{(3)} = x_1x_2 + x_1x_3 + x_2x_3$.

Example 8.47

For any n, the first elementary symmetric function is

$$\sigma_1^{(n)} = x_1 + x_2 + \cdots + x_n,$$

and the last is

$$\sigma_n^{(n)} = x_1 x_2 \cdots x_n.$$

Elementary symmetric functions arise naturally from polynomials — the coefficients of a polynomial are elementary symmetric functions in the roots. More precisely, we have

$$(z - x_1)(z - x_2)\cdots(z - x_n) = z^n - \sigma_1 z^{n-1} + \sigma_2 z^{n-2} - \cdots + (-1)^n \sigma_n$$

(which follows immediately from considering what happens when we expand the left-hand side). Meanwhile, they are extremely useful for the following reason:

Theorem 8.48

We have $R_n = \mathbb{Z}[\sigma_1, \sigma_2, \ldots, \sigma_n]$ — in other words, every symmetric polynomial can be written (uniquely) as a polynomial in the elementary symmetric functions.

Example 8.49

When $n = 3$, we can write

$$x_1^2 + x_2^2 + x_3^2 = \sigma_1^2 - 2\sigma_2,$$
$$x_1^3 + x_2^3 + x_3^3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3.$$

Before we prove the theorem, we'll describe a useful observation — the formulas in the above example don't actually depend on n, and for *any* $n \geq 3$ it's true that

$$x_1^2 + \cdots + x_n^2 = \sigma_1^2 - 2\sigma_2,$$
$$x_1^3 + \cdots + x_n^3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3.$$

In order to see this, we can consider the obvious homomorphism $r_n : \mathbb{Z}[x_1, \ldots, x_n] \to \mathbb{Z}[x_1, \ldots, x_{n-1}]$ sending $x_n \mapsto 0$ (so we essentially just eliminate the extra variable). This homomorphism sends $R_n \to R_{n-1}$ (since if a polynomial was invariant under permutations of all n variables, then when we eliminate one variable, it's still invariant under permutations of the remaining $n - 1$).

This homomorphism interacts well with the elementary symmetric functions — it's clear that $\sigma_i^{(n)} \mapsto \sigma_i^{(n-1)}$ for all $1 \leq i < n$, while $\sigma_n^{(n)} \mapsto 0$ (since we essentially just eliminate all of the terms containing x_n when performing the homomorphism).

Meanwhile, the power sums $x_1^d + \cdots + x_n^d$ have the same property — the homomorphism sends $x_1^d + \cdots + x_n^d \mapsto x_1^d + \cdots + x_{n-1}^d$. This immediately proves that if we have a formula for $x_1^d + \cdots + x_n^d$ in terms of the elementary symmetric functions for some number of variables n, then we can deduce the same formula for all $m < n$.

On the other hand, we also have degree considerations — σ_i is a homogeneous polynomial in x_1, …, x_n of degree i, so the possible i for which σ_i can be in the formula is restricted by the degree of our original expression. For example, the formula for $x_1^3 + x_2^3 + x_3^3$ must have degree 3, so it can only contain σ_1, σ_2, and σ_3. This means if we have a formula for $n = 3$, then the same formula must hold for all $n \geq 4$ as well. (It's not enough to have a formula for $n = 2$ in this case, since when $n = 2$, σ_3 has been sent to 0.)

We'll now prove the theorem.

Proof of Theorem 8.48. We use induction on n (the number of variables). We want to check both that every symmetric polynomial can be expressed as a polynomial in the elementary symmetric functions, and that this expression is unique. In other words, we can define the map $\varphi_n : \mathbb{Z}[t_1, \dots, t_n] \to R_n$ sending $t_i \mapsto \sigma_i^{(n)}$; then we want to check that φ_n is an isomorphism.

First we'll check that φ_n is injective; it suffices to show that it has trivial kernel. Suppose we have a polynomial Q such that $\varphi_n(Q) = 0$. Then write $Q = t_n^d Q'$, where $t_n \nmid Q'$. We then have

$$0 = \varphi_n(Q) = \varphi_n(t_n)^d \varphi_n(Q') = (x_1 x_2 \cdots x_n)^d \varphi_n(Q').$$

Since the first factor is nonzero, this means we must have $\varphi_n(Q') = 0$; so

$$Q'(\sigma_1, \dots, \sigma_n) = 0.$$

Now we can use our observation to eliminate one variable — let

$$\overline{Q}(t_1, \dots, t_{n-1}) = Q'(t_1, \dots, t_{n-1}, 0) = r_n(Q').$$

Note that $\overline{Q}$ is nonzero, as Q' is not divisible by t_n. But then we have

$$\varphi_{n-1}(\overline{Q}) = \overline{Q}(\sigma_1^{(n-1)}, \dots, \sigma_{n-1}^{(n-1)}) = 0,$$

contradicting the inductive assumption. (The base case of induction is $n = 1$, which is trivial.)

Now we'll check that φ_n is surjective, or in other words, that every $P \in R[x]$ can be written as a polynomial in $\sigma_1, \dots, \sigma_n$. We can assume that P is homogeneous of degree d; we now use induction on d as well.

First, by the inductive assumption, we can write

$$r_n(P) = T(\sigma_1^{(n-1)}, \dots, \sigma_{n-1}^{(n-1)})$$

for some polynomial T. Then we have

$$P - T(\sigma_1^{(n)}, \dots, \sigma_{n-1}^{(n)}) \in \ker(r_n) = (x_n),$$

which means that x_n must divide this difference. But since it's symmetric, then all other variables must divide the difference as well, and by unique factorization their product must divide the difference as well. So we can write

$$P - T(\sigma_1^{(n)}, \dots, \sigma_{n-1}^{(n)}) = \sigma_n^{(n)} \cdot Q$$

for a polynomial Q, which must be a symmetric polynomial which is homogeneous and has degree $d - n$. By the inductive assumption, we have $Q = S(\sigma_1, \dots, \sigma_n)$, which means

$$P = T(\sigma_1, \dots, \sigma_{n-1}) + \sigma_n S(\sigma_1, \dots, \sigma_n)$$

is a polynomial in the σ_i, as desired. □

§8.6.3 The Discriminant

An especially important example of a symmetric polynomial is the discriminant:

Definition 8.50. The *discriminant* of a polynomial $P(z) = (z - \alpha_1) \cdots (z - \alpha_n)$ is

$$D(P) = \prod_{i<j} (\alpha_j - \alpha_i)^2.$$

In particular, $D(P) = 0$ iff P has a double root.

Note that the discriminant is a symmetric polynomial in the roots of P. By Theorem 8.48, we can write

$$\prod_{i<j}(x_i - x_j)^2 = \Delta_n(\sigma_1, \dots, \sigma_n)$$

for some polynomial Δ_n; then if $P(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0$, we have

$$D(P) = \Delta(-a_{n-1}, a_{n-2}, \dots, (-1)^n a_0).$$

Example 8.51

The discriminant of $x^2 + bx + c$ is the familiar expression

$$D = b^2 - 4c.$$

Example 8.52

The discriminant of $x^3 + px + q$ is

$$D = -4p^3 - 27q^2.$$

Proof. It's possible to explicitly compute D, but we can simplify the computation greatly by using degree considerations — we know that D is a polynomial in p and q; but D has degree 6 in the roots, while p and q have degree 2 and 3 respectively. The only ways to write 6 as a sum of 2's and 3's are $2+2+2$ and $3+3$; therefore we must have

$$D = ap^3 + bq^2$$

for some a and b. We can now plug in different values of p and q to find a and b — if we take $P(x) = x^3 - x$, then $D = 4$, $p = -1$, and $q = 0$, which implies that $a = -4$. Then if we take $P(x) = (x-1)^2(x+2) = x^3 - 3x + 2$, we have $D = 0$, $p = -3$, and $q = 2$, which gives $b = -27$. □

Remark 8.53. Note that $x^3 + px + q$ is a *depressed cubic*, since its x^2 coefficient is 0. The formula for a general cubic is more complicated; but it's actually enough to consider this case, since it's always possible to eliminate the x^2 coefficient by shifting the variable.

The discriminant plays an important role in Galois theory — consider the expression

$$\delta_n = \prod_{i<j}(x_j - x_i),$$

where the discriminant is $D = \delta_n^2$. Note that δ_n isn't symmetric; in particular, swapping two adjacent variables will flip its sign. Since transpositions of adjacent elements generate S_n, this means for any permutation $\tau \in S_n$, we have

$$\delta_n(x_{\tau(1)}, \dots, x_{\tau(n)}) = (-1)^{\operatorname{sgn}(\tau)}\delta(x_1, \dots, x_n).$$

Now consider a polynomial $P \in F[x]$, and suppose P has roots $\alpha_1, \dots, \alpha_n$ which are all distinct (although P does not necessarily have to be irreducible). Let the splitting field of P be $E = F(\alpha_1, \dots, \alpha_n)$, and let its Galois group be $G = \operatorname{Gal}(E/F)$. Since all elements of G must permute the roots α_i, we know that $G \subset S_n$.

Proposition 8.54

We have $G \subset A_n$ iff the discriminant of P is a square in F.

Proof. Consider the element

$$\delta = \prod_{i<j}(\alpha_j - \alpha_i) \in E,$$

so that $\delta^2 = D$ is the discriminant of P (and in particular, lies in F). An element $\sigma \in G$ then fixes δ if σ is even, and sends $\delta \mapsto -\delta$ if σ is odd.

But by Theorem 8.18, $\delta \in F$ iff δ is fixed by all elements $\sigma \in G$ (since F is exactly the fixed field corresponding to G). So $\delta \in F$ iff all such σ are even, or equivalently if $G \subset A_n$. □

§8.6.4 Cubic Polynomials

We've seen that the discriminant can be used to gain information about the Galois group. In the case $n = 3$, we can make this extremely concrete:

Corollary 8.55

If P is an irreducible polynomial of degree 3, then $\mathrm{Gal}(P)$ is A_3 if the discriminant of P is a square, and S_3 otherwise.

Proof. The Galois group must be a transitive subgroup of S_3, and the only two transitive subgroups are A_3 and S_3. Then this follows immediately from the fact that $\mathrm{Gal}(P) \subset A_3$ iff $D(P)$ is a square. □

Example 8.56

Find the Galois group of $x^3 - 3x - 1$.

Solution. The discriminant is

$$D = 4 \cdot 27 - 27 = 81,$$

which is a square; so the Galois group is $A_3 \cong \mathbb{Z}/3\mathbb{Z}$. □

Example 8.57

If F contains a primitive cube root of unity ω, find the Galois group of $x^3 - a$.

Solution. The discriminant is $D = -27a^2$. But since F contains $\omega = -\frac{1}{2} \pm \frac{\sqrt{-3}}{2}$, then -3, and therefore -27, is a square in F. So D is a square, and $G = \mathbb{Z}/3\mathbb{Z}$. (Note that this is a special case of Proposition 8.31, on Kummer extensions.) □

So this gives a nice and effective way to find the Galois group of a cubic polynomial. We'll now discuss more comcretely how to actually solve the polynomial.

Proposition 8.58

If F contains a primitive cube root of unity ω, and E/F is a Galois extension with $\mathrm{Gal}(E/F) = \mathbb{Z}/3\mathbb{Z}$, then $E = F(\alpha)$ for some α with $\alpha^3 \in F$.

Proof. It's enough to find some nonzero $\alpha in E$ such that $\sigma(\alpha)$ is either $\omega\alpha$ or $\omega^2\alpha$, where σ is a generator of $\mathrm{Gal}(E/F)$. Then we have $\sigma(\alpha^3) = \alpha^3$; since σ generates $\mathrm{Gal}(E/F)$, this implies that α^3 is fixed by all elements of $\mathrm{Gal}(E/F)$, so $\alpha^3 \in F$. Meanwhile α itself cannot be in F (as it is not fixed by σ); so $[F(\alpha) : F] = 3$. But we have $[E : F] = 3$ as well, so we must then have $E = F(\alpha)$.

In order to construct such an α, pick any $\beta \in E$ with $\beta \notin F$, and consider the two elements

$$\alpha_1 = \beta + \omega\sigma(\beta) + \omega^2\sigma^2(\beta),$$
$$\alpha_2 = \beta + \omega^2\sigma(\beta) + \omega\sigma^2(\beta).$$

It's clear (since σ^3 is the identity) that $\sigma(\alpha_1) = \omega^2\alpha_1$ and $\sigma(\alpha_2) = \omega\alpha_2$, so it remains to check that one of α_1 and α_2 is nonzero.

But if both were zero, then $(\beta, \sigma(\beta), \sigma^2(\beta))$ would be a solution to the linear system

$$a + \omega b + \omega^2 c = 0$$
$$a + \omega^2 b + \omega c = 0.$$

But the orthogonality of characters for $\mathbb{Z}/3$ (from representation theory) implies that the only solution is $a = b = c$. This would mean β is fixed by σ, and therefore by all elements of $\mathrm{Gal}(E/F)$, contradicting the fact that $\beta \notin F$.

So one of α_1 and α_2 is nonzero, and is therefore a generator with $\alpha^3 \in F$. □

Remark 8.59. The same proof works if we replace 3 by any prime, and with a bit more work, it can be generalized to work for any positive integer n.

This can be used to prove the converse to Proposition 8.37 — we've seen that every radical extension is solvable, but this can be used to prove that if E/F has solvable Galois group, then it is radical.

We know that E contains $F(\delta)$, where $\delta = \sqrt{D}$ (it's possible $\delta \in F$, but it usually isn't), and

$$\mathrm{Gal}(E/F(\delta)) = \mathbb{Z}/3\mathbb{Z}.$$

(We are assuming F contains a cube root of unity. We're also assuming $\mathrm{char}(F) = 0$, although similar ideas work as long as $\mathrm{char}(F) \neq 2, 3$.) The proposition implies that $E = F(\delta)(\alpha)$ for some α with $\alpha^3 \in F(\delta)$. This shows theoretically that it's possible to express all roots of P in terms of radicals (in fact, in terms of square roots and cube roots), but it's actually possible to turn our construction into an explicit formula for the roots.

Let β_1, β_2, and β_3 be the roots of P, with $\beta_i \in E$ for all i. Then $\mathrm{Gal}(E/F(\delta))$ must contain the cycle (123), which means we can take

$$\alpha = \beta_1 + \omega\beta_2 + \omega^2\beta_3$$

in the above construction. We've already seen that $\alpha^3 \in F(\delta)$, but it's actually possible to find an explicit expression for it using symmetric polynomials. We can compute

$$\alpha^3 = \beta_1^3 + \beta_2^3 + \beta_3^3 + 6\beta_1\beta_2\beta_3 + \omega(\beta_1^2\beta_2 + \beta_2^2\beta_3 + \beta_3^2\beta_1) + \omega^2(\beta_1\beta_2^2 + \beta_2\beta_3^2 + \beta_3\beta_1^2).$$

We've seen earlier how to compute $\beta_1^3 + \beta_2^3 + \beta_3^3$ and $\beta_1\beta_2\beta_3$ from symmetric polynomials — if $P(x) = x^3 + px + q$, then these are $-3q$ and $-q$ respectively. Let the remaining two expressions be A and B. Then $A + B$ is symmetric, and we can write

$$A + B = \sigma_1\sigma_2 - 3\sigma_3 = 3q.$$

Meanwhile, $A - B$ is not symmetric, but we have

$$A - B = (\beta_1 - \beta_2)(\beta_1 - \beta_3)(\beta_2 - \beta_3) = \delta.$$

With this, we're almost done — we can calculate A and B in terms of the coefficients of P and δ, and use this to compute α, as well as its counterpart

$$\alpha' = \beta_1 + \omega^2\beta_2 + \omega\beta_3.$$

Then we have

$$\beta_1 = \frac{\alpha + \alpha'}{3},$$

and we can calculate the other two roots similarly. Writing out these computations more explicitly results in Cardano's formula.

Remark 8.60. The fact that α^3 could be expressed using symmetric polynomials and δ can be shown more theoretically as well — in fact, we always have

$$\mathbb{Q}[x_1, \dots, x_n]^{A_n} = \mathbb{Q}[\sigma_1, \dots, \sigma_n] + \delta\mathbb{Q}[\sigma_1, \dots, \sigma_n].$$

In other words, we know that if instead of considering polynomials fixed by *all* permutations (which are the symmetric polynomials), we instead consider polynomials fixed by *even* permutations, then we get the new element δ. But this states that δ essentially accounts for *all* the new elements.

Remark 8.61. Cardano's formula was discovered in 1545, well before Galois theory. The method described here was found by Legendre — it's possible to find without using Galois theory, by just noticing that if $\alpha = \beta_1 + \omega\beta_2 + \omega^2\beta_3$, then α^3 can be written in terms of the symmetric polynomials and δ.

But the formula was discovered not only before Galois theory, but also before complex numbers. Working with roots of negative numbers previously gave people a lot of trouble, and was considered controversial until the 19th century.

In fact, suppose P has 3 real roots, so δ is real. But then $\alpha = \beta_1 + \omega\beta_2 + \omega^2\beta_3$ is *not* real — when you write down the formula in radicals, the *final* answer will be real, but inside the formula, you'll need to work with the cube root of a complex number. (This was known as *casus irreducibilis*.)

A book by Barry Mazur, called *Imagining Numbers, Especially* $\sqrt{-15}$, discusses the history of how the understanding of such concepts developed — people were working with complex numbers centuries before they were fully realized and accepted as existing.

§8.6.5 Quartic Equations

In the case of quartic equations, it's again possible to analyze solutions by analyzing the Galois group. We know the Galois group is a subgroup of S_4, and S_4 has the normal subgroup K_4 (the Klein 4-group — as a group, it's $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$), consisting of the elements (12)(34), (13)(24), and (14)(23). Meanwhile, $S_4/K_4 \cong S_3$.

We can use this to construct the *resolvent cubic*:

Definition 8.62. Given a quartic polynomial P with roots $\alpha_1, \dots, \alpha_4$, its *resolvent cubic* is the (monic) cubic polynomial with roots $\alpha_1\alpha_2 + \alpha_3\alpha_4$, $\alpha_1\alpha_3 + \alpha_2\alpha_4$, and $\alpha_1\alpha_4 + \alpha_2\alpha_3$ (which we denote by β_1, β_2, and β_3).

Clearly, every element of the Galois group permutes β_1, β_2, and β_3. We know that the coefficients of Q are symmetric polynomials in the roots of P, so if $P(x) = x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, then we can write $Q(x) = x^3 + b_2x^2 + b_1x + b_0$ where the b_i are polynomials in the a_j (concretely, the exact formulas are $b_2 = -a_2$, $b_1 = a_1a_3 - 4a_0$, and $b_0 = 4a_0a_2 - a_1^2 - a_0a_3^2$).

In order to compute the roots α_i of P, we first compute the roots of β_i of Q. Then we have

$$(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4) = \beta_1 + \beta_3,$$
$$(\alpha_1 + \alpha_2) + (\alpha_3 + \alpha_4) = -a_3.$$

This gives a quadratic equation which can be used to compute $\alpha_1 + \alpha_2$, and we can compute the other pairwise sums in the same way. Then we have a system of linear equations, which can be used to solve for the roots α_i.

Remark 8.63. The fact that all we have to do (once we've solved the cubic) is solve a few quadratics follows more theoretically from Galois theory as well — this happens because $K_4 \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

This shows how to find formulas for the roots of P, but similarly to the case of cubics, we can also ask about the Galois group:

Question 8.64. For a given irreducible polynomial P of degree 4 with splitting field E, how can we compute $\mathrm{Gal}(E/F)$?

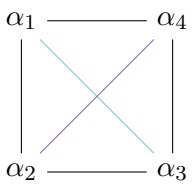
In degree 3, there were only two possibilities, and we could differentiate between them by considering whether the discriminant was a square or not. In this case, the process is longer, but uses similar ideas.

First, there are five transitive subgroups of S_4 — these are S_4, A_4, K_4 (consisting of (12)(34), (13)(24), and (14)(23)), C_4 (the cyclic group, generated by (1234)), and D_4 (the dihedral group, generated by (1234) and (24), which we can think of as the set of symmetries of a square).

It's still possible to calculate the discriminant from the coefficients of the polynomial (although the expression is very messy). Then as we've seen earlier, $\sqrt{D} \in F$ iff the Galois group is a subgroup of A_4; these subgroups are K_4 and A_4 itself.

We can obtain more information by looking at the resolvent cubic. Note that $Q(x)$ splits completely in F iff $G = K_4$ (since the permutations which preserve each of β_1, β_2, and β_3 are exactly K_4).

Meanwhile, if $Q(x)$ has exactly one root in F (meaning that the Galois group preserves exactly one of β_1, β_2, and β_3), then the Galois group must be C_4 or D_4. Intuitively, this is because C_4 and D_4 both represent a set of symmetries of a square:



Any symmetry of the square must preserve or swap the diagonals, and therefore must fix $\alpha_1\alpha_3 + \alpha_2\alpha_4$ (while the other two roots β_i may be permuted).

This information allows us to distinguish between all cases except C_4 and D_4, which we won't discuss.

§8.7 The Main Theorem of Algebra

Earlier, we saw a proof of the Main Theorem of Algebra using winding numbers. We'll now give another proof using Galois theory.

First, we need the following group-theoretic result:

Proposition 8.65

Every p-group is solvable. Furthermore, if G is a finite group of order p^n, then there exists a chain of subgroups

$$G = G_0 \supset G_1 \supset \cdots \supset G_n = \{1\}$$

such that for all i, G_{i+1} is a normal subgroup of G_i with $G_i/G_{i+1} \cong \mathbb{Z}/p\mathbb{Z}$.

Proof. We'll start constructing the sequence from the rightmost end.

Claim — G has nontrivial center.

Proof. Consider its class equation — every conjugacy class has size p^m for some m, and in particular, $p \mid |C|$ for every conjugacy class C with size greater than 1. But we have

$$p^n = 1 + \sum |C_i|,$$

where the sum is over all conjugacy classes except the identity. So we must have $|C_i| = 1$ for some i (or else the right-hand side would be 1 mod p), and the element of that conjugacy class is then in the center. □

Now we can induct on n. Pick an element $g \in Z$ of order p (which exists because the center is a p-group, so every non-identity element has a power with order p). Then $\overline{G} = G/\langle g \rangle$ has order p^{n-1} (note that we can take the quotient because g is in the center of G, and therefore $\langle g \rangle$ is normal). By the inductive assumption, we can then write

$$\overline{G} = \overline{G}_0 \supset \cdots \supset \overline{G}_d = \{1\}$$

for a chain of subgroups with $\overline{G}_i/\overline{G}_{i+1} \cong \mathbb{Z}/p\mathbb{Z}$ for all i.

Now let G_i be the pre-image of $\overline{G}_i$ (under the homomorphism sending $G \mapsto \overline{G}$ corresponding to quotienting out by $\langle g \rangle$), and take $G_{d+1} = \{1\}$. This works by the homomorphism theorem (and the fact that the pre-image of $G_d = \{1\}$ is $\langle g \rangle = \mathbb{Z}/p\mathbb{Z}$). □

Using this, we can prove the Main Theorem of Algebra (in a slightly different formulation, which is equivalent to the usual formulation that every polynomial has a complex root):

Theorem 8.66

$\mathbb{C}$ is the only finite extension of $\mathbb{R}$.

Proof. Let $E/\mathbb{R}$ be a finite extension, and WLOG assume that E is a splitting field (we know E is obtained by adjoining the roots of *some* polynomial, and we can add all the other roots as well), and is therefore a Galois extension; let $G = \mathrm{Gal}(E/\mathbb{R})$.

Claim — $|G|$ is a power of 2.

Proof. Let $H \subset G$ be a Sylow 2-subgroup (if the greatest power of 2 dividing $|G|$ is k, then H is a subgroup of order 2^k — this argument still works if $k = 0$ and H is trivial). Then

$$[E^H : F] = \frac{|G|}{|H|}$$

is odd, so $[E^H : F]$ has odd degree. But every odd-degree polynomial over $\mathbb{R}$ has a real root, so this is a contradiction unless $H = G$. (For any $\alpha \in E^H$, we've seen that $\deg(\alpha) \mid [E^H : F]$, so $\deg(\alpha)$ would have to be odd, and therefore the minimal polynomial of α would have odd degree. This is impossible unless it has degree 1.) □

Then we can use the previous proposition — we know we can write

$$G = G_0 \supset G_1 \supset \cdots \supset G_n = \{1\},$$

and considering the fixed fields of these subgroups gives a chain of extensions

$$\mathbb{R} = F_0 \supset F_1 \supset \cdots \supset F_n = E$$

(where F_i is the fixed field of G_i for all i), where $[F_i : F_{i-1}] = 2$ for all i.

But it's clear that $\mathbb{C}$ is the only *quadratic* extension of $\mathbb{R}$, and there are no quadratic extensions of $\mathbb{C}$ — any quadratic extension is obtained by adding a square root, and it's easy to check that every complex number has a square root in $\mathbb{C}$ (by using its trigonometric form).

So G must be $\{1\}$ or $\mathbb{Z}/2\mathbb{Z}$, and therefore E must be $\mathbb{R}$ or $\mathbb{C}$. □

§8.8 Galois Theory for Finite Fields

We've seen that when F is a number field (a finite extension of $\mathbb{Q}$), Galois groups $\mathrm{Gal}(E/F)$ can be complicated.

But $\mathbb{Q}$ is only one of the primary fields; the others are $\mathbb{F}_p$ for primes p. So we can also consider Galois extensions for finite extensions of $\mathbb{F}_p$. Any finite extension of $\mathbb{F}_p$ must be of the form $\mathbb{F}_q$ where $q = p^m$ for some m, and a Galois extension of $\mathbb{F}_q$ must again be of the form $\mathbb{F}_{q^n}$ for some n.

In this case, the answer is much easier, and in some sense we've seen it already.

Theorem 8.67

The extension $\mathbb{F}_{q^n}/\mathbb{F}_q$ is always Galois. Its Galois group is cyclic, and is generated by the *Frobenius automorphism* $\mathrm{Fr}_q : x \mapsto x^q$.

Proof. We've already seen that Fr_q is an automorphism, since

$$(a + b)^q = a^q + b^q.$$

Meanwhile, its fixed points are the set $\{x \mid x^q = x\}$, which is exactly $\mathbb{F}_q$. This means $\mathrm{Fr}_q \in \mathrm{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)$.

But we can also compute the order of Fr_q — we have $\mathrm{Fr}_q^a : x \mapsto x^{q^a}$. This means $\mathrm{Fr}_q^n = \mathrm{Id}$ (since $x^{q^n} = x$ for all $x \in \mathbb{F}_{q^n}$). Meanwhile, for all $1 \le a < n$, there exist elements of $\mathbb{F}_{q^n}$ with $x^{q^a} \neq x$, which means $\mathrm{Fr}_q^a \neq \mathrm{Id}$. So then $\mathrm{ord}(\mathrm{Fr}_q) = n$.

But we know $\langle \mathrm{Fr}_q \rangle \subset \mathrm{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)$, so $|\mathrm{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)| \ge n$. On the other hand, we have

$$|\mathrm{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)| \le [\mathbb{F}_{q^n} : \mathbb{F}_q] = n.$$

So equality must hold; this means $\mathbb{F}_{q^n}/\mathbb{F}_q$ is Galois, and $\mathrm{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q) = \langle \mathrm{Fr}_q \rangle$. □

Remark 8.68. The Frobenius automorphism is quite important. In particular, in order to count the number of solutions to a system of polynomial equations over $\mathbb{F}_q$, one first considers solutions over its algebraic closure

$$\overline{\mathbb{F}_q} = \bigcup_n \mathbb{F}_{q^n}.$$

Then the solutions in $(\mathbb{F}_q)^n$ (allowing polynomials in multiple variables) are exactly the fixed points of $\mathrm{Fr}_q : (x_1, \dots, x_n) \mapsto (x_1^q, \dots, x_n^q)$.

It's possible to use intuition from a similar problem in topology, of counting te number of fixed points of an automorphism of a geometric shape X (which relates to the Lipschitz Fixed Points Theorem and the Weil conjectures).

§9 Final Remarks

To conclude the class, we'll review the topics covered and discuss further directions in which they can be taken.

§9.1 Representation Theory

The first topic we discussed was the representations of finite groups; the class **18.715** develops this topic further.

We've seen the main *general* theorem about the representations of finite groups. But one further direction is to study the characters of a *specific* group — in particular, the group S_n.

In general, we know that the number of irreducible representations equals the number of conjugacy classes. In the case of S_n, it's actually possible to index both by the same set:

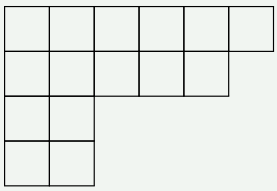
Definition 9.1. A *partition* of n is a nondecreasing sequence of positive integers $(\lambda_1, \dots, \lambda_k)$ with $\lambda_1 + \cdots + \lambda_k = n$.

As we've seen earlier, the conjugacy classes of S_n correspond to cycle types, which can be described as partitions of n — for example, the conjugacy class of $(12)(345)$ can be described by the partition $5 = 3 + 2$. Meanwhile, it turns out that the irreducible representations are *also* in bijection with partitions, where this bijection has nice properties.

One way to describe partitions is by Young diagrams, where for a partition $\lambda = (\lambda_1, \dots, \lambda_k)$, we draw λ_1 boxes in the first row, λ_2 in the second, and so on.

Example 9.2

The Young diagram



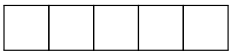
corresponds to the partition $15 = 6 + 5 + 2 + 2$.

The properties of this correspondence between irreducible representations and Young diagrams can be used to answer questions about these representations — for example, a question we answered earlier by looking directly at the characters.

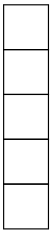
Example 9.3

For which n is $\tau \otimes \mathrm{sgn} = \tau$? (Here τ is the tautological representation of S_n, where elements of S_n act on the space with $x_1 + \cdots + x_n = 0$ by permuting the coordinates; and $\tau \otimes \mathrm{sgn}$ is the representation $\sigma \mapsto \tau(\sigma)\,\mathrm{sgn}(\sigma)$.)

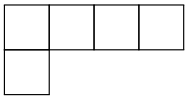
Sketch of Solution. In the correspondence between Young diagrams and irreducible representations, the Young diagram



corresponds to the trivial representation; the Young diagram



corresponds to sgn; and the Young diagram



corresponds to τ. In fact, for *any* irreducible representation ρ, the diagram corresponding to $\rho \otimes \mathrm{sgn}$ is the transpose of the diagram for ρ. So $\tau \otimes \mathrm{sgn} = \tau$ if and only if the last Young diagram shown, corresponding to $(n-1, 1)$, is its own transpose, which occurs if and only if $n = 3$. □

§9.1.1 Compact Lie Groups

We studied representations of *finite* groups, but it's also possible to study representations of compact Lie groups.

Definition 9.4. A *compact Lie group* is a closed compact subgroup of $\mathrm{GL}_n(\mathbb{C})$.

Example 9.5

Examples of compact Lie groups include the unitary groups $\mathrm{U}(n)$, the special unitary groups $\mathrm{SU}(n)$, the special orthogonal groups $\mathrm{SO}(n)$, and the quaternionic unitary groups $\mathrm{Sp}(n)$.

There's actually a classification of all compact Lie groups, with some exceptions — the exceptions are named G_2, F_4, E_6, E_7, and E_8 (the largest, E_8, has dimension 248). These groups are studied further in **18.745** and **18.755**.

For each compact Lie group, there's also a classification of its irreducible representations. For example, the irreducible representations of $\mathrm{U}(n)$ can be indexed by sequences of n integers $d_1 \geq \cdots \geq d_n$.

Example 9.6

The irreducible representations of $\mathrm{SU}(2)$ are indexed by nonnegative integers n. For each n, the corresponding irreducible representation acts on the space V_n of homogeneous polynomials of degree n in two variables. (This space has basis x^n, $x_{n-1}y$, …, y^n, and therefore has dimension $n+1$.)

Since $\mathrm{SO}(3) = \mathrm{SU}(2)/\{\pm 1\}$, then for even n, this construction can be used to form an irreducible representation of $\mathrm{SO}(3)$ as well. This actually relates to the structure of the periodic table — the rows of the tables have lengths 2, 8, 8, 18, 18, 32, 32. These are all of the form $2n^2$, and n^2 arises as $1 + 3 + \cdots + (n-1)$ — where the odd numbers come from the dimensions of irreducible representations of $\mathrm{SO}(3)$.

To describe the connection further, recall a problem from an optional problem set, on greedy monsters:

Problem 9.7. At each vertex of a cube, there is a greedy monster. Each monster starts with some amount of gold. Every minute, the gold of each monster is evenly distributed among its three neighbors. After a long time, what does the distribution of gold look like?

The process described here can be generalized:

Definition 9.8. Given a graph with certain weights assigned to each vertex, the *Laplace operator* redistributes the weight of each vertex equally among its neighbors.

Here, the operator is defined in *discrete* settings; but it's also possible to define it in *continuous* ones. For example, in $\mathbb{R}^2$, the Laplace operator is

$$\Delta = \frac{d^2}{dx^2} + \frac{d^2}{dy^2}.$$

In the discrete version, the operator vanishes (i.e., does nothing) when the weight at each point is the average of its neighbors; a similar statement is true for Δ.

In certain situations, it's possible to understand the eigenvalues and eigenvectors of the operator using representation theory (as we did in the solution to the greedy monsters problem). In particular, for the Laplace operator on the sphere S^2, its eigenvalues can be analyzed using the representation theory of $\mathrm{SO}(3)$. These eigenvalues also have connections to quantum physics.

On a different note, another important identity we saw in the representation theory of finite groups was

$$|G| = \sum d_i^2,$$

where d_i are the dimensions of the irreducible representations of G. To prove this, we looked at the regular representation on $\mathbb{C}[G]$, and decomposed it as

$$\mathbb{C}[G] = \bigoplus V_i^{d_i}.$$

We can think of $V_i^{d_i}$ as $\mathrm{End}(V_i)$ (since V_i has dimension d_i, so specifying a linear operator on V_i is the same as specifying the images of the d_i basis vectors), so then we can write

$$\mathbb{C}[G] \cong \bigoplus \mathrm{End}(V_i).$$

This generalizes to compact groups — a typical function can't be written as a finite sum, but it *can* be written as an infinite series.

Example 9.9

In the case of $\mathrm{U}(1)$ (which is just the unit circle), the irreducible representations are indexed by integers, where $\rho_n : \theta \mapsto e^{n\theta}$. Decomposing functions in this way turns into the theory of Fourier series.

Another possible direction is to study $\bigoplus \mathrm{End}(V_i)$ — we then have

$$\mathrm{MSpec}\left(\bigoplus \mathrm{End}(V_i)\right) = G_{\mathbb{C}},$$

where $G_{\mathbb{C}}$ is an *algebraic group* — this is studied in **18.737**.

Beyond the theory of representatinos of compact groups, one can also work with *non-compact* Lie groups — closed but not necessarily compact subgroups of $\mathrm{GL}_n(\mathbb{C})$, such as $\mathrm{SL}_n(\mathbb{R})$. In this case, most representations are infinite-dimensional. This is also studied in **18.755** and its continuations.

§9.2 Factorization

We've seen a story about factorization in quadratic number fields, or more precisely, in their rings of algebraic integers. This generalizes to factorization in rings of algebraic integers in more general number fields. This is studied in number theory, especially by using the action of the Galois group. An important question is often the following:

> **Question 9.10.** Given a prime ideal in a number field, how does it factor in a *larger* number field?

One example of such a result is quadratic reciprocity. Quadratic reciprocity is a classical result in elementary number theory; for example, it is taught in **18.781**. One part of the theorem is the following:

> **Theorem 9.11**
> If p and q are primes with $p \equiv 1 \pmod 4$, then p is a square mod q if and only if q is a square mod p.

Quadratic reciprocity has many proofs, including elementary ones. But there's also a proof that connects well to algebraic number theory in general, which actually uses ideas similar to ones we've seen in class — the main idea is to consider $\mathbb{Q}(\zeta_p)$, where ζ_p is a primitive pth root of unity. As we proved in class, $\mathbb{Q}(\zeta_p)$ contains $\mathbb{Q}(\sqrt{\pm p})$ (where we have $\sqrt{p}$ if $p \equiv 1 \pmod 4$, and $\sqrt{-p}$ if $p \equiv 3 \pmod 4$). Quadratic reciprocity can be proven by analyzing the factorization of q in the rings of algebraic integers of these two fields $\mathbb{Q}(\sqrt{\pm p})$ and $\mathbb{Q}(\zeta_p)$, and using the description of the Galois group to look at this factorization in two different ways.

In number theory, this generalizes to higher reciprocity laws as well.

§9.3 Rings and Modules

When discussing rings and modules, one of the main theorems we saw was the classification of the finitely generated modules over a PID. THe class **18.705** on commutative algebra develops this much further.

Commutative algebra is also closely related to algebraic geometry. If we have a ring R which is a quotient of $\mathbb{C}[x_1, \dots, x_n]$, then as mentioned earlier, we can study its maximal spectrum $\mathrm{MSpec}(R)$, which (as we've seen before) is a subset of $\mathbb{C}^n$.

Now suppose we have a R-module M and an element $x \in \mathrm{MSpec}(R)$ corresponding to the maximal ideal $\mathfrak{m}_x$. Then by Nullstelensatz, we have $R/\mathfrak{m}_x = \mathbb{C}$. This means $M/\mathfrak{m}_x M$ is a $\mathbb{C}$-vector space, and if M is finitely generated, then this vector space is finite-dimensional. So if we fix the module M, then we get a family of vector spaces corresponding to M, indexed by elements $x \in \mathrm{MSpec}(R)$. This idea is also studied in topology and differential geometry.

§9.4 Galois Theory

We've seen a story relating field extensions to groups — our key examples here were extensions of number fields, and extensions of $\mathbb{C}(t)$ (the latter was only sketched, but it's an important example as well).

As we've seen, Galois theory can be used to prove the impossibility of solving a general quintic in radicals. At around the same time Galois worked on this problem, another mathematician, Abel, also worked on the same problems, but from a different perspective — involving Riemann surfaces (which relate to complex analysis). Galois theory allows us to say that for a *specific* polynomial equation, there's no formula for the solution in radicals; Abel's proof showed that there's no *universal* formula (such as the quadratic equation).