Ring Theory (Math 113), Summer 2016

James McIvor
University of California, Berkeley
August 9, 2016

Abstract

These are some informal notes on rings and fields, for the course Math 113 at UC Berkeley, Summer 2016. We go through the basic stuff: rings, homomorphisms, isomorphisms, ideals and quotient rings, division and (ir)reducibility, all heavy on the examples, mostly polynomial rings and their quotients. The aim of the course is to introduce the minimal amount of technical baggage in order to expose the beautiful idea of Galois Theory, in which the course culminates.

Contents

1 Basic Examples and Definitions

Next lecture

1.1 Preliminary Examples

A ring is just a set where you can add, subtract, and multiply. In some rings you can divide, and in others you can't. There are many familiar examples of rings, the main ones falling into two camps: "number systems" and "functions".

1. $\mathbb{Z}$: the integers $\ldots, -2, -1, 0, 1, 2, \ldots$, with usual addition and multiplication, form a ring. Note that we cannot always divide, since $1/2$ is no longer an integer.

2. Similarly, the familiar number systems $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all rings[1].

3. $2\mathbb{Z}$: the even integers $\ldots, -4, -2, 0, 2, 4, \ldots$. This ring is "non-unital", to be defined shortly, and as such will not be under much consideration in our course.

4. $\mathbb{Z}[x]$: this is the set of polynomials whose coefficients are integers. It is an "extension" of $\mathbb{Z}$ in the sense that we allow all the integers, plus an "extra symbol" x, which we are allowed to multiply and add, giving rise to x^2, x^3, etc., as well as $2x, 3x$, etc. Adding up various combinations of these gives all the possible integer polynomials.

5. $\mathbb{Z}[x, y, z]$: polynomials in three variables with integer coefficients. This is an extension of the previous ring, too. In fact you can continue adding variables to get larger and larger rings.

6. $\mathbb{Z}/n\mathbb{Z}$: The integers mod n. These are equivalence classes of the integers under the equivalence relation "congruence mod n". If we just think about addition (and subtraction), this is exactly the cyclic group of order n, as discussed a long time ago. However, when we call it a ring, it means we are also using the operation of multiplication.

7. $C[0, 1]$: This is my notation for the set of all continuous real-valued functions on the interval $[0, 1]$. For example, $f(x) = 2x$ and $g(x) = \sin x$ are in $C[0, 1]$. They can be added and multiplied to give $(f + g)(x) = 2x + \sin x$ and $(fg)(x) = 2x \sin x$, which are also elements of $C[0, 1]$. This is a very large ring, since there are lots and lots of continuous functions. Notice also that the polynomials from example 2 are contained as a proper subset of this ring. We will see in a bit that they form a "subring".

8. $M_n(\mathbb{R})$ (non-commutative): the set of $n \times n$ matrices with entries in $\mathbb{R}$. These form a ring, since we can add, subtract, and multiply square matrices. This is the first example we've seen where the order of multiplication matters: AB is not always equal to BA (usually it's not).

9. $\mathbb{Q}[[x]]$: this ring consists of what are called "formal power series" with entries in $\mathbb{Q}$ (the rational numbers). A power series is just a polynomial with (possibly) infinitely many terms, such as you see in a calculus course. The word "formal" means that we don't care whether they converge or not, so that the series $\Sigma n! x^n$ is perfectly good, even though you never talk about it in calculus because it only converges when $x = 0$. Because of this possible non-convergence, we can't think of these power series as functions, and we think of the x as a "formal variable", rather

[1] In fact they're **fields**, to be defined shortly.

than something for which we can substitute a numerical value. We are restricting the coefficients to be rational numbers for the sake of example, but you could just as well consider $\mathbb{Z}[[x]]$ or $\mathbb{R}[[x]]$.

10. $\mathbb{Z}[\frac{1}{p}]$, where p is a prime: We take the integers, and adjoin the fraction $\frac{1}{p}$. But since we can multiply elements in a ring, we can also obtain such fractions as $\frac{1}{p} \cdot \frac{1}{p} = \frac{1}{p^2}$. Iterating, we get all fractions $\frac{1}{p^k}$, for $k \geq 0$. Since we can also add, we get things like $\frac{1}{p} + \frac{1}{p} = \frac{2}{p}$. For example, if $p = 3$, then $\frac{2}{27}$ and $-\frac{40}{81}$ are in $\mathbb{Z}[\frac{1}{3}]$, whilst $\frac{1}{2}$ is not (only powers of 3 can be in the denominator, assuming the fraction is completely reduced). Note that all these elements live inside $\mathbb{Q}$. It's an example of a subring, to be defined tomorrow.

11. $\mathbb{R}[x]/(x^2+1)$. Here's some new notation. It means take the polynomial ring $\mathbb{R}[x]$ as above, and "mod out" by the polynomial x^2+1, meaning that this polynomial gets set to zero. So in this ring, the polynomial $(x+1)^2$ is the same as $2x$, since $(x+1)^2 = x^2+2x+1 = 2x+(x^2+1) = 2x+0 = 2x$. Another way of thinking about this is that x^2 is the same as -1. So there are never any powers of x larger than 1, since whenever we get to x^2 we just swap it out for -1. So every polynomial in here is going to have a constant term and an x term and that's it. This should remind you of the complex numbers, which each have a real part (the constant term) and an imaginary part (the x term), but usually when we work with complex numbers, we use the letter i instead of x. But it's essentially the same ring. Note that for complex numbers, we can always divide (except by zero, of course), so that shows that in this weird polynomial ring, we can divide as well, which is a bit strange, since in the usual polynomial ring we can almost never divide (since, for example, $\frac{1}{x}$ doesn't count as a polynomial). This is an example of a quotient ring, which is the ring version of a quotient group, and which is a very very important and useful concept.

12. Here's a really strange example. Consider a set S (finite or infinite), and let R be the set of all subsets of S. We can make R into a ring by defining the addition and multiplication as follows. For two subsets A, B, define $A + B = A \cup B \setminus A \cap B$ (sometimes people call this the symmetric difference, or "exclusive or"). Define subtraction by $-A = S \setminus A$ (the set-theoretic complement). Thus $A - B = (A \cup (S \setminus B)) \setminus (A \cap S \setminus B)$. This example shows you that addition and multiplication needn't be the usual operations we know from grade school. But luckily, in most of our examples, like above, they will be.

1.2 Definition of a Ring

As the preceding examples indicate, a ring is basically a set in which we have a way of adding, subtracting, multiplying, but not necessarily dividing[2] Of course, depending on the ring, the addition and multiplication may not seem like the ordinary operations we are used to. So here's the formal definition:

Definition 1.2.1. A **ring** is a set R endowed with two binary operations, usually denoted $+$ and $\cdot$, such that

- R1: R is an abelian group with respect to $+$
- R2: For any a, b, c in R, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ (associativity of $\cdot$)
- R3: For any a, b, c in R, $a \cdot (b + c) = a \cdot b + a \cdot c$ (left-distributivity)
- R3': For any a, b, c in R, $(a + b) \cdot c = a \cdot c + b \cdot c$ (right-distributivity)

Most often we will also impose some additional conditions on our rings, as follows:

- R4: There exists an element, denoted 1, which has the property that $a \cdot 1 = 1 \cdot a = a$ for all a in R (multiplicative identity)

[2]We will see later that a ring in which we can always divide is called a *field*.

- R5: $a \cdot b = b \cdot a$ for all a, b in R (commutativity of $\cdot$)

Notice that since R forms an abelian group under $+$, the addition is *always* commutative, and that there is also an additive identity, which we will usually denote by 0. So axioms 4 and 5 impose extra conditions on the multiplicative structure of R. A ring satisfying $R4$ is called a **ring with unity** (or sometimes a **unital ring**), where unity is just a fancy name for the multiplicative identity. A ring satisfying R5 is called a **commutative ring**.

From now on, except in certain specific examples, if the term "ring" is used, it will mean a ring satisfying R1-R5. That is to say, unless stated otherwise, all our rings will be unital rings.

As usual we use exponents to denote compounded multiplication; associativity guarantees that the usual rules for exponents apply. However, with rings (as opposed to multiplicative groups), we must use a little caution, since a^k may not make sense for $k < 0$, as a is not guaranteed to have a multiplicative inverse.

In most of the examples above it is easy to see what the additive and multiplicative identities are. What are they for example 10?

The axioms are just a *minimal* list of properties of the addition and multiplication. Others can be deduced from these, e.g.,

Lemma 1.2.2. *Let R be a ring, with additive and multiplicative identities 0 and 1, respectively. Then for all a, b in R,*

1. $0a = a0 = 0$;
2. $(-a)b = a(-b) = -(ab)$;
3. $(-a)(-b) = ab$;
4. $(na)b = a(nb) = n(ab)$ *for any n in $\mathbb{Z}$.*

In 4, note that n is not to be thought of as an element of R: the notation na just means $a + \cdots + a$, where there are n copies of a in the sum.

Proof.

1. Exercise
2. To show that $(-a)b = -(ab)$ is to show that the element $(-a)b$ is the additive inverse of ab; so we add them together, and hope to get zero. So $(-a)b + ab = ((-a) + a)b = (0)b = 0$ (by 1). The equality of $a(-b)$ and $-(ab)$ is similar.
3. Exercise
4. $(na)b = (a + \cdots + a)b = (ab + \cdots + ab) = n(ab) = a(b + \cdots + b) = a(nb)$

□

Example 1.2.3. (The **zero ring**) The axiom R4 begs the question: can 0 and 1 be the same? The answer is yes, but in that case it turns out that there is only one element in our ring, which is 0 (which is equal to 1). We call this the zero ring, and sometimes write it just as 0. Here's the reason: suppose 1=0 in a ring, and now pick any element r in this ring. Since $r = 1 \cdot r = 0 \cdot r = 0$, we find that every element is 0.

1.3 Special elements in a ring

Here we pick out some types of elements that can occur in rings:

Definition 1.3.1. Let a be an element of a ring R. We say that a is:

1. a **unit** if a has a multiplicative inverse, i.e., if there exists an element b in R such that $ab = ba = 1$; in this case, a is also said to be **invertible**, and b the **inverse** of a (and vice versa). Note also that b is a unit as well - units come in pairs. Of course, it's possible that $b = a$, i.e., an element may be its own inverse. The set of units in R is denoted $R^\times$ - the **group of units** of R;

2. a **zerodivisor** if $a \neq 0$ and there is a nonzero element b in R such that $ab = ba = 0$;

3. **nilpotent** if $a^k = 0$ for some $k \in \mathbb{N}$;

4. **idempotent** if $a^2 = a$.

Example 1.3.2. 1. In any ring 0 and 1 are (trivially) idempotent, and 0 is trivially nilpotent. 1 is always a unit ("unity is a unit")

2. In $\mathbb{Z}$, the units are ± 1, there are no zerodivisors, no nilpotent elements, and only 1 is idempotent.

3. In $\mathbb{Q}[x]$, the units are the nonzero constant polynomials, there are no zerodivisors, and no nontrivial idempotent or nilpotent elements.

4. In $M_n(\mathbb{R})$, the units are just the invertible matrices, which is just the multiplicative group $GL_n(\mathbb{R})$. There are plenty of zerodivisors: any strictly upper-triangular matrix multiplied by a strictly lower-triangular matrix is zero, so there are already lots of them. In fact, the zero-divisors are precisely the non-invertible matrices (except for 0, which never counts as a zerodivisor). This doesn't usually happen: in general, rings can contain many elements that are neither units nor zerodivisors. Nilpotents must have 0 as their only eigenvalue. Idempotents must be diagonalizable and have 0 or 1 as their only eigenvalue.

5. In $\mathbb{Z}/n\mathbb{Z}$, the units are those classes $\overline{m}$ for which $\gcd(m, n) = 1$. The zerodivisors are those for which $\gcd(m, n) \neq 1$. This is another ring in which every nonzero element is either a unit or a zerodivisor, but again do not be tempted to believe that this holds for all rings!

Definition 1.3.3. A nonzero ring in which every nonzero element is a unit is called a **field**.

Fields include many familiar number systems, such as $\mathbb{Q}, \mathbb{R}, \mathbb{C}$; $\mathbb{Z}$, on the other hand, is not a field.

2 Subrings; Homomorphisms

Previous lecture - Next lecture

IMPORTANT: FOR THE REMAINDER OF THESE LECTURES, ALL RINGS WILL BE ASSUMED COMMUTATIVE WITH UNITY, WITHOUT FURTHER MENTION, UNLESS EXPLICITLY STATED OTHERWISE.

In this lecture we will discuss how to easily get new rings out of other rings by finding rings inside others (subrings) or by "combining rings together", and also learn about homomorphisms, which just as for the groups, are the only types of functions which are of much interest in ring theory.

2.1 Subrings; Adjoining Elements

Definition 2.1.1. If R is a ring, and S is a subset of R we will say that S is a **subring** of R if

1. S is a subgroup of R under $+$.
2. S is closed under multiplication, and
3. 1 is in S.

We've seen subrings already: in yesterday's examples, $\mathbb{Z}$ is a subring of $\mathbb{Z}[x]$, which is in turn a subring of $\mathbb{Z}[x, y, z]$, etc. Similarly, $\mathbb{Z}[\{\frac{1}{p}\}_{p \text{ is prime}}]$ is a subring of $\mathbb{Q}$, which is a subring of $\mathbb{R}$, etc.

In general, there is a construction called **adjoining an element** defined as follows: start with a ring R, and add a new element x (of course, if x is already being used as the name of an element in R, then you'd better choose a different letter...). This x could be a "formal variable", or it could be a known element of some other ring containing R. We build a ring $R[x]$ (read aloud as "R adjoin x"), which contains all elements of R as well as the new element x. Since we allow addition and multiplication we must also include x^2, x^3, etc, as well as any product rx^k (and $x^r k$, if we want it to be commutative), where $r \in R$. Since we should be able to add, we also obtain sums of the form $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$, where the $a_i \in R$. Thus, as you can see, $R[x]$ is simply a polynomial ring with coefficients in R. But of course, the elements of R could be all sorts of things, in which case elements of $R[x]$ may not look like ordinary polynomials, which typically have some sorts of numbers ($\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$) as their coefficients. Thus if you wanted, you could build a ring whose elements were polynomials with matrix coefficients.

The construction of $R[x]$ realizes R as a subring of a larger ring, and is in some sense the "smallest" way to do so, as we only added one element, and those which were forced upon us by the ring axioms.

Example 2.1.2.

1. The **Gaussian integers** are defined as the subring of $\mathbb{C}$ given by adjoining i to the integers, namely $\mathbb{Z}[i]$. They can be pictured as a square lattice in the complex plane. They contain $\mathbb{Z}$ as a proper subring.
2. Rings such as $\mathbb{Q}[\sqrt{2}]$, where we adjoin an irrational square root to the rational numbers, are of great importance in number theory. They are called **quadratic number fields**. Since $(\sqrt{2})^2 \in \mathbb{Q}$, we don't need any higher powers of the new element $\sqrt{2}$, so actually $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$. In other words, if we think of this as some set of polynomial ring, where the "variable" is $\sqrt{2}$, then actually the only polynomials we need are linear.

2.2 Products of Rings

Definition 2.2.1. Let R and S be two rings. Their **product**, sometimes called the **direct product**, denoted $R \times S$, is the ring

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

So as a set, $R \times S$ is just the Cartesian product. It's made into a ring by defining addition and multiplication componentwise:

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2); \quad (r_1, s_1) \cdot (r_2, s_2) = (r_1 r_2, s_1 s_2)$$

The zero element (additive identity) of this ring is just $(0, 0)$ - note that the first zero lives in R, but the second lives in S, so it's bad notation. The multiplicative identity is $(1, 1)$.

Inside $R \times S$, there is "a copy" of R, namely the set

$$R \times \{0\} = \{(r, 0) \mid r \in R\}$$

Correction (7/19): This is a NOT subring of $R \times S$, as it doesn't contain the multipplicative identity $(1, 1)$. However, it is a ring in its own right, it just doesn't share the same multiplicative identity as $R \times S$. It's not exactly the same as R, since things in $R \times \{0\}$ are still ordered pairs, whereas elements of R are not. But you can see that they're basically the same. In fancy jargon, R is "isomorphic" to $R \times \{0\}$.

Example 2.2.2.

1. Products of rings always have lots of zerodivisors. For example, in $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$, if a, b are nonzero integers, then $(a, 0)$ and $(0, b)$ are nonzero elements whose product is zero, so they are zerodivisors.

2. $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ is a ring with 6 elements. So is $\mathbb{Z}/6$. Your intuition from our study of them as abelian groups may tell you that they are the same. We will see that they're **isomorphic** as rings, just as they were isomorphic as groups. But it's crucial that 2 and 3 are coprime: $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is not the same as $\mathbb{Z}/4\mathbb{Z}$, not least because they're not even the same as groups!

3. Products of rings also have nontrivial idempotents, which is a comparatively rare phenomenon[1]. For instance, $\mathbb{Z}$ has no nontrivial idempotents, whereas $\mathbb{Z} \times \mathbb{Z}$ has the idempotents $(1, 0)$ and $(0, 1)$.

2.3 Homomorphisms

Just as with groups, when we study rings, we are only concerned with functions that "preserve the structure" of a ring, and these are called ring homomorphisms. Maybe you can guess what the definition should be, by analogy with the case of groups.

Definition 2.3.1. Let R and S be rings, and $\phi\colon R \to S$ be a function. We say ϕ is a **ring homomorphism** if, for all a, b in R,

1. $\phi(a + b) = \phi(a) + \phi(b)$,
2. $\phi(a \cdot b) = \phi(a) \cdot \phi(b)$, and
3. $\phi(1) = 1$.

Note that in (3), the first 1 is in R, while the second 1 is in S.

Notice that we explicitly require that ϕ sends 1 to 1. What about the additive identity 0? Why don't we have to require that $\phi(0) = 0$? This is because, as we proved a while ago, since R and S are groups under addition, it follows automatically. But since neither R or S are groups under multiplication, we have to add in this condition separately.

Definition 2.3.2.

1. Let R and S be two rings. The set of all homomorphisms from R to S is denoted $\mathrm{Hom}(R, S)$.
2. A homomorphism is **injective** or **surjective** if it so as a map of sets (i.e., the usual definitions apply)

Example 2.3.3.

1. $f\colon \mathbb{Z} \to \mathbb{Z}$ given by $f(n) = n + 1$ is not a ring homomorphism. It fails conditions 1,2, and 3.

[1] In fact, you can show that the existence of nontrivial idempotents in a ring R allows one to decompose R into a product of subrings.

2. In fact, there is only one ring homomorphism from $\mathbb{Z}$ to $\mathbb{Z}$, the identity map, which sends each integer to itself. This is one of the reasons why $\mathbb{Z}$ is a very important ring.

3. The map $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ which sends an integer m to its congruence class $\overline{m}$ mod n is a ring homomorphism.

4. $\mathrm{Ev}_a \colon \mathbb{Z}[x] \to \mathbb{Z}$ given by $\mathrm{Ev}_a(p(x)) = p(a)$ is a ring homomorphism, called the **evaluation map** at a. It means simply "plug in a". This type of homomorphism is ubiquitous, since polynomials can be viewed as functions, and for functions we just "plug in" an element. One can define the same sort of map with $\mathbb{Z}$ replaced by an arbitrary ring R.

5. The map $\mathbb{Z} \to \mathbb{Z}$ sending n to n^k is not a ring homomorphism unless $k = 1$. For example, if $k = 2$, then since $2 = 1^2 + 1^2 \neq (1+1)^2 = 4$, it is not additive.

6. If p is a prime, and R is a ring in which $p = 1 + 1 + \cdots + 1 = 0$, then the map $R \to R$ which sends r to r^p *is* a ring homomorphism. In other words, when we work mod p, the pth power map is a ring homomorphism.[2] Contrast this with the example above.

7. If R is any ring, then there are lots of homomorphisms from $\mathbb{Z}[x]$ to R. All we have to do is pick an element r in R, and send x to r. The rest of the map is forced by the definition of homomorphism (see following proposition).

Proposition 2.3.4. *For any ring R, the set* $\mathrm{Hom}(\mathbb{Z}[x], R)$ *is in bijection with R.*

Proof. We will define a bijection from R to $\mathrm{Hom}(\mathbb{Z}[x], R)$. For each r in R, define a map f_r from $\mathbb{Z}[x]$ to R by

$$f_r(a_n x^n + \cdots + a_1 x + a_0) = a_n r^n + \cdots + a_1 r + a_0 \cdot 1$$

Since the term on the right is just various powers of r added together (remember that the a_i are integers), it makes sense as an element of R. Now we have to check some things: a) f_r is a homomorphism, b) the correspondence between the rs and f_rs is one-to-one, and c) it's onto.

To see that f_r is a homomorphism, it's convenient to write $f_r(p(x)) = p(r)$. This is just a shorthand notation for the formula above, where we've written $p(x)$ for $a_n x^n + \cdots + a_1 x + a_0$. Then

$$f_r(p(x) + q(x)) = p(r) + q(r) = f_r(p(x)) + f_r(q(x))$$

and

$$f_r(p(x)q(x)) = p(r)q(r) = f_r(p(x))f_r(q(x)),$$

which proves the first two properties. Finally, $f_r(1) = 1$, simply because there is nowhere to substitute r for x.

Now we have to show the one-to-one part. So suppose r and r' are two distinct elements of R. Then we get two maps f_r and $f_{r'}$ and we have to show they're different. Well, $f_r(x) = r$, and $f_{r'}(x) = r'$. Since these two homomorphisms send the polynomial x to different elements of R, they must be different homomorphisms, which establishes that the correspondence is one-to-one.

To show the correspondence is onto, suppose we start with a homomorphism f from $\mathbb{Z}[x]$ to R. We have to show it has the form f_r for some element r of R. First plug in the polynomial x to our map f. That gives us our element r. This r in turn gives rise to a map f_r. Now we have to show that f and the new f_r are the same map. By the definition of f_r, we have $f_r(p(x)) = p(r)$. And writing $p(x) = a_n x^n + \cdots + a_1 x + a_0$, we have

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

[2] In fact, in the case where $R = \mathbb{Z}/p\mathbb{Z}$ (for which the condition $p = 0$ indeed holds), Fermat's little theorem says that $n^p \equiv n \mod p$, so this map is none other than the identity map.

This shows that when applied to any polynomial $p(x)$, f and f_r give the same result, so they're the same homomorphism. Thus any given f can be written as f_r (where we choose r to be $f(x)$), so the correspondence is onto.

□

The proposition can be restated as follows: a homomorphism out of $\mathbb{Z}[x]$ is uniquely determined by where it sends x. An analogous statement is true for maps out of $\mathbb{Z}[x, y]$, etc.

The last part of the proof suggests why polynomial rings are so ubiquitous in ring theory: polynomials are built entirely out of addition and multiplication, and homomorphisms "pass through" addition and multiplication. So it is very easy to define homomorphisms on polynomials.

2.4 Isomorphisms

Just as for groups, bijective homomorphisms are called isomorphisms, and they tell us when two rings "have the same structure".

Definition 2.4.1. An **isomorphism** from a ring R to another ring S is a bijective homomorphism. If an isomorphism between R and S exists, then we say R and S are **isomorphic** and we write $R \cong S$.

There is an alternative way to characterize isomorphisms, using inverse functions.

Proposition 2.4.2. *Let $f : R \to S$ be a homomorphism. Then f is an isomorphism if and only if there exists a homomorphism $g : S \to R$ such that $g \circ f$ is the identity map on R and $f \circ g$ is the identity map on S.*

Proof. Exercise.

□

Example 2.4.3. Inside the matrix ring $M_2(\mathbb{R})$, there is a subring

$$R = \left\{ \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} \;\middle|\; n \in \mathbb{Z} \right\}.$$

Even though $M_2(\mathbb{R})$ is not a commutative ring, the subring R is commutative, and it is isomorphic to $\mathbb{Z}$, which is probably not surprising. To prove this, we define a map $\phi : \mathbb{Z} \to R$ by, for n in $\mathbb{Z}$,

$$\phi(n) = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}.$$

First we check it's a homomorphism:

1. Additivity:

$$\phi(m+n) = \begin{pmatrix} m+n & 0 \\ 0 & m+n \end{pmatrix} = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} + \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} = \phi(m) + \phi(n)$$

2. Multiplicativity:

$$\phi(mn) = \begin{pmatrix} mn & 0 \\ 0 & mn \end{pmatrix} = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} = \phi(m)\phi(n)$$

3. Finally, ϕ sends 1 in $\mathbb{Z}$ to 1 in R, the role of 1 in R being played by the identity matrix.

So ϕ is a homomorphism. To see that it's an isomorphism, we can either check injectivity and subjectivity, or go by the definition and produce an inverse. We'll go by the definition. The inverse is $\psi : R \to \mathbb{Z}$, given by

$$\psi \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} = n.$$

This is one of those points where authors tend to say "it's obvious that ϕ and ψ are inverses", but I'll write out the details as a model for future reference. We need to check that $\psi(\phi(n)) = n$ and $\phi\left(\psi\begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}\right) = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}$. Well,

$$\psi(\phi(n)) = \psi\begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} = n,$$

and

$$\phi\left(\begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}\right) = \phi(n) = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix},$$

using only the definitions of the two maps.

3 Kernels and Ideals

Previous lecture - Next lecture

3.1 The kernel of a homomorphism

Definition 3.1.1. Let $\phi\colon R \to S$ be a homomorphism of rings. The **kernel** of ϕ, denoted $\ker\phi$, is the subset $\{r \in R \mid \phi(r) = 0\}$ of R. In other words, it's the pre-image of 0 under ϕ.

The **image of** ϕ is the set $\operatorname{im}\phi = \{s \in S \mid s = \phi(r) \text{ for some } r \in R\}$.

The image of a homomorphism is a subring of the codomain; the kernel is not generally a subring, it's what's called an ideal (to be defined shortly). Because of the prominence of ideals in the theory, the kernel tends to be a more prominent player than the image in our analysis of homomorphisms. However, as is the case for groups, there is an "isomorphism theorem" saying that the two are "balanced" (see lecture 5).

Example 3.1.2.

1. The kernel of $\phi\colon \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ sending m to $[m]$ is $\{kn \mid k \in \mathbb{Z}\}$, in other words, the set of all multiples of n.
2. The kernel of $\phi\colon \mathbb{Z}[x] \to \mathbb{Z}[x]$ sending $p(x)$ to $p(1)$ is the set of all multiples of $x - 1$.
3. The kernel of the map $\mathbb{Z}[x] \to \mathbb{Q}$ which sends x to $1/2$ is the set of all multiples of $2x - 1$.
4. The kernel of the map $\mathbb{Z}[x] \to \mathbb{Z}/2\mathbb{Z}[x]$ which sends a polynomial $p(x)$ to the congruence class $\overline{p(x)}$ (in other words, "reduce the coefficients mod 2") is the set of all polynomials with even coefficients.

Just as for groups, the kernel and image detect injectivity and surjectivity: to be injective means to have a trivial kernel, so maps with a large kernel can be thought of as "very un-injective".

Proposition 3.1.3. *Let $f : R \to S$ be a homomorphism of rings. Then f is injective if and only if* $\ker f = \{0_R\}$*; f is surjective if and only if* $\operatorname{im} f = S$.

Proof. Worksheet problem. □

3.2 Ideals

Notice that in the examples above, the kernels were not even subrings, as can be easily seen by the fact that 1 was not in the kernel. So what were they? They were ideals, which is in some ways like a normal subgroup, and in some ways like a vector subspace in linear algebra.

Let's look back at example (1) to motivate the definition (we'll assume $n > 1$ here). The kernel of the map $m \mapsto \overline{m}$ is the set of multiples of n, which we denote by (n). What properties does this set have? It doesn't include 1, so it's not a subring, but it *is* closed under addition, so its "nearly" a subring. It's also closed under multiplication, since if an and bn are two elements of (n), then $an \cdot bn = abn \cdot n$, which is again a multiple of n. So it's really "nearly" a subring. But in many ways it's better: closure under multiplication means if we take any two elements of (n) and multiply together, we stay in (n). But in fact, something more is true: if we take one integer which is in (n), and another one which might not even be in (n), and multiply, we still stay in (n). In symbols: $a \cdot bn = ab \cdot n$, which is still in (n). So an ideal (in this case, the kernel) is closed even under multiplication when **one of the factors need not lie in the ideal**. This is stronger than the usual closure under multiplication, and is what makes ideals so special.

Definition 3.2.1. Let R be a commutative ring. A subset I of R is called an **ideal** if it satisfies the following conditions:

1. I is an additive subgroup of R under $+$ (additive subgroup)

2. For any i in I and any element r of R, $r \cdot i$ is in I (closed under scaling)

I call the second condition "closed under scaling" to distinguish it from "closed under multiplication". I borrow the term from linear algebra, because for ideals, I sometimes think of the elements of I as vectors, which are the important bits, and all the elements of r as scalars, which are used to multiply up the vectors. Of course, they all live in R, since I is a subset of R, but when dealing with an ideal, you usually keep your attention focused on the elements of the ideal rather than all the elements of the ring. So in some ways you can regard the definition of an ideal as being analogous to the definition of a vector subspace, which is (a) an additive subgroup, and (b) closed under scaling. Hopefully this helps you keep track of the difference between ideals and subrings, which confused me when I first learned this.

Notice that in the definition we required the ring R to be commutative. This isn't a big deal for us since most of our rings are commutative anyway. In a noncommutative ring there are notions of left ideals, right ideals, and two-sided ideals, depending on wether I is closed under scaling on the left, the right, or both. As you can imagine, tis complicates things quite a bit, and is one of the reasons why noncommutative rings are much harder to study.

Proposition 3.2.2. *If $f : R \to S$ is a homomorphism,* $\ker f$ *is an ideal of R.*

Proof. Exercise. □

Example 3.2.3.

1. Every ring has at least one ideal: the subset consisting of only 0. It's an additive subgroup, and closed under scaling because anything times zero is zero. We denote it by either (0) or just simply 0

2. As we've seen just before the definition, in $\mathbb{Z}$, for any n, the set (n) of all multiples of n is an ideal.

3. Similarly, in a polynomial ring $\mathbb{Z}[x]$ (or $\mathbb{R}[x]$, etc), for any polynomial $p(x)$, the set $(p(x))$ of all multiples of p is an ideal.

4. In a polynomial ring $\mathbb{Z}[x]$ (or $\mathbb{R}[x]$, etc), for any element a of $\mathbb{Z}$ (or $\mathbb{R}$, etc) the set of polynomials which go to zero when you plug in a for x forms an ideal. You might recognize this as the kernel of the map $p(x) \mapsto p(a)$.

5. In any ring R, the entire ring R is itself ideal, called the "unit ideal". The reason for this name is: if an ideal I contains 1, then it is equal to the entire ring, because if 1 is in I, then for any r, $r = r \cdot 1$ is also in I by closure under scaling. This is the largest ideal in R.

6. OMITTED - this example didn't make sense as previously posted.

7. What are all possible ideals in $R = \mathbb{Z}/6\mathbb{Z}$? This is a classic exam-style question. Let's work it out - let I be a "mystery" ideal. I will just write elements as integers, but they're all to be taken mod 6. First of all, if 1 is in I, then $I = R$, so from now on let's assume that 1 is not in I. We know 0 must be in I, and that if nothing else is, then I is just the zero ideal. So now let's assume that there is at least one other element a in I, which is nonzero, and not equal to 1. If $a = 5$, then since $5 \cdot 5 = 5 \cdot 5 = 25 = 1$, I would contain 1, and hence be the unit ideal again. So let's assume $\neq 5$, either. If $a = 3$, then the only multiples of a mod 6 are 3 and 0, so in this case $I = \{0, 3\}$. If $a = 2$, then all the multiples of 2 mod 6 are 0,2,4, we get $I = \{0, 2, 4\}$. Similarly if $a = 4$. What if I contains 2 and 3? Then by closure under addition it contains 5 as well, and then I would be the unit ideal as above. Similarly, if I contains one even and one odd number (mod 6), then we just get the unit ideal. So the possible ideals in $\mathbb{Z}/6\mathbb{Z}$ are

$$0, R, (\overline{2}), (\overline{3})$$

Note we have used the notation $(\cdots)$ as in the examples above to denote the set of multiples of a given element. These types of ideals come up so often that they have a special name:

Definition 3.2.4. Let X be any subset of a ring R. The **ideal generated by** X is the smallest ideal containing all elements of X, and is denoted (X). It can also be described as the set of all "R-linear combinations" of elements of X with coefficients in R. Elements of X are called **generators** of the ideal. If X is a finite set, say $X = \{x_1, \dots, x_k\}$, we will write this ideal as $(x_1, \dots, x_k)$. In the case when X is a singleton, $X = \{x\}$, the ideal (x) is called **principal**.

Note that there may be many different choices of generators for a given ideal. For example, $(n) = (-n)$ inside $\mathbb{Z}$; $(x) = (5x)$ inside $\mathbb{R}[x]$ (but *not* when considered as ideals in $\mathbb{Z}[x]$ - do you see why not?); and $(2, x) = (4, x - 2)$ inside $\mathbb{Z}[x]$. In fact, even the number of generators can be ambiguous: the ideals (2) and $(4, 6)$ are equal in $\mathbb{Z}$.

Principal ideals are the nicest possible ideals you can have, and there are lots of rings in which *every* ideal is principal, including $\mathbb{Z}$ and $\mathbb{Q}[x]$. These nice rings are called principal ideal domains, and we will study them more next week.

Notice also, in example 7, that we found as soon as our ideal contained a unit (in that case 1 or 5), it was the unit ideal. The next lemma says that this is always the case, thus justifying the term "unit ideal".

Lemma 3.2.5. *Let I be an ideal in R. I contains a unit if and only if $I = R$.*

The proof is an exercise.

3.3 Operations on Ideals

The following proposition gives ways to produce new ideals, both larger and smaller, from given ideals.

Proposition 3.3.1. *Let $\{I_a\}_{a \in A}$ be an arbitrary collection of ideals in R, indexed by a set A. Then*

1. *$\bigcap_{a \in A}$ is an ideal in R.*
2. *$I_a + I_b = \{x + y \mid x \in I_a, y \in I_j\}$ is an ideal, called the* **sum of I_a and I_b**. *It is the smallest ideal in R which contains both I_a and I_b.*

Proof. Exercise □

Example 3.3.2.

1. Let $I = (x^2, y)$, $J = (x, y^2)$ in $\mathbb{C}[x, y]$. Then $I \cap J = (x^2, xy, y^2)$, while $I + J = (x, y)$.
2. Let $m, n \in \mathbb{Z}$ be coprime. Then $(m) \cap (n) = (mn)$, while $(m) + (n) = (1) = \mathbb{Z}$. This second is another way of stating the fact that if $\gcd(m, n) = 1$, then there are integers a, b such that $am + bn = 1$.

4 Quotient Rings

Previous lecture - Next lecture

The concept of a quotient ring is one the most important concepts in the class, so pay attention!

4.1 Review: Cosets

We saw in the very beginning of the class that if you have a set S, and you choose an equivalence relation on it, you can partition your set into equivalence classes, which are disjoint and whose union is the entire set. Elements are in the same equivalence class if and only if they are equivalent under the relation you chose at the beginning. We also saw that conversely, any way you partition your set gives rise to an equivalence relation - you simply *define* two elements to be equivalent whenever they live in the same piece of the partition.

The idea of a coset in ring theory is basically the same as in group theory, except with ideals replacing subgroups. Let's go through it: let R be a ring, and I an ideal in R. First we define our equivalence relation: elements a, b of R are equivalent, written $a \sim b$, when $a - b$ is in I. Then by the general theory discussed in the previous paragraph, this partitions our ring into disjoint subsets, called cosets. One of the cosets is just I itself: this is the equivalence class of 0, since $a \sim 0$ exactly when $a = a - 0 \in I$. The other cosets look similar to I, they've just been "translated" by elements of R. We will write R/I for the set of cosets of I.

For a familiar example, take R to be $\mathbb{Z}$, and I to be the ideal (3). Then the cosets are (3) itself, and the two other subsets $\{\dots, -2, 1, 4, \dots\}$ and $\{\dots, -1, 2, 5, \dots\}$. They look the same (integers spaced 3 apart), but just "shifted".

You may object: this is exactly the same example we did in the group theory part - what do rings and ideals have to do with it? Nothing, yet, we'll only use the fact that I is an ideal in the following section. So far the definition of cosets only uses the fact that R is an abelian group under addition.

4.2 Quotient Rings

Now that we've defined the cosets, we want a way to turn the set of cosets R/I into a ring itself, and this is where the fact that I is an ideal comes into play. To fix notation, R is our ring, I our ideal in R, and if r is an element of R we write $\overline{r}$ for the coset containing r - another way of writing it is $r + I = \{r + i \mid i \in I\}$, namely the elements of I all shifted by r. We sometimes call r a **representative** of $\overline{r}$. The annoying part is that there may be many choices of representative for the same coset - this is where that stuff about things being "well-defined" comes in to play. We can add cosets in the same way as for groups: $\overline{r}_1 + \overline{r}_2 = \overline{r_1 + r_2}$. One has to fuss about to prove that this addition rule is well-defined, but I'll skip that since we did it for groups and we're about to do it again for the multiplication of cosets.

We multiply cosets in the obvious way: $\overline{r_1} \cdot \overline{r_2} = \overline{r_1 r_2}$. Now let's check this is well-defined. Suppose r_1 and s_1 are both representatives of the same coset (so $\overline{r_1} = \overline{s_1}$), and that r_2 and s_2 also represent the same coset. We need to check that whether we multiply the cosets using the r's or the s's we'll get the same answer. That's what it means to check that the multiplication is well-defined. In other words, we must show that $\overline{r_1} \cdot \overline{r_2} = \overline{s_1} \cdot \overline{s_2}$. Since r_1 and r_2 represent the same coset, they are equivalent (since cosets are equivalence classes), which means there is an element i_1 of I such that $r_1 = s_1 + i_1$. Similarly there is an i_2 such that $r_2 = s_2 + i_2$. This means that

$$\overline{r_1} \cdot \overline{r_2} = \overline{r_1 r_2} = \overline{(s_1 + i_1)(s_2 + i_2)} = \overline{s_1 s_2 + s_1 i_2 + s_2 i_1 + i_1 i_2} = \overline{s_1 s_2} = \overline{s_1} \cdot \overline{s_2}$$

You should be able to go through and justify why each equality is true. The trickiest one is the fourth: the cosets $\overline{s_1 s_2 + s_1 i_2 + s_2 i_1 + i_1 i_2}$ and $\overline{s_1 s_2}$ are the same because the two representatives are equivalent: $(s_1 s_2 + s_1 i_2 + s_2 i_1 + i_1 i_2) - (s_1 s_2) = s_1 i_2 + s_2 i_1 + i_1 i_2$, which is in I. You can see here why we need property 2 of the definition of an ideal. We want $s_1 i_2 + s_2 i_1 + i_1 i_2$ to be in I, and it wouldn't be if I were only closed under multiplication, since s_1 and s_2 do not necessarily live in I.

There's a lot more left to show that this addition and multiplication make R/I into a (commutative) ring (with unity): we have to check all the axioms for a (commutative) ring (with unity)! I won't type that all out in great detail. First of all, R/I has a zero element, which is just the coset I (usually

represented by the element 0 in R. Often we will just sloppily write 0 instead of the coset notation $\overline{0}$. Secondly, as you might guess, R/I has a multiplicative identity, which is just the coset $\overline{1}$. Again we may often write simply 1 instead of the coset notation $\overline{1}$.

We don't really need to check the axioms stating that R/I is an additive abelian group, since we've already done that when we discussed quotient groups. The other axioms basically follow from the corresponding property for R. For instance, here's a proof that left-distributivity holds. Let a, b, c be elements of R. Then

$$\overline{a} \cdot (\overline{b} + \overline{c}) = \overline{a \cdot (b + c)} = \overline{a \cdot b + a \cdot c} = \overline{a} \cdot \overline{b} + \overline{a} \cdot \overline{c}$$

4.3 The easy way to think about quotients of a polynomial ring

If all this seems vile and abstract, fear not: quotient rings are simple and quite fun to work with in practice. We've already seen one example: the integers mod n. Take $R = \mathbb{Z}$ and $I = (n)$. Then $R/I = \mathbb{Z}/n\mathbb{Z}$, and this is just the usual integers mod n where we do all the addition and multiplication mod n.

Now let's look at quotients of polynomial rings, as this is the fun bit. We'll just do examples until you get the idea:

Examples

1. Take $R = \mathbb{C}[x]$, and $I = (x - 2)$. We said before that the coset I acts as the zero element in R/I, so $\overline{x - 2} = 0$. This is the same, using the addition rule for cosets, as saying that $\overline{x} = \overline{2}$. So how does R/I look? Well, it's just like the polynomial ring, except we've replaced x with 2 and put bars over everything. Since 2 is already in $\mathbb{C}$, we haven't really added anything, so in fact this quotient ring is just like $\mathbb{C}$.

2. Take $R = \mathbb{R}[x]$, $I = (x^2 + 1)$. In the quotient ring $\mathbb{R}[x]/(x^2 + 1)$, we have cosets of polynomials, but with the rule that $\overline{x^2 + 1} = \overline{0}$, which we rewrite as $\overline{x}^2 = -\overline{1}$. So we think of elements of R/I here as polynomials, but with the extra rule that $\overline{x}^2 = -1$. This is something we've seen before: it's basically the same as $\mathbb{C}$, where we take $\mathbb{R}$ and adjoin a square root of -1, which in the current setting is being denoted by $\overline{x}$.

3. Let's work backwards a bit. Say you want to work with a number system that's like the rational numbers, but also for some reason includes the element $\sqrt{2}$. We can build this number system, which before we've called $\mathbb{Q}[\sqrt{2}]$, using quotient rings. Start with $\mathbb{Q}$. We'll need to adjoin a variable (call it x), and then set this variable equal to $\sqrt{2}$. We cannot just quotient out by $(x - \sqrt{2})$. because that's not a polynomial with rational coefficients. So the correct quotient ring seems like it should be $\mathbb{Q}[x]/(x^2 - 2)$. So the ring $\mathbb{Q}[x]/(x^2 - 2)$ is like the rationals, but with an extra symbol $\overline{x}$, which we know should behave like $\sqrt{2}$, namely $\overline{x}^2 = \overline{2}$. We can say that $\mathbb{Q}[x]/(x^2 - 2) \cong \mathbb{Q}[\sqrt{2}]$, where the latter is to be thought of as a subring of $\mathbb{R}$ (or $\mathbb{C}$).

4. Let $R = \mathbb{R}[x]$ and I be the ideal generated by $x^3 - 1$. Now I'm going to stop writing the bars on top of everything, for simplicity. So we think of elements of this ring as polynomials in x, but subject to the relation $x^3 = 1$. What is $(x^4 - x^2)$ times $(x^2 + x + 1)$ in this ring? It doesn't have degree 6, like you'd expect in a normal polynomial ring. First we can simplify the first one to $x - x^2$ using the relation, and then multiply them out to give

   $$(x - x^2)(x^2 + x + 1) = x^3 + x^2 + x - x^4 - x^3 - x^2 = x - x^4,$$

   and since $x^3 = 1$, $x^4 = x$, so this is just zero! So the first thing we notice is that this ring has zero divisors - we just found two of them - $x^4 - x^2$ and $x^2 + x + 1$. There is a more enlightening way to see that their product is zero, however, by diligently factoring everything first:

   $$(x^4 - x^2)(x^2 + x + 1) = x^2(x + 1)(x - 1)(x^2 + x + 1) = x^2(x + 1)(x^3 - 1).$$

This shows that their product is a multiple of x^3-1, hence lies in I. So $(x^4-x^2)(x^2+x+1)$ and x^3-1 are equivalent, hence the coset $[(x^4-x^2)(x^2+x+1)]$ is the same as the coset $[x^3-1]$, which is just I, and we know that this functions as the zero element in the quotient ring.

Notice that the zerodivisor phenomenon basically happened because the generator for I factored: $x^3-1=(x-1)(x^2+x+1)$. That means right away that the two polynomials $x-1$ and x^2+x+1 are zerodivisors. We'll come back to this when we discuss prime ideals.

5. Here's a cool example that shows how you can do calculus without limits. Start with the ring $\mathbb{R}[x]$, and adjoin a new element ϵ, giving the ring $\mathbb{R}[x][\epsilon]=\mathbb{R}[x,\epsilon]$. Finally, take the quotient by the ideal (ϵ^2), this forces ϵ to satisfy the relation $\epsilon^2=0$. Let $R=\mathbb{R}[x,\epsilon]/(\epsilon^2)$ be this new ring. We can compute derivatives in this ring as follows: pick a polynomial whose derivative you want to compute, say $f(x)=x^3$. Now look at

$$f(x+\epsilon)=(x+\epsilon)^3=x^3+3x^2\epsilon+3x\epsilon^2+\epsilon^3=x^3+3x^2\epsilon.$$

All the higher ϵ terms disappear because of the relation $\epsilon^2=0$, and the coefficient of the remaining ϵ term is just $3x^2$, the derivative! In a HW exercise you will be asked to prove that in general,

$$f(x+\epsilon)=f(x)+f'(x)\epsilon$$

If you work instead with the relation $\epsilon^n=0$, when you expand $f(x+\epsilon)$, you will get the first $n+1$ terms of the Taylor series for f.

4.4 The Isomorphism Theorem

This section contains an extremely powerful result, which I remember as saying "image = source/kernel". This is entirely analogous to the isomorphism theorem we saw for groups. Recall that the **image** of a map of rings $\phi\colon R\to S$ is the subring

$$\text{im }\phi=\{s\in S\mid \text{ there exists an } r\in R \text{ such that } \phi(r)=s\}.$$

Proposition 4.4.1. *Let $\phi\colon R\to S$ be a homomorphism of rings, and let $I\subset R$ be its kernel. Then* $\text{im }\phi\cong R/I$.

Proof. As usual we have to define an isomorphism $f\colon R/I\to \text{im }\phi$. Let $\overline{r}\in R/I$ be a coset, and define $f(\overline{r})=\phi(r)$. Since we're dealing with cosets, there may be many different representatives for the same coset, which means we have to check it's well-defined. So suppose $\overline{r}=\overline{r}'$, which is to say that r and r' both represent the same coset. We need to check that $f(\overline{r})=f(\overline{r}')$. Since $\overline{r}=\overline{r}'$, r and r' are equivalent, so their difference lives in I, which means that $\phi(r-r')=0$ (since I is the kernel of ϕ). But then $f(\overline{r})-f(\overline{r}')=\phi(r)-\phi(r')=\phi(r-r')=0$ (using that ϕ is a homomorphism). This shows that $f(\overline{r})=f(\overline{r}')$, so f is well-defined.

Next we check f is a homomorphism. Let $\overline{a},\overline{b}$ be two elements of R/I.
1. Additivity. $f(\overline{a}+\overline{b})=f(\overline{a+b})=\phi(a+b)=\phi(a)+\phi(b)=f(\overline{a})+f(\overline{b})$.
2. Multiplicativity. $f(\overline{a}\cdot\overline{b})=f(\overline{ab})=\phi(ab)=\phi(a)\phi(b)=f(\overline{a})f(\overline{b})$.
3. Sends 1 to 1. $f(\overline{1})=\phi(1)=1$.

Notice that each of the steps only used the rules for addition and multiplication in the quotient, and the fact that ϕ is a homomorphism. So it's easy, as long as you know how to set it up.

Finally we check that f is injective and surjective. For injectivity, suppose $f(\overline{r})=0$. We need to show that $\overline{r}=0$ in R/I. Since $f(\overline{r})=0$, $\phi(r)=0$, by the definition of f, so r is in the kernel of ϕ, which is I. But this means that $\overline{r}=I$, which is the zero coset in R/I. For surjectivity, pick any s in $\text{im }\phi$. We need to find a coset that maps to s under f. By the definition of $\text{im }\phi$, this means there is an r in R such that $\phi(r)=s$. But then the coset $\overline{r}$ maps to s under f, since $f(\overline{r})=\phi(r)=s$, so f is surjective. □

Example 4.4.2. 1. First we'll use the proposition to give a proof that $\mathbb{R}[x]/(x^2+1) \cong \mathbb{C}$, which we showed in the previous section. The idea is to first give a map from $\mathbb{R}[x]$ to $\mathbb{C}$, find its kernel, and apply the theorem. The map we'll use is $\phi\colon \mathbb{R}[x] \to \mathbb{C}$ given by $p(x) \mapsto p(i)$. Like we did before, but now x really means x, and not the coset $[x]$. It's surjective, because for any $a+bi$ in $\mathbb{C}$, we have $\phi(a+bx) = a+bi$; so $\Im\phi = \mathbb{C}$. To find the kernel, suppose $\phi(p) = 0$. Then $p(i) = 0$, so p has i as a root; since it's a real polynomial, it must also have $-i$ as a root, too (complex roots of real polynomials always come in conjugate pairs...). This means that if we factor p over $\mathbb{C}$, it has factors $x-i$ and $x+1$, so over $\mathbb{R}$, it has a factor x^2+1. Thus p is in the ideal generated by x^2+1. Conversely, every element of the ideal (x^2+1) is in kernel, so this shows $\ker\phi = (x^2+1)$. Applying the theorem then gives $R[x]/(x^2+1) \cong \mathbb{C}$.

2. For an easier example, consider the ring $\mathbb{R}[x]/(x-1)$. We know that the quotient here basically means "set x equal to 1", and since 1 is already in $\mathbb{R}$, this quotient ring should just be $\mathbb{R}$. Let's use the proposition to prove this rigorously. The philosophy behind using this proposition is to find a map whose kernel is the thing you're taking the quotient by (in this case the ideal $(x-1)$). So define $\phi\colon \mathbb{R}[x] \to \mathbb{R}$ by sending x to 1; that is to say, $\phi(p(x)) = p(1)$ (so ϕ is just the evaluation map Ev_1). Next we compute its kernel and image. The image is all of $\mathbb{R}$ since for any a in $\mathbb{R}$, $\phi(ax) = a$. To find the kernel, note that $p(x)$ is in the kernel if and only if $p(1) = 0$, that is, if and only if 1 is a root of p. Since roots correspond to linear factors, this happens if and only if $p(x) = (x-1)q(x)$ for some q in $\mathbb{R}[x]$ (in other words, we can factor out at least one copy of $(x-1)$ from p). But this exactly means that p is in the ideal $(x-1)$, so this ideal must be the kernel. Then the proposition says that

$$\mathbb{R} = \operatorname{im}\phi \cong \mathbb{R}[x]/\ker\phi = \mathbb{R}[x]/(x-1)$$

3. Let's see what $\mathbb{R}[x,y]/(xy)$ is isomorphic to. We'll first define a homomorphism from $\mathbb{R}[x,y]$ to $\mathbb{R}[x] \times \mathbb{R}[y]$, and then check its image and kernel. The map we'll use, let's call it f, is:

$$f(p(x,y)) = (p(x,0), p(0,y))$$

First, check it makes sense: the output is an ordered pair, and in the first slot we've substituted $y = 0$, giving a polynomial in x only; in the second slot we've substituted $x = 0$, giving a polynomial in y. Thus the output is indeed an element of $\mathbb{R}[x] \times \mathbb{R}[y]$. Let's first find the image. Notice that both terms in the outputs come from the same polynomial p. So if we set $x = 0$ in the first slot, and $y = 0$ in the second slot, we get the same result, namely $p(0,0)$. So elements of the image are pairs $(p(x), q(y))$ which satisfy $p(0) = q(0)$. Let's check that this is enough to describe the image. Specifically, if we have a pair satisfying $p(0) = q(0)$, is that enough to guarantee it's in the image? Yes, because such a pair occurs as an output under f of the polynomial $p(x) + q(y) - p(0)$ (weird, but it works...). Second, let's determine the kernel. It had better be (xy) or else this map won't tell us anything about the quotient ring we're interested in! Suppose $p(x,y)$ is in the kernel, then $p(x,0) = 0$, and $p(0,y) = 0$. The first condition says that when you set y to 0, it kills the polynomial, so y divides $p(x,y)$; similarly x divides $p(x,y)$. Combining these, we find that xy divides the kernel, hence $\ker f = (xy)$ (the other inclusion is clear since $f(xy) = (0,0)$). Thus applying the isomorphism theorem above gives

$$\mathbb{R}[x,y]/(xy) \cong \{(p(x), q(y)) \in \mathbb{R}[x] \times \mathbb{R}[y] \,\big|\, p(0) = q(0)\}.$$

This has a nice geometric interpretation: elements of $\mathbb{R}[x,y]/(xy)$ are functions on the zero set of xy, which is just the union of x- and y-axes (the "cross"). To give a function on the cross, we just have to give a function on the x-axis, and another function on the y-axis, and they can be whatever we like, as long as they match up at the origin. That's exactly what the right hand side of the isomorphism describes!

5 Factorization; Reducibility; Roots

Previous lecture - Next lecture

This lecture is a comparison of some structural properties relating to factorization in the ring of integers and the ring of polynomials in one variable.

First we recall mostly without proof the basic multiplicative properties of $\mathbb{Z}$.

5.1 Division and Factorization in $\mathbb{Z}$

Here we simply restate various familiar facts about division in $\mathbb{Z}$.

Proposition 5.1.1. *If a and b are two integers, with $b \neq 0$, then there exist unique integers q and r, with $0 \leq r < |b|$, such that $a = bq + r$*

We say that b **divides** a if $a = bk$ for some integer k, and this happens if and only if $r = 0$ in the proposition. Using this, one can prove that any pair of nonzero integers a, b has a gcd, and that the gcd can be expressed as an integer linear combination of a and b, in particular

Proposition 5.1.2. *If a and b are coprime, then there exist integers x and y such that $ax + by = 1$.*

The definition of a prime integer is an integer greater than one which has no factors except 1 and itself. The above proposition shows that

Proposition 5.1.3. *If p is prime and p divides ab, then $p|a$ or $p|b$.*

We will see that it is this property of primes which generalizes well to more abstract settings (in particular, it suggests a notion of prime ideals).

Proof. Assume that $p|ab$. We show that if $p \not| a$, it forces $p|b$. So assume that $p \not| a$. Then p and a are coprime (since p has no nontrivial factors), so we can find x, y with $px + ay = 1$. Multiply by b to give $pbx + aby = b$. Since p divides both terms on the left, it divides b. □

Finally, the most famous result about primes is that they are "multiplicative building blocks" for $\mathbb{Z}$.

Proposition 5.1.4 (Fundamental Theorem of Arithmetic)**.** *If n is any integer, then n can be written as*

$$n = cp_1 \cdots p_r,$$

where the p_i are primes and $c = \pm 1$. This expression is unique except for the order of the prime factors.

5.2 Division and Factorization of Polynomials

To keep the theory of polynomials manageable, we will be considering mostly polynomials whose coefficients live in $\mathbb{Q}$ or $\mathbb{R}$ or $\mathbb{C}$ (as opposed to $\mathbb{Z}$, where there aren't enough units, or say $\mathbb{Z}/6\mathbb{Z}$, where there are zero divisors). These rings $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$,are called fields:

Definition 5.2.1. A **field** is a ring in which every nonzero element is a unit.

We will denote by k an arbitrary field, but you should keep the examples $k = \mathbb{Q}, \mathbb{R}, \mathbb{C}$ in your mind. In the division algorithm for $\mathbb{Z}$, the remainder r is "small", i.e. less than $|b|$. For this we had to measure the "size" of b using absolute value (since b might be negative). Polynomials also have a measure of size, namely their degree.

Definition 5.2.2. Let $p \in k[x]$ be a *nonzero* polynomial, written as $p(x) = a_n x^n + \cdots + a_1 x + a_0$, where $a_n \neq 0$. The **degree of** p is n, written $\deg p = n$. In other words, the degree of p is the largest exponent of x which appears with nonzero coefficient.

Note that we did not define the degree of the zero polynomial, for technical reasons. Polynomials of degree zero (and also the zero polynomial, whose degree is undefined) are called constant polynomials; degree one, linear; degree two; quadratic; then cubic, quartic, etc.

As for integers, we say f **divides** g in $k[x]$, written $f|g$, if there is an $h \in k[x]$ such that $g = fh$. Then f is said to be a **divisor** or **factor** of g The notion of irreducibility, like primeness for integers, is a sort of minimality condition for the divides relation:

Definition 5.2.3. A polynomial $p \in k[x]$ is **irreducible** if p is non-constant and the only divisors of p which have lower degree are the constant polynomials.

The main results about divisibility and factorization for polynomials in $k[x]$ are the following:

Theorem 5.2.4. *1. If $f, g \in k[x]$ and $g \neq 0$, then there exist polynomials $q, r \in k[x]$ such that*

$$f = gq + r,$$

with either $r = 0$ or $0 \leq \deg r < \deg g$.

2. *If f and g have no nonconstant common factor then there are polynomials r, s such that $fr+gs = 1$.*
3. *If $p \in k[x]$ is irreducible and $p|fg$, then $p|f$ or $p|g$.*
4. *Every nonzero polynomial f can be written as*

$$f = cp_1 \cdots p_r,$$

where c is a unit (a nonzero constant polynomial) and the p_i are irreducible polynomials.

In the factorization in 4, we can even omit the unit term c, since cp_1 will still be irreducible, so we could just rename cp_1 as p_1. However, we will often try to work with **monic** polynomials, i.e., those whose leading coefficient is 1. In this case, if we take all the p_i to be monic, then the unit factor c (which would just be the product of the leading terms of the p_i) appears out front.[1]

Proof. Omitted. Part 1 is basically just long division of polynomials as you learned in high school. Parts 2, 3, and 4 follow from 1 using the same arguments as for integers. □

The urge to factorize polynomials is intimately relating to the arguably more basic urge to solve polynomials, which is basically the origin of all algebra. This can be made precise as follows:

Definition 5.2.5. Let $f \in k[x]$. An element $\alpha \in k$ is called a **root** (or **zero**) of f is $f(\alpha) = 0$ in k. Equivalently, α is a root of f if $f \in \ker \mathrm{Ev}_\alpha$. If K is a larger field containing k, we also say that $\alpha \in K$ is a **root** of f in K if $f(\alpha) = 0$ in K. We sometimes say that f **cancels** α.

Note that it makes sense to evaluate $f(\alpha)$, even when $\alpha \in K \setminus k$, because since $k \subset K$, $k[x] \subset K[x]$, so we can regard f also as an element of $K[x]$ and apply the evaluation map $\mathrm{Ev}_\alpha \colon K[x] \to K[x]$ to f. The next proposition says that finding roots is the same as finding linear factors.

Proposition 5.2.6. *Let $f \in k[x]$. Then f has a root in k if and only if f has a linear factor.*

Proof. First suppose that f has a linear factor, so we can write $f(x) = (x - c)g(x)$ for some $c \in k$, $g \in k[x]$. Then $f(c) = (c - c)g(c) = 0$, so f has the root c in k. Conversely, suppose f has a root c in k. Then $f \in \ker \mathrm{Ev}_c \colon k[x] \to k$. Also, the polynomial $x - c$ is in $\ker f$. Now apply the division algorithm to f and $x - c$, giving $f = (x - c)q + r$, where $r = 0$ or $\deg r < \deg(x - c) = 1$, so r must be a constant (either it's 0 or it has degree zero, meaning it's a nonzero constant). Call this constant a. Then $f(x) = (x - c)q(x) + a$. Plugging in $x = c$ gives $0 = f(c) = (c - c)q(c) + a = 0 + a = a$, so $a = 0$. Thus $f = (x - c)q$, so f has a linear factor.

□

Now let's see how these factorizations look over various fields of interest.

[1] On a related note, you might notice that, unlike in $\mathbb{Z}$, here we made no claim about the uniqueness of the factorization. This is because there are lots of units in k, so we can always "tweak" the factorization, by say multiplying p_1 by a unit and then multiplying p_2 by its inverse. One can formulate a uniqueness claim roughly as: the factorization is unique up to the ordering of the irreducibles and multiplication by units. We won't need any uniqueness results in our analysis, so I've not made any such statement in the theorem.

Example 5.2.7. 1. Over $\mathbb{C}$, every nonconstant polynomial has a root[2]. So it has a linear factor. Proceeding inductively, this shows that every complex polynomial factors into linear factors, i.e. every nonzero $f \in \mathbb{C}[x]$ can be written as $f(x) = c(x-a_1)^{r_1}\cdots(x-a_n)^{r_n}$ for some $c \neq 0$ and a_i in $\mathbb{C}$, and $r_i \in \mathbb{N}$. If some exponent $r_i > 1$, a_i is called a **multiple root**, and r_i its **multiplicity**. If $r_i = 1$, a_i is called a **simple root**. In other words, the irreducible polynomials over $\mathbb{C}$ are the linear ones.

2. Over $\mathbb{R}$ things are slightly more complicated. There are two types of irreducibles in $\mathbb{R}[x]$: linear ones and irreducible quadratics. You can tell whether a quadratic ax^2+bx+c is irreducible by looking at the **discriminant** b^2-4ac. If $b^2-4ac<0$, it's irreducible, and vice versa. The way to see that these are the only irreducibles is by taking a real polynomial $f \in \mathbb{R}[x]$, and viewing it temporarily as an element of $\mathbb{C}[x]$ (which contains $\mathbb{R}[x]$). It factorizes completely into linear factors over $\mathbb{C}$, and it's easy to check that the non-real roots come in complex conjugate pairs $a_i, \overline{a_i}$. For each such pair we obtain a real irreducible quadratic, because

$$(x-a)(x-\overline{a}) = x^2 - (a+\overline{a})x + |a|^2,$$

and you can check that the discriminant of this thing is negative.

3. Over $\mathbb{Q}$ things are even worse! Not only are there linear and quadratic irreducibles, like in $\mathbb{R}$, but in fact there are irreducibles of *every* degree! For example, if p is prime, the polynomial $x^{p-1}+x^{p-2}+\cdots+x+1$ is irreducible. This is a very interesting polynomial, called a **cyclotomic polynomial**, which can be thought of as $\dfrac{x^p-1}{x-1}$. It's roots are the pth roots of unity (except 1 itself).

4. Irreducibility is fussy over $\mathbb{Z}$ as well. For example, the polynomial $3x-3$ is irreducible over $\mathbb{Q}$, because even though we can factor it as $3(x-1)$, 3 is a unit in $\mathbb{Q}$. But over $\mathbb{Z}$ 3 is not a unit, so $3x-3$ is reducible. Thus the issue of (ir)reducibility is complicated by the existence of fewer units. You will explore the relation between reducibility over $\mathbb{Z}$ and $\mathbb{Q}$ using Gauss' Lemma in a HW problem.

Luckily there is a useful result called Eisenstein's Criterion that helps allows us to identify some of the irreducible polynomials over $\mathbb{Q}$:

Theorem 5.2.8 (Eisenstein's Criterion)**.** *Let $f(x) = a_nx^n+\cdots+a_1x+a_0 \in \mathbb{Z}[x]$. If there is a prime p such that $p^2 \nmid a_0$, $p|a_i$ for $i = 0,\dots,n-1$ and $p \nmid a_n$, then f is irreducible over $\mathbb{Q}$. If $\gcd(a_0,\dots,a_n)=1$, then f is also irreducible over $\mathbb{Z}$.*

Note: we can interpret the conditions on the coefficients in terms of reduction mod p, as follows: letting $\overline{f}$ be the reduction of f mod p (or mod p^2), i.e., the image of f under the map $\mathbb{Z}[x] \to \mathbb{Z}/p\mathbb{Z}[x]$ (or $\mathbb{Z}[x] \to \mathbb{Z}/p^2\mathbb{Z}[x]$), the conditions say that $\overline{f} = \overline{a_n}x^n$, with $\overline{a_n} \neq 0$ mod p, and $\overline{f} = \overline{a_n}x^n + \overline{a_0}$ mod p^2, with $\overline{a_n} \neq 0$ and $\overline{a_0}$ a multiple of p mod p^2.

Proof. Suppose the conditions of the theorem hold, and assume for a contradiction that f is reducible in $\mathbb{Q}[x]$, so that $f = gh$ for some non-unit $g, h \in \mathbb{Q}[x]$. By a corollary of Gauss' Lemma (see HW6), we find that f is reducible in $\mathbb{Z}[x]$ as well. Then denoting by $\overline{f}$, $\overline{g}$, $\overline{h}$ the reductions mod p of these polynomials, we have $\overline{g}\,\overline{h} = \overline{f} = \overline{a_n}x^n$, which means that $\overline{g} = \overline{a}x^k$ and $\overline{h} = bx^{n-k}$. This shows that all the other coefficients of g and h are zero mod p, so they're all divisible by p. But since their constant terms are both divisible by p, the constant term of f is divisible by p^2, which contradicts the hypothesis of the theorem.

□

Example 5.2.9. 1. The polynomial $f(x) = 2x^3+6x^2+12x+6$ is irreducible over $\mathbb{Q}$, since the prime 3 divides all but the leading coefficient, but 3^2 does not divide the constant term. Yet it is not irreducible over $\mathbb{Z}$ because we can factorize it as $2(x^3+3x^2+4x+3)$, and this is a nontrivial factorization since 2 is not a unit in $\mathbb{Z}$.

[2]This highly nontrivial statement is called the **Fundamental Theorem of Algebra**; we won't prove it in this course, but will use it when dealing with examples over $\mathbb{C}$.

2. The theorem does not apply to the polynomial $x^2 + 5x + 6$, since there is no prime which divides the lowest two coefficients. However, this polynomial is nevertheless reducible, since it factorizes as $(x + 2)(x + 3)$.

3. The theorem does not apply either to the polynomial $2x^2 + x + 3$, but this one happens to be irreducible, since its discriminant is $1^2 - 4 \cdot 2 \cdot 3 = -11$. This means it has two complex roots, so cannot have any integer roots. These two examples show that if we cannot find a prime p as in the statement, no conclusion whatsoever can be drawn about the irreducibility of the polynomial.

4. Here is a famous application. We prove that the **cyclotomic polynomial** $\Phi_p(x) = x^{p-1} + \cdots + x + 1$ (where p is some prime) mentioned above is irreducible in $\mathbb{Z}[x]$. First off, a trick: $\Phi_p(x)$ is irreducible if and only if $\Phi_p(x+1)$ is (reason: $f(x) \mapsto f(x+1)$ is an automorphism of rings). But since $\Phi_p(x) = \dfrac{x^p - 1}{x - 1}$, so replacing x by $x + 1$ we have

$$\begin{aligned}\Phi_p(x+1) &= \frac{1}{x}\left((x+1)^p - 1\right) \\ &= \frac{1}{x}\left(x^p + px^{p-1} + \binom{p}{2}x^{p-2} + \cdots + px\right) \\ &= x^{p-1} + px^{p-2} + \cdots p\end{aligned}$$

   Thus Eisenstein's criterion applies and shows the irreducibility of $\Phi_p(x + 1)$, hence of $\Phi_p(x)$.

6 Special Classes of Rings

Previous lecture - Next lecture

The previous lecture illustrated some striking similarities between the seemingly very different rings $\mathbb{Z}$ and $k[x]$. Today we introduce some terminology to describe these similarities. We define various classes of rings, and investigate the relationships between them.

6.1 Fields

We have seen fields already, but I include them again here for completeness.

Definition 6.1.1. A **field** is a nonzero ring in which every nonzero element is a unit. As before, we use the letter k throughout to denote an arbitrary field.

For example, $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are fields, while $\mathbb{Z}$ is not. You proved in HW5 that $\mathbb{Q}[\sqrt{2}]$ is a field. $\mathbb{Z}/5$ is a field, while $\mathbb{Z}/6$ is not. In general, $\mathbb{Z}/n/Z$ is a field if and only if n is a prime (you should prove this as an exercise - it follows from an analysis of the units and zero divisors in $\mathbb{Z}/n\mathbb{Z}$).

Proposition 6.1.2. *Let R be a nonzero ring. R is a field if and only if it has exactly two ideals (which must be the zero and unit ideals).*

6.2 Integral Domains ("Domains")

Definition 6.2.1. A nonzero ring is called an **integral domain** (or usually simply a **domain**) if it contains no zerodivisors.

Examples

1. $\mathbb{Z}$ is a domain.
2. $\mathbb{Z}/n\mathbb{Z}$ is not a domain when n is composite. It is a domain when n is prime.
3. $\mathbb{Q}[x]/(x^2-1)$ is not a domain, but $\mathbb{Q}[x]/(x^2+1)$ is.

6.3 Principal Ideal Domains

Definition 6.3.1. An integral domain is called a **principal ideal domain** (PID for short) if every ideal in it is principal (can be generated by a single element).

Examples

1. $\mathbb{Z}$ is a principal ideal domain. Question: what about the ideal $(2, 5)$ generated by 2 and 5? It has two generators, so it doesn't seem like it's a principal ideal... Answer: since 5-2-2=1 is also in this ideal, the ideal is actually the unit ideal, generated by 1, hence principal. Some ideals may not seem principal, even though they actually are.
2. $k[x]$ is a principal ideal domain.
3. $\mathbb{Z}[x]$ is *not* a principal ideal domain: the ideal $(2, x)$ cannot be generated by one element.
4. $\mathbb{C}[x, y]$ is *not* a principal ideal domain: the ideal (x, y) cannot be generated by one element.

6.4 Euclidean Domains

This one is more complicated. A Euclidean domain is one where there is a division algorithm. To define such an algorithm, we need to equip the ring with some measure of size to be able to say that the remainder upon division is smaller than the divisor.

Definition 6.4.1. An integral domain R is called a **Euclidean domain** if there is a **norm** on R, which is a function n from $R \setminus \{0\}$ to $\mathbb{N}$, which satisfies

1. For any f, g in R, with g nonzero, there exist q, r in R such that $f = qg + r$, and either $r = 0$ or $n(r) < n(g)$.
2. For any nonzero f, g in R, $n(f) \leq n(fg)$.

The definition says, in simple terms, that any element f can be divided by a nonzero element g, giving a remainder r that is smaller than the thing you divided by (g). The second property says that when you multiply elements, the norms get bigger (or stay the same).

Example 6.4.2.
1. $\mathbb{Z}$ is a Euclidean domain. The norm function is just the absolute value. The fact that it satisfies the two properties is clear.
2. $\mathbb{Q}[x]$ is a Euclidean domain. The norm function is the **degree**: the highest power of x appearing in the polynomial. Just as for integers, we can do long division with polynomials, and the remainder always has degree strictly less than the thing we divided by. So property 1 holds. Property 2 holds, and in fact $n(f) = n(g)$ if and only if they differ by a unit (recall that the units in the this ring are just the nonzero constant polynomials).
3. The Gaussian integers $\mathbb{Z}[i]$ are a Euclidean domain with norm function given by $n(a+bi) = a^2+b^2$.
4. It can be quite hard to decide whether a given ring is a Euclidean domain or not. For how we can we prove that *there does not exist* a norm function? Maybe there is one, but we weren't smart enough to find it! Anyways, $\mathbb{Z}[x]$ is not a Euclidean domain. The easiest way to see this will come soon, when we prove that every Euclidean domain is a PID. Since we know that $\mathbb{Z}[x]$ isn't a PID, it cannot be a Euclidean domain, either.
5. If R is a Euclidean domain, and I an ideal of R, can the quotient ring R/I be made into a Euclidean domain? The obvious guess for a norm function would be this: $n([r]) = n_R(r)$, where n_R is the norm on R. But is it well-defined? Probably not, unless all the elements in our ideal have the same norm.

6.5 Unique Factorization Domains

Unique factorization domains (or UFDs, for short) are another class of very nice rings, again modeled on the integers and polynomials, in which we can factor things into irreducible pieces. We have defined irreducibility for polynomials, but the definition can easily be extended to more general rings:

Definition 6.5.1. A nonzero element r of a ring R is called **irreducible** if it is not a unit and the only way to factor $r = ab$ is by taking either a or b to be a unit.

Thus an irreducible element is one which is "unfactorable", except for "trivial" factorizations obtained by pulling out units. For example, prime numbers in $\mathbb{Z}$, or linear polynomials in $\mathbb{C}[x]$. Of course, in $\mathbb{R}[x]$, there are irreducible quadratics, too, such as $x^2 + 1$.

Definition 6.5.2. An integral domain R is called a **unique factorization domain** if every nonzero element that is not a unit can be written as a product of finitely many irreducible elements, and this factorization is unique, meaning that given any two factorizations $r = a_1 \cdots a_k = b_1 \cdots b_m$, k and m must be the same, and after possibly reordering the factors, each a_i is a unit times b_i.

Example 6.5.3.
1. $\mathbb{Z}$ is a UFD - this is the statement of the fundamental theorem of arithmetic.
2. $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, and $\mathbb{C}[x]$ are all UFDs ($\mathbb{Z}[x]$ is the hardest one to prove).
3. $\mathbb{Z}/6\mathbb{Z}$ is not a UFD, because for example there are two ways to factor 3: $3 = 3 \cdot 3$ and $3 = 3 \cdot 5$. But there's a simpler reason: $\mathbb{Z}/6\mathbb{Z}$ is not even a domain! 3 is a zero divisor: if you multiply it up by 2, you get 0, so there should be many ways to factor it, by adding copies of 2 to the second factor. This is one reason why we restrict our attention to domains when talking about factorization.
4. The most famous example of a ring that *is* a domain but is *not* a UFD is $\mathbb{Z}[\sqrt{-5}]$, which is a subring of $\mathbb{C}$. In this ring there are two ways to factor 6: $6 = 2 \cdot 3$ and $6 = (1+\sqrt{-5})(1-\sqrt{-5})$. One has to check that $2, 3$ and $1 \pm \sqrt{-5}$ are all irreducible and do not differ by units to make sure that these really count as *distinct* factorizations.

6.6 Relationships Between the Types of Rings

We know establish some containments between these various types of rings. The results are summarized in the following theorem.

Theorem 6.6.1. $\{$ *fields* $\} \subset \{$ *Euclidean domains* $\} \subset \{$ *principal ideal domains* $\}$ $\subset \{$ *unique factorization domains* $\} \subset \{$ *integral domains* $\}$

Proof. 1. **Fields are Euclidean domains** Given any field $\mathbb{F}$, define a norm n on $\mathbb{F}$ by setting $n(a) = 0$ for all $a \neq 0$ in $\mathbb{F}$. This satisfies the division algorithm because we can always divide without even needing remainders (it's a field).

2. **Euclidean domains are PIDs**

   Start with any Euclidean domain R, and pick an arbitrary ideal I in R (can assume it's nonzero since the zero ideal is already principal). We need to show it can be generated by one element. First we have to choose this element. Well, it should be the "smallest" (think about the ideal (3) in $\mathbb{Z}$: the generator is the smallest, since it has to divide all the other multiples). So let f be a nonzero element of I of smallest norm (there may be more than one choice for f, as in using 3 or -3 to generate (3) in $\mathbb{Z}$). How do we know that there *is* an element of smallest norm? Well, look at the set of all norms of all nonzero elements of R. It's a subset of $\mathbb{N} \setminus \{0\}$, so it has a smallest element ($\mathbb{N}$ is "well-ordered").

   Now we've got our putative generator, and we have to show that everything else in I can be written as a multiple of f. So pick any other nonzero g in I, and do the division algorithm on g by f (property (1) of Euclidean domains). This gives a quotient q and remainder r satisfying $g = qf + r$, where r has smaller norm than f (or else is 0). But since f and g are both in the ideal I, and $r = g - qf$, r must also be in I. Since f has smallest norm in I, we cannot have $n(r) < n(f)$, and therefore $r = 0$, which shows that f divides g, hence g is a multiple of f. Since g was arbitrary in I, this shows that $I = (f)$.

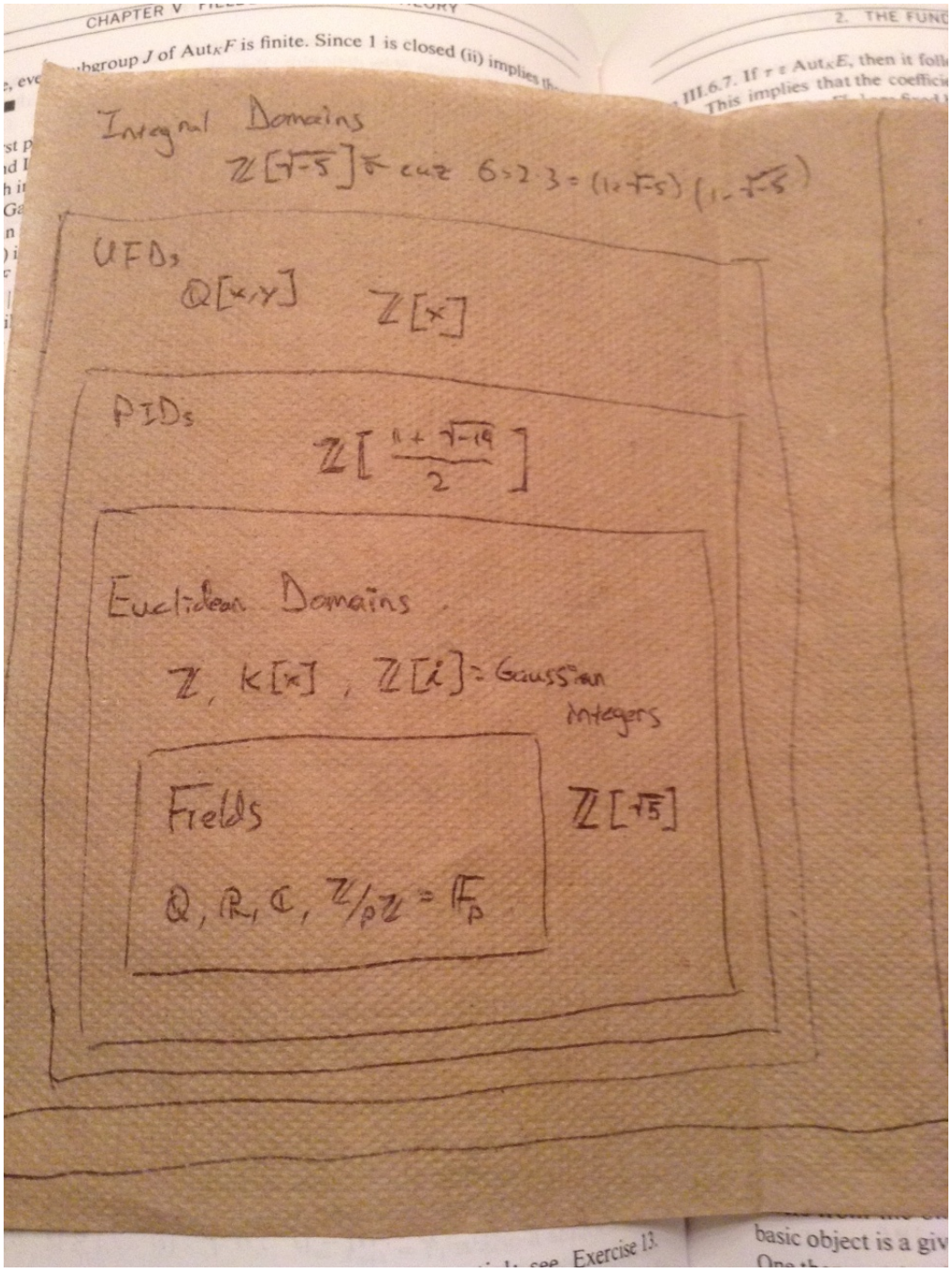
3. **PIDs are UFDs** This is the hardest bit, which I only do part of. We start with a PID R, and we have to show it's a UFD, which breaks into two parts: 1. show that any element can be factorized into finitely many irreducibles, and 2. show that this factorization is unique (in the sense of the definition of UFD). Here's a sketch of step 1: Pick an element r of R. If it's zero or a unit, we've no need to factorize. So assume r is nonzero and not a unit. It's either irreducible, or it's not. If irreducible, we're done. If not, it has two proper non-unit factors. Repeat the process for these factors. Keep repeating this process until you reach irreducible elements. Problem: what if the process doesn't stop?[1] To see that it stops, suppose for contradiction that a certain element could be factored indefinitely: in this case we would be able to produce a chain of distinct ideals $(a_1) \subset (a_2) \subset \cdots$ (by taking as generator, one of the new factors in each successive factorization). Then if we take the union of all these ideals, it's (a) still an ideal (check!) and (b) principal (we're in a PID). Then you'd have to check that its generator is equal to one of the a_i (up to a unit). But it would also be equal to a_{i+1}, etc, (up to units), and we supposed the ideals were all distinct - contradiction. This shows that in a PID, the process of breaking off irreducibles stops after finitely many steps. For the uniqueness, one makes an inductive argument on the number of irreducibles in a factorization.

4. **UFDs are integral domains** This is built into the definition: UFDs, PIDs, and EDs all have a "D" for domain.

□

[1] This can actually happen: Consider the subring $\mathbb{Z}[\{\sqrt[n]{2}\}_{n\in\mathbb{N}}]$ of $\mathbb{C}$, obtained by adding all nth roots of 2, as n varies from 1 to ∞. Try to factor 2 in this ring. It factors as $2 = (\sqrt{2})^2 = (\sqrt[4]{2})^4 = (\sqrt[8]{2})^8 = \cdots$, and this process of factoring can go on forever, each radical being broken down further and further, whilst never reaching an irreducible. The infinite chain of ideals in this case would be $(\sqrt{2}) \subset (\sqrt[4]{2}) \subset (\sqrt[8]{2}) \subset \cdots$

I conclude this lecture with a picture, patiently etched on a Caffe Strada napkin, showing the inclusions amongst the various rings, along with examples. If a ring lives in a certain box, it's that sort of ring; if not, it's not one of those rings. So, for example, $\mathbb{Z}[x]$ is a UFD but not a PID.



7 Prime and Maximal Ideals

Previous lecture - Next lecture

Today we discuss two very important types of ideals, and learn how to use them to check whether certain quotients are fields or integral domains.

7.1 Prime numbers and prime ideals in $\mathbb{Z}$

The crucial connection between (principal) ideals and elements is this: **divides corresponds to contains**. This means: if $f|g$ in a ring (meaning there is an h such that $g = fh$), then $g \in (f)$. If f divides g, then g is in the ideal generated by f.

What does this mean in the simple case of the ring $\mathbb{Z}$? Prime numbers are ones that have no proper divisors. Another property, often taken as the definition, is that if a prime p divides a product, then it divides one of the factors: if $p|ab$, then $p|a$ or $p|b$. This is the property we want to generalize to ideals. So let's translate the statement. $p|ab$ means $ab \in (p)$, and similarly $p|a$ means $a \in (p)$, etc. So in terms of ideals, p is a prime number means that if ab is in (p), then a or b (or both) must be in (p). So for an ideal in $\mathbb{Z}$ generated by a prime number, we have the following slogan: **if a product is in it, one of the factors must be, also**. This is basically the definition of a prime ideal, and it makes sense in any ring.

7.2 Prime Ideals

Definition 7.2.1. Let I be a proper ideal in a ring R. We say I is a **prime ideal** if whenever ab is in I, either a or b (or both) is in I.

Examples

1. In $\mathbb{Z}$, an ideal (n) is prime if and only if the integer $|n|$ is prime (I put the absolute value since primes are required to be positive, but the generator may not be), or $n = 0$.
2. More generally, in any PID, a nonzero ideal is prime if and only if it is generated by an irreducible element.
3. In the ring $\mathbb{Z}$, the zero ideal is prime, but in the ring $\mathbb{Z}/6$, the zero ideal is not prime, since $2 \cdot 3 \in (0)$ but $2 \notin (0)$ and $3 \notin (0)$.

7.3 Maximal Ideals

A maximal ideal is what the name suggests: the biggest possible ideal. But that would be the entire ring, and there would be only one. So we require maximal ideals to be proper, and then it turns out that there can be many of them.

Definition 7.3.1. An ideal $\mathfrak{m}$ in a ring R is called **maximal** if it is not the unit ideal and there are no other ideals I such that $\mathfrak{m} \subsetneq I \subsetneq R$.

Examples

1. In $\mathbb{Z}$ and $\mathbb{R}[x]$, every nonzero prime ideal is maximal.
2. In $\mathbb{R}[x, y]$, (x, y) is maximal, (x) is prime but not maximal, and (x^2) is neither prime nor maximal.
3. In $\mathbb{C}[x, y]$, any ideal of the form $(x - a, y - b)$ (where $a, b \in \mathbb{C}$) is maximal.[1]
4. In $\mathbb{Z}[x]$, (2) is prime but not maximal. $(2, x)$ is maximal. $(4, x)$ is neither prime nor maximal.

The relationship between prime and maximal ideals is as follows:

[1] Notice that these maximal ideals correspond bijectively to points in $\mathbb{C}^2$. This is the beginning of algebraic geometry, and is one version of a theorem called **Hilbert's Nullstellensatz**.

Proposition 7.3.2. *Any maximal ideal is prime.*

This proposition follows immediately from the results of the next section, but let's prove it directly for some practice working with the definitions.

Proof. Let $\mathfrak{m}$ be a maximal ideal. To show it's prime, assume $a, b \in R$, with $ab \in \mathfrak{m}$ and $a \notin \mathfrak{m}$. We must show that $b \in \mathfrak{m}$. Since $a \notin \mathfrak{m}$, the ideal sum $\mathfrak{m} + (a)$ is strictly larger than $\mathfrak{m}$, and since $\mathfrak{m}$ is maximal, it must be equal to the unit ideal R. So $1 \in \mathfrak{m} + (a)$, which means we can write $1 = x + ra$ for some $x \in \mathfrak{m}$ and $r \in R$. Then $b = xb + rab$, and since both x and ab are in $\mathfrak{m}$, this shows that $b \in \mathfrak{m}$. □

That the converse is not true is demonstrated by example 4 in the previous section: the ideal (2) is prime but not maximal. It's not maximal because it's contained in the larger (proper) ideal $(2, x)$. But it's prime because if $fg \in (2)$, then fg has only even coefficients. But an analysis of the formula for the coefficients of a product shows inductively that either f or g must then have only even coefficients, hence be in (2). Notice that the ring in which we found such a counterexample, $\mathbb{Z}[x]$, is not a PID. This is no coincidence:

Proposition 7.3.3. *Let R be a principal ideal domain, and $I = (a)$ a nonzero ideal. Then I is prime if and only if I is maximal if and only if a is an irreducible element.*

Proof. For the first equivalence, we need only show that prime implies maximal. So assume I is prime, and suppose J is another ideal properly containing I. We'll show that J must be the unit ideal, and this will show that I is maximal. Write J as $J = (b)$ (since we're in a PID). Since J properly contains I, $b \notin J$. But since $(a) \subset (b)$, $\in (b)$, so $a = rb$ for some r. Since $a = rb \in I$, and I is prime, $r \in I$ or $b \in I$; but we said $b \notin I$, so it must be that $r \in I$. Thus we can write $r = sa$, so $a = rb = sab$, hence $a - sab = a(1 - sb) = 0$. Since we're in a domain and $a \neq 0$, $1 - sb = 0$, so $sb = 1$ hence b is a unit so J is the unit ideal.

For the second equivalence, first assume $I = (a)$ is maximal, and suppose for a contradiction that a is reducible, say $a = bc$, with b, c nonunits. Then $(a) \subset (b) \neq R$, which contradicts maximality of I; so a must be irreducible. Conversely, assume a is irreducible. Then for any b, b properly divides a implies b is a unit. In terms of ideals this says that $(a) \subset (b)$ (proper inclusion) implies $(b) = R$. Since ideals of the form (b) are the only ideals in R, this exactly says that (a) is maximal.

□

7.4 Relations between these ideals and their quotients

Now we prove a few useful things about these ideals.

Before proving the next very useful proposition, we need a basic result about ideals in quotient rings. Recall that the **canonical homomorphism** $\pi\colon R \to R/I$ from a ring to its quotient is the map sending x to the coset $\overline{x}$. Note that π is surjective.

Proposition 7.4.1. *Let R be a ring and I an ideal of R. Then the ideals of R/I are in bijection with the ideals of R which contain I.*

Proof. First pick an ideal $\overline{J}$ in R/I. Its pre-image $J = \pi^{-1}(\overline{J})$ is an ideal in R, since pre-images of ideals are again ideals. Moreover, it contains I since everything in I goes to zero, and zero is in $\overline{J}$. Conversely, suppose given an ideal J of R which contains I. Then we claim[2] that $\pi(J)$ is an ideal in R/I. We know that it's an additive subgroup, since it's the image of a group homomorphism. It's closed under scaling: let $\overline{x} \in \pi(J)$, for some $x \in J$, and $\overline{y}$ be any other element of R/I. Since π is surjective, write $\overline{y} = \pi(y)$. Then $xy \in J$ (since it's an ideal), and $\pi(xy) = \overline{xy} = \overline{x}\,\overline{y}$, so $\overline{x}\,\overline{y}$ is in $\pi(J)$.

To establish the bijection, we must check that $\pi(\pi^{-1}(\overline{J})) = \overline{J}$ and $\pi^{-1}\pi(J) = J$. I'll leave this to you...

□

Proposition 7.4.2. *Let R be a ring.*

[2] This is special - usually images of ideals are not ideals.

1. *An ideal $\mathfrak{p}$ is prime if and only if $R/\mathfrak{p}$ is an integral domain.*

2. *An ideal $\mathfrak{m}$ is maximal if and only if $R/\mathfrak{m}$ is a field.*

Proof. 1. The condition for $\mathfrak{p}$ to be prime is "$ab \in \mathfrak{p}$ implies a or b is in $\mathfrak{p}$". In terms of the quotient ring $R/\mathfrak{p}$, this is exactly the same as saying "$\overline{a}\overline{b} = 0$ implies $\overline{a}$ or $\overline{b}$ is zero." This is exactly the statement that there cannot be zerodivisors in $R/\mathfrak{p}$.

2. If $R/\mathfrak{m}$ is a field then the only ideals are (0) and $R/\mathfrak{m}$. So the only ideals in R containing $\mathfrak{m}$ are their two pre images, namely $\mathfrak{m}$ and R. So $\mathfrak{m}$ is maximal. Conversely, if $\mathfrak{m}$ is maximal, then there are no ideals in $R/\mathfrak{m}$ containing (0) besides $R/\mathfrak{m}$ itself. This means (0) is the unique maximal ideal of $R/\mathfrak{m}$, which implies that everything outside (0) is a unit (HW exercise), hence $R/\mathfrak{m}$ is a field. Alternatively, this implies $R/\mathfrak{m}$ has only two ideals, 0 and $R/\mathfrak{m}$ itself, so 'tis a field by 6.1.2. □

8 Field Extensions

Previous lecture - Next lecture

We now begin a study of fields which will lead us into Galois theory, one of the most famous successes of algebra. We will see that certain fields "sit inside" each other (an extension of fields) in such a way that the fields "sitting in-between" them correspond to subgroups of a certain group, the Galois group of the field extension . Thus one can use some ideas from group theory to obtain information about these fields.

8.1 Field Extensions and Examples

We begin with a simple but possibly surprising observation which makes fields rather interesting rings.

Proposition 8.1.1. *If K and L are fields, then any homomorphism between them must be injective.*

Proof. Let $f\colon K \to L$ be a homomorphism. Then $\ker f$ is an ideal of K, and there are only two such: the zero and unit ideals. $\ker f$ cannot be the unit ideal, or else f would be the zero map, which is not a homomorphism ($f(1) = 1$ for homomorphisms, and $1 \neq 0$ in a field). Thus $\ker f$ is trivial, so f is injective. □

Definition 8.1.2. A **field extension** is an injective map $K \to L$ of fields. Equivalently (by identifying K with its image in L, which is a subring), it is a subring K of a field L which is also a field. In this situation, we will also speak of K being a **subfield of** L, or of L as an **extension of** K.

Example 8.1.3.

1. $\mathbb{Q} \subset \mathbb{Q}[\sqrt{2}]$, $\mathbb{Q}[\sqrt{2}] \subset \mathbb{R}$ and $\mathbb{R} \subset \mathbb{C}$ are all field extensions.
2. For p prime, $\mathbb{Z}/p\mathbb{Z}$ is a field with p elements, which we will hereafter denote $\mathbb{F}_p$. The polynomial ring $\mathbb{F}_p[x]$ is not a field, but it is possible to find a polynomial $q(x)$ such that the quotient $\mathbb{F}_p[x]/(q)$ is also a field, in which case $\mathbb{F}_p \subset \mathbb{F}_p[x]/(q)$ is a field extension.
3. If K is a field, and x an indeterminate (or "formal") variable, the set $K(x) = \left\{ \dfrac{p(x)}{q(x)} \;\middle|\; q(x) \neq 0 \right\}$ is a field; it is an extension of k (a very large one, as we will see). We say that $K(x)$ is the field obtained by **adjoining** x to the field K. This is different from the formation of $K[x]$, in which we adjoin the element x to form a new ring (the polynomial ring), which is not a field. The field $K(X)$ is often called the **field of rational functions** over K, to distinguish it from the polynomial ring. It's important to pay attention to the square brackets "[ ] " denoting "ring adjoin" and the parentheses "( )" denoting "field adjoin".
4. By contrast with the previous example, if we define $\mathbb{Q}(\sqrt{2}) = \left\{ \dfrac{p(\sqrt{2})}{q(\sqrt{2})} \;\middle|\; q(\sqrt{2}) \neq 0 \right\}$ (so the "indeterminate" x has been replaced by $\sqrt{2}$), we find that actually $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}[\sqrt{2}]$. Thus in some cases, the notion of "field adjoin" and "ring adjoin" coincide. We will soon see a large class of examples in which this is the case, but it is important to keep in mind, that in general, they are different. This construction is generalized in the following definition.

Definition 8.1.4. Let $K \subset L$ be a field extension. If $\alpha \in L$, we let $K(\alpha)$ be the intersection of all subfields of L containing K and α, and call this the **field extension of** K **generated by** α. It is the smallest subfield of L which contains both K and α. More generally, if A is a subset of L, we define $K(A)$ to be the intersection of all subfields of L which contain both K and A. It is the smallest subfield of L containing K and A.

One should check that this coincides with the definition of $\mathbb{Q}(\sqrt{2})$ given in example 4 above. Essentially it comes down to the following observation: every subfield of L containing K and α must contain all rational functions in α. This is forced by the field axioms. Since $K(\alpha)$ is the smallest, it contains nothing more.

The main advantage of working with fields is that one can apply results from linear algebra, an approach popularized by Emil Artin, who to some extent "rewrote the book" on Galois theory. In particular, whenever we have a field extension $K \subset L$, we may regard L as a vector space over K. In particular, we have a notion of the dimension of L as a K-vector space, which gives an idea of the "size" of the extension.

Definition 8.1.5. Let $K \subset L$ be a field extension. The **degree** of the extension, denoted $[L : K]$ is the dimension of L as a vector space over K. It may be infinite. We say that L is a **finite extension** (resp. **infinite extension**) of K to mean that its degree over K is finite (resp. infinite).

Example 8.1.6.

1. $[\mathbb{Q}[\sqrt{2}] : \mathbb{Q}] = 2$, because the set $\{1, \sqrt{2}\}$ is a basis.
2. $[\mathbb{C} : \mathbb{R}] = 2$, because $\{1, i\}$ is a basis.
3. Let $\omega = e^{\frac{2\pi i}{3}}$, a **cube root of unity**. Then $[\mathbb{Q}(\omega) : \mathbb{Q}] = 2$ because $\{1, \omega\}$ is a basis. Note that $1 + \omega + \omega^2 = 0$ (draw a picture...) so the set $\{1, \omega, \omega^2\}$ is linearly dependent.
4. $[\mathbb{R} : \mathbb{Q}] = \infty$, because, for example, the set $\{\sqrt{p} \mid p \text{ is prime}\}$ is linearly independent (and infinite), and a finite-dimensional space cannot contain an infinite set of linearly independent elements.
5. $K \subset K(x)$ is an infinite extension: the elements $1, x, x^2, \ldots$ are linearly independent.

The next result is sometimes abbreviated by saying that "degree is multiplicative in towers":

Proposition 8.1.7. *Let $K \subset L$ and $L \subset M$ be field extensions. Then in particular $K \subset M$ is also a field extension; it is finite if and only if $K \subset L$ and $L \subset M$ are both finite, in which case the degrees of the three extensions are related by the formula*

$$[M : K] = [M : L][L : K]$$

Proof. Let $[M : L] = m$ and $[L : K] = n$, and pick bases $a_1, \ldots, a_m$ for M over L and $b_1 \ldots, b_n$ for L over K. Then the set $\{a_i b_j \mid i = 1, \ldots, m, j = 1, \ldots, n\}$ is a basis for M over K (check).

□

8.2 Algebraic vs. Transcendental Extensions

In this section we distinguish two important types of field extensions, which exhibit fundamentally different phenomena.

Definition 8.2.1. Let $K \subset L$ be a field extension. An element $\alpha \in L$ is called **algebraic over** K if there is a nonzero polynomial $f \in K[x]$ such that $f(\alpha) = 0$. In this case we say that α **satisfies** or **is a root of** the polynomial f, and that f **cancels** α. If there is no such polynomial, α is called **transcendental over** K. If every element of L is algebraic over K, L is called an **algebraic field extension**. Otherwise it is called a **transcendental field extension**.

Note we have defined two slightly different usages of the terms algebraic and transcendental, depending on whether they describe *elements* or *extensions*.

Example 8.2.2.

1. $\mathbb{Q}[\sqrt{2}]$ is an algebraic extension of $\mathbb{Q}$. Every element $a + b\sqrt{2}$ satisfies the polynomial relation $(a + b\sqrt{2})^2 - 2a(a + b\sqrt{2}) + (a^2 - 2b^2) = 0$, which is a rational polynomial.
2. $\mathbb{Q}[\omega]$ is an algebraic extension of $\mathbb{Q}$, because the new element ω is algebraic over $\mathbb{Q}$; it is a root of $x^3 - 1$.
3. $\mathbb{Q} \subset \mathbb{R}$ is a transcendental extension: for example, π is not algebraic over $\mathbb{Q}$ (this is a difficult result).
4. $\mathbb{C}$ is an algebraic extension of $\mathbb{R}$, essentially just because i satisfies the polynomial $x^2 + 1$. Notice the the transcendence of π (over $\mathbb{Q}$) is not an issue because we are allowed real number coefficients, so for example π satisfies the polynomial $x - \pi$. In fact, one can show that, up to isomorphism, $\mathbb{C}$ is the *only* algebraic extension of $\mathbb{R}$ (excluding the "trivial" extension $\mathbb{R} \subseteq \mathbb{R}$).

5. $K \subset K(X)$ is a transcendental extension: x does not satisfy any nonzero polynomial relation.

6. Looking at the degrees of these examples, you might be tempted to think that the transcendental extensions are precisely those whose degree is infinite, but to see that this is not the case consider the extension $\mathbb{Q}[\{\sqrt{n} \mid n \in \mathbb{N}\}]$. This has infinite degree, because the square roots of all (non-square) are integers are linearly independent, but each of these elements is algebraic over $\mathbb{Q}$.

8.3 Simple Extensions

Simple extensions are those in which the large field can be obtained from the small field by adjoining just one element.

Definition 8.3.1. A field extension $K \subset L$ is a **simple extension** if there is an element $\alpha \in L$ such that $L = K(\alpha)$.

Example 8.3.2. 1. $\mathbb{Q} \subset \mathbb{Q}[\sqrt{2}]$, $\mathbb{Q} \subset \mathbb{Q}[\omega]$, and $\mathbb{R} \subset \mathbb{C}$ are simple extensions.

2. $K \subset K(X)$ is a simple extension.

3. $\mathbb{Q} \subset \mathbb{Q}(\sqrt{2}, \omega)$ does not appear to be a simple extension, but in fact it is. We will see why soon - it turns out you can get of both the "adjoined" elements by simply adjoining the one element $\omega\sqrt{2}$.

4. $\mathbb{Q} \subset \mathbb{R}$ is not a simple extension.

In what follows we will classify all simple extensions. Note that an isomorphism of fields is just an isomorphism of rings in which the two rings involved are actually fields. Also, fix some terminology: a polynomial is **monic** is its leading coefficient is 1.

Theorem 8.3.3 (Classification of Simple Field Extensions)**.** *Let K and L be subfields of some other field F (possibly equal to L), and suppose that $K \subset L$ is a simple extension, so that $L = K(\alpha)$ for some $\alpha \in F$.*

1. *If α is transcendental over K, then $L \cong K(x)$, the field of rational functions.*

2. *If α is algebraic over K, then the following hold:*

   (a) *$K(\alpha) = K[\alpha]$ as subrings of F; in particular, $K[\alpha]$ is itself a field.*

   (b) *There is a monic irreducible polynomial $f \in K[x]$ such that $K(\alpha) \cong K[x]/(f)$. This f is the unique monic polynomial of minimal degree having α as a root.*

   (c) *If $\deg f = n$ (where f is as in (b)), then $[K(\alpha) : K] = n$, and $\{1, \alpha, \alpha^2, \ldots, \alpha^{n-1}\}$ is a basis for $K(\alpha)$ over K.*

The theorem can be roughly summarized as saying that "every simple extension of K is, up to isomorphism, either $K(x)$ or a quotient of $K[x]$". The polynomial f in part 2(b) is referred to as the **minimal polynomial** of α (or of the extension $K(\alpha)$). So (c) says that **the degree of the extension is equal to the degree of the minimal polynomial**. We will use this often in computations.

Proof. 1. Since α is transcendental, for any nonzero polynomial $g \in K[x]$, $g(\alpha) \neq 0$, so we can define a map $\phi\colon K(x) \to K(\alpha)$ by $\dfrac{f(x)}{g(x)} \mapsto \dfrac{f(\alpha)}{g(\alpha)}$. It is surjective, essentially by definition of $K(\alpha)$, and injective because it is a homomorphism between fields (check the homomorphism property yourself...). Thus it's an isomorphism.

2. We prove (a) and (b) together, by considering the map $\phi\colon K[x] \to K[\alpha]$ given by evaluation at α. It is surjective (again, essentially by definition of $K[\alpha]$). It's not injective, since α is algebraic, so there is some nonzero polynomial in $\ker\phi$, hence $\ker\phi \neq 0$. Since $K[x]$ is a PID, $\ker\phi = (f)$ for

some nonzero[1] $f \in K[x]$. Let c be the leading coefficient of f. Since $c \in K$ is a unit, $(f) = (\frac{1}{c}f)$, so replacing f by $\frac{1}{c}f$ we may assume f is monic. Applying the isomorphism theorem gives

$$K[\alpha] = \text{im } \phi \cong K[x]/(f),$$

and since $K[\alpha]$ is an integral domain, (f) is prime, so f is irreducible. Moreover, f, being the generator of $\ker \phi$, has minimal degree out of all those polynomials which have α as a root (can check this with division algorithm, as we have done many times before). It is the unique such because all generators for this ideal differ by a unit, so there is only one which is monic. Finally, since $K[x]$ is a PID, and (f) a nonzero prime ideal, it is also maximal. Therefore $K[x]/(f)$ is a field, so $K[\alpha]$ is a subfield of F containing K and α. But $K(\alpha)$ is the smallest subfield of F containing K and α, so $K(\alpha) \subseteq K[\alpha]$. On the other hand it is clear from the definitions that $K[\alpha] \subseteq K(\alpha)$; thus they're equal. This proves (a) and (b).

To prove (c), note that $1, \alpha, \cdots, \alpha^{n-1}$ are linearly independent, for any dependence amongst them would result in a polynomial of degree less than that of f with α as a root. But f has minimal degree with this property. To show that it spans $K(\alpha)$ (which we now know is equal to $K[\alpha]$), pick any nonzero element $g(\alpha) \in K[\alpha]$; we wish to write it as a linear combination of $1, \alpha, \ldots, \alpha^{n-1}$. Apply the division algorithm to g by f to get $g = fq + r$ with $r = 0$ or $\deg r < \deg f$. Plugging in α gives $g(\alpha) = f(\alpha)q(\alpha) + r(\alpha) = r(\alpha)$. Then $r \neq 0$, or else $g(\alpha) = 0$, so $\deg r < n$, so we have $g(\alpha) = a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}$, as desired. Thus $\{1, \alpha, \ldots, \alpha^{n-1}\}$ is a basis, hence $[K(\alpha) : K] = \deg f = n$.

□

The next simple corollary gives an alternative characterization of the minimal polynomial, in terms of irreducibility.

Corollary 8.3.4. *The monic polynomial f of a simple extension $K(\alpha)$ is the unique monic irreducible polynomial canceling α.*

Proof. f is irreducible, because if f were reducible, say $f = gh$, then g or h would both have smaller degree, and one of them would cancel α, contradicting the minimality of f. Of course, they may not be monic, but could be made to be so by dividing out by the leading coefficient. Conversely, any irreducible polynomial canceling α will be in the ideal (f) as above, and since it's irreducible, will be a generator for that ideal, so will differ from f by a unit only. Being monic forces it to equal f.

□

Definition 8.3.5. The results above show that, for any element $\alpha \in L$ which is algebraic over K, there is a unique monic irreducible polynomial canceling α. The other roots of f will be called the **conjugates of** α.

Note that the terminology is consistent with the terminology for complex conjugation: we already call $-i$ the complex conjugate of i. Since i is a root of its minimal polynomial $x^2 + 1$, and the other root is $-i$, $-i$ is also the conjugate in the sense just defined.

[1] Moreover, f is nonconstant, for if it were constant, then being nonzero, it would be a unit, hence $\ker \phi$ would be the unit ideal, which is impossible for a ring homomorphism.

9 Minimal Polynomials of Finite Algebraic Extensions

Previous lecture - Next lecture

Last lecture we saw that any simple algebraic extension $K(\alpha)$ of K can be written as $K[x]/(f)$ with f being the unique monic polynomial of lowest degree canceling α. Today we will learn how to find this f, and thus how to find the degree of extensions; then we will see that every finite algebraic extension is simple, so our methods will apply to other extensions which at first seemed more complicated.

9.1 Computing the Minimal Polynomial

The minimal polynomial f of a simple extension $K(\alpha)$ is necessarily irreducible, so if we can find an irreducible polynomial canceling α, as long as we make it monic, it will be the minimal polynomial. This is useful when working over $\mathbb{Q}$, because we have a test for irreducibility (Eisenstein's Criterion). Here are some examples.

Example 9.1.1.

1. Consider the extension field $\mathbb{Q}(\sqrt{2})$ of $\mathbb{Q}$. The monic polynomial $f(x) = x^2 - 2$ is irreducible over $\mathbb{Q}$ (either by Eisenstein, or by noting that its two roots $\pm\sqrt{2}$ do not lie in $\mathbb{Q}$), so f is the minimal polynomial for this extension. We see from this that the degree of the extension is 2, but we already knew this since we have seen that $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}[\sqrt{2}]$, and every element of $\mathbb{Q}[\sqrt{2}]$ can be written uniquely as $a + b\sqrt{2}$ for $a, b \in \mathbb{Q}$, so the set $\{1, \sqrt{2}\}$ is a basis over $\mathbb{Q}$.

2. Now consider the simple extension $\mathbb{Q}(i\sqrt[3]{2})$ of $\mathbb{Q}$. What is its minimal polynomial? What is its degree? First of all, look at the powers of $i\sqrt[3]{2}$, namely

$$1, i\sqrt[3]{2}, (i\sqrt[3]{2})^2 = -\sqrt[3]{4}, (i\sqrt[3]{2})^3 = -2i, (i\sqrt[3]{2})^4 = 2\sqrt[3]{2}, (i\sqrt[3]{2})^5 = 2i\sqrt[3]{4}, (i\sqrt[3]{2})^6 = -4.$$

   So we find that $(i\sqrt[3]{2})^6$ is in $\mathbb{Q}$, and in particular that $i\sqrt[3]{2}$ is a root of $x^6 + 4$. But is this the minimal polynomial? How can we be sure that there's not a lower-degree polynomial cancelling $i\sqrt[3]{2}$? So far we know that the degree (let's call it n) of the extension is at most 6. But looking at the list above, we see that $-2i$ is in this field, hence so is i, so $\mathbb{Q}(i)$ is an intermediate subfield. Since $\mathbb{Q} \subset \mathbb{Q}(i)$ is a degree two extension, $2|n$, by proposition 8.1.7. Similarly, since $2\sqrt[3]{2}$ is in this field, $\mathbb{Q}(\sqrt[3]{2})$ is an intermediate subfield, and as the degree of $\mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2})$ is 3, $3|n$. Thus n must be 6, and so $x^6 + 4$ is indeed the minimal polynomial of the extension.

3. Now consider $\mathbb{Q} \subset \mathbb{Q}(\alpha)$, where $\alpha = \sqrt{2}+\sqrt{3}$. Let's find the minimal polynomial of this extension. First compute $\alpha^2 = 5+\sqrt{6}$. Now move everything that's rational to one side, giving $\alpha^2 - 5 = \sqrt{6}$. Square again: $(\alpha^2 - 5)^2 = 6$. Thus the polynomial $f(x) = (x^2 - 5)^2 - 6$ is monic and cancels α, so it's certainly a good candidate for the minimal polynomial. Now f has degree 4, but how can we be sure there isn't another one with lower degree? One way to show it's the minimal polynomial is to show that the degree of the extension must be at least 4, by just writing down some elements. The set

$$1, \alpha, \alpha^2 = 2\sqrt{6} + 5, (\alpha^2 - 5)\alpha = 4\sqrt{3} + 6\sqrt{2}$$

   is independent over $\mathbb{Q}$, so the degree is at least 4, hence exactly 4, and our f then must be the minimal polynomial.

Can we extend any of the above procedures to some sort of algorithm? Yes, but it's not always the most efficient - irreducibility inspection often tends to be the most effective route, so Corollary 8.3.4 is a good ally...

Proposition 9.1.2. *Let $K \subset K(\alpha)$ be a simple algebraic extension of degree n. The list*

$$1, \alpha, \alpha^2, \ldots, \alpha^{n-1}, \alpha^n,$$

being linearly dependent, satisfies some dependence relation

$$a_0 + a_1\alpha + \ldots + a_{n-1}\alpha^{n-1} + a_n\alpha^n = 0,$$

with $a_n \neq 0$. Then the polynomial $f(x) = \frac{a_0}{a_n} + \frac{a_1}{a_n}x + \cdots + \frac{a_{n-1}}{a_n}x^{n-1} + x^n$ is the minimal polynomial for this extension.

Proof. There's not much to prove. $a_n \neq 0$ because if it's zero, then there is a dependence amongst $1 \ldots, \alpha^{n-1}$, but the previous theorem said they're independent. Then after dividing through by a_n we obtain a monic polynomial of degree n canceling α, which by uniqueness of the minimal polynomial must *be* the minimal polynomial. □

The algorithm works well when the adjoined element is not a sum, as in example 2 above: we just kept taking powers of $i\sqrt[3]{2}$ until we dropped back into the rationals. However, in example 3, we saw it was easier to keep moving everything but the radical over to one side until we got something rational. In that example, then, the algorithm above wasn't necessarily the best choice.

9.2 Theorem of the Primitive Element

The previous section contained an extremely powerful result which gives us a complete understanding of the structure of a simple extension. It gets better: here we show that every finite extension is algebraic, and state a result which says any finite algebraic extension is actually simple. Thus the understanding of extensions furnished by Theorem 8.3.3 applies to a seemingly larger class of extensions.

Proposition 9.2.1. *Every finite field extension $K \subset L$ is algebraic.*

Proof. Suppose for contradiction that L were transcendental over K. Then there would be an element $\alpha \in L$ which satisfied no polynomial relation over K. This would imply that the set $1, \alpha, \alpha^2, \alpha^3, \ldots$ would be linearly independent (a finite dependence relation amongst them would result in a polynomial canceling α). This would contradict that L is finite over K.

□

So any finite extension is algebraic; the next result says moreover that it's *simple*. We will postpone the proof until the next lecture when we have developed some more tools. But we state it here for thematic completeness.

Theorem 9.2.2 (Theorem of the Primitive Element)**.** *Every finite algebraic extension of K is a simple extension. In other words, given a finite algebraic extension L of K, there is an element $\alpha \in L$ such that $L = K(\alpha)$.*

Definition 9.2.3. The element α of the theorem is called a **primitive element** for the finite algebraic extension $K \subset L$.

10 Field Automorphisms and The Galois Group

Previous lecture - Next lecture

From here until the end of the course, we will assume that all fields are subfields of $\mathbb{C}$, to simplify the results and their proofs. We will also assume all field extensions are finite.

10.1 Automorphisms and Galois Groups

The fundamental idea of Galois Theory is to introduce some group theory into the study of polynomials and field extensions. The way to do this is to look at automorphism groups. These automorphisms can be thought of as a group of "symmetries of the field". On the other hand, given a polynomial has a set of roots, and we will see that these automorphisms can also be thought of as a group of "symmetries of the roots". We will eventually see the relation between these two types of symmetry.

Definition 10.1.1. An **automorphism** of a field L is just an isomorphism of fields $L \to L$. The set $\operatorname{Aut} L$ of all automorphisms of L forms a group under composition, called the **automorphism group** of L. If $K \subset L$ is a field extension, we say an automorphism $\alpha \in \operatorname{Aut} L$ **fixes** K **pointwise** (or sometimes just sloppily **fixes** K) if $\alpha(x) = x$ for every $x \in K$. The subset of $\operatorname{Aut} L$ consisting of those automorphisms of L which fix K pointwise is a subgroup, called the **Galois group of the extension** $K \subset L$, and is denoted $\operatorname{Gal}(L : K)$.

Since we have been thinking of groups as symmetries of various objects, you can think of $\operatorname{Aut} L$ as "symmetries of the field L". Requiring them to be isomorphisms (as opposed to arbitrary permutations) is reasonable, since we are viewing L not just as a set but as a field. Similarly, elements of $\operatorname{Gal}(L : K)$ can be thought of as "symmetries of L which don't do anything to K."

We will be working over fields which are contained in $\mathbb{C}$, and the following basic observations will help us narrow down the possible automorphisms of such fields.

Proposition 10.1.2. *Every subfield L of $\mathbb{C}$ contains $\mathbb{Q}$, and every automorphism of L fixes $\mathbb{Q}$ pointwise.*

Proof. Begin with an arbitrary subfield L of $\mathbb{C}$. Then it contains 1, so it contains all the integers since it's closed under addition and subtraction. Since it's closed under taking multiplicative inverses (it's a field), we get all the rational numbers.

Now suppose σ is an arbitrary automorphism of L. Then since $\sigma(1) = 1$, $\sigma(n) = n$ for any $n \in \mathbb{Z}$, as we've already shown for ring homomorphisms. Now pick any $\frac{a}{b} \in \mathbb{Q}$. Then since $b\frac{a}{b} = a$, applying σ gives $\sigma(b)\sigma(\frac{a}{b}) = \sigma(a)$, so since $a, b \in \mathbb{Z}$, $b\sigma(\frac{a}{b}) = a$, so $\sigma(\frac{a}{b}) = \frac{a}{b}$. Thus σ fixes every element of $\mathbb{Q}$. □

Corollary 10.1.3. *The automorphism group of $\mathbb{Q}$ is trivial.*

Example 10.1.4. We will compute two automorphism groups and one Galois group; take $K = \mathbb{Q}(\sqrt{2})$ and $L = \mathbb{Q}(\sqrt{2}, i)$. Then the automorphism group of K consists of two elements: the identity map, and the "conjugation" map sending $\sqrt{2}$ to $-\sqrt{2}$. To see this, suppose $\sigma \in \operatorname{Aut} L$, and pick $a+b\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. Then $\sigma(a + b\sqrt{2}) = \sigma(a) + \sigma(b)\sigma(\sqrt{2}) = a + b\sigma(\sqrt{2})$, so σ is determined by what it does to $\sqrt{2}$. This is another instance of the general philosophy that "homomorphisms are determined by what they do to the generators", which we have seen repeatedly in this course. Let's see what the possible values for $\sigma(\sqrt{2})$ are. Since $(\sqrt{2})^2 = 2$, we must have $2 = \sigma(2) = \sigma(\sqrt{2}^2) = \sigma(\sqrt{2})^2$, so $\sigma(\sqrt{2}) = \pm\sqrt{2}$. This gives the two automorphisms mentioned above.

Similarly, looking at $\operatorname{Aut} L$, we find that any such map σ on L is determined by what it does to $\sqrt{2}$ and i. By the same argument, we find that $\sigma(\sqrt{2}) = \pm\sqrt{2}$, and since $i^2 = -1$, $\sigma(i)^2 = -1$, so that $\sigma(i) = \pm i$. Thus there are four possible automorphisms of L. Since each one has order 2 (meaning if you compose it with itself, you get the identity map), we find that $\operatorname{Aut} L \cong K_4$, the Klein four group. Since L is an extension field of K, we can compute the Galois group $\operatorname{Gal}(L : K)$. Automorphisms in $\operatorname{Gal}(L : K)$ are required to fix K, so in particular they must fix $\sqrt{2}$. Thus out of our four automorphisms above, only two fix $\sqrt{2}$, hence $\operatorname{Gal}(L : K)$ is a group with 2 elements, namely the identity map, and the map $\sigma(a + b\sqrt{2} + ci) = a + b\sqrt{2} - ci$.

The main calculations of the previous example generalize:

Proposition 10.1.5. *Let $K \subset L$ and $f \in K[x]$. If $\sigma \in \mathrm{Gal}(L:K)$ and α is a root of f (in K or L), then $\sigma(\alpha)$ is a root of f, and is in K if and only if α is in K.*

Proof. Write $f(x) = a_n x^n + \cdots + a_1 x + a_0$, with $a_i \in K$, so that $\sigma(a_i) = a_i$. Then

$$f(\sigma(\alpha)) = a_n\sigma(\alpha)^n + \cdots + a_1\sigma(\alpha) + a_0 = s\sigma(a_n\alpha^n + \cdots + a_1\alpha + a_0) = \sigma(0) = 0,$$

so $\sigma(\alpha)$ is a root of f. It is clear since σ fixes K pointwise that if $\alpha \in K$, so is $\sigma(\alpha)$ (which is just α). Conversely, if $\sigma(\alpha) \in K$, then applying σ^{-1}, which also fixes K, we get that $\alpha \in K$. □

10.2 Embeddings of Subfields

The previous result says that elements of $\mathrm{Gal}(L:K)$ "permute the roots" of polynomials over K. Thus $\mathrm{Gal}(L:K)$ can be thought of as "symmetries of the roots of a polynomial" (the minimal polynomial for the extension, say). We'd like to say that the elements of $\mathrm{Gal}(L:K)$ correspond exactly to permutations of the roots of a polynomial. The problem with this is that the roots do not necessarily live in the same subfields. Permutations of roots do give rise to maps out of L, but they're not always automorphisms of L because the image may not be L. These maps are what are called **embeddings** of L in $\mathbb{C}$.

For example, consider the extension $\mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2})$ of $\mathbb{Q}$ inside $\mathbb{C}$. $\sqrt[3]{2}$ is a root of x^3-, its minimal polynomial, and the other two roots are $\sqrt[3]{2}\omega$ and $\sqrt[3]{2}\omega^2$, where $\omega = \frac{-1+i\sqrt{3}}{2}$ is an imaginary cube root of 1. If we permute the roots according to $\sqrt[3]{2} \mapsto \sqrt[3]{2}\omega \mapsto \sqrt[3]{2}\omega^2 \mapsto \sqrt[3]{2}$, a "cyclic permutation", what do we get? Not an automorphism of $\mathbb{Q}(\sqrt[3]{2})$, but rather isomorphisms

$$\mathbb{Q}(\sqrt[3]{2}) \to \mathbb{Q}(\sqrt[3]{2}\omega) \to \mathbb{Q}(\sqrt[3]{2}\omega^2) \to \mathbb{Q}(\sqrt[3]{2}).$$

These maps are not automorphisms for the simple reason that their domains and codomains do not coincide. It may seem like a nitpicky technical point, but this small detail constitutes the primary technical obstacle in understanding the relation between the field structures of extensions and the group structures arising from symmetries of roots.

So we briefly investigate these embeddings, before figuring out what conditions on the extension ensure that they actually *are* automorphisms.

Definition 10.2.1. If K is a subfield of $\mathbb{C}$, an **embedding** of K is just an injective map $K \to \mathbb{C}$. Equivalently, it is an isomorphism of K to some other subfield of $\mathbb{C}$ (this other subfield just being the image of the injective map).

To prove the next two theorems, we will occasionally make use of the formal derivative of a polynomial:

Definition 10.2.2. If $f = \sum_{k=0}^{n} a_k x^k \in K[x]$, the **formal derivative of** f is $Df = \sum_{k=1}^{n} k a_k x^{k-1}$.

So this is just the usual formula for derivatives of polynomials, but defined "formally", without reference to limits. We need this primarily to detect multiple roots of f:

Lemma 10.2.3. *If $f \in K[x]$ has a repeated linear factor $(x-\alpha)$ (where $\alpha \in \mathbb{C}$ need not lie in K), then $(x-\alpha)$ is also a factor of Df.*

Proof. Write $f(x) = (x-\alpha)g(x)$, where $g(x)$ also has a factor $(x-\alpha)$, hence $g(\alpha = 0$. Then $Df(x) = D(x-\alpha)g(x) + (x-\alpha)Dg(x) = g(x) + (x-\alpha)Dg(x)$ (the usual product rule holds also for this formal derivative). Plugging in α gives $Df(\alpha) = 0$, so α is a root of Df.

□

Lemma 10.2.4. *Let f be an irreducible polynomial over K. Then the linear factors of f (over $\mathbb{C}$) are distinct.*

Proof. Suppose for a contradiction that $(x-\alpha)$ is a repeated linear factor of f. Then $(x-\alpha)$ is a factor of Df as well. But the ideal (f, Df) generated by f and Df is the unit ideal in $K[x]$ (check this as an exercise), so we can write $1 = gf + hDf$ for some $g, h \in K[x]$. Plugging in α gives $1 = g(\alpha)f(\alpha) + h(\alpha)Df(\alpha) = 0$, a contradiction.

□

With these simple tools, we are ready to prove two powerful theorems:

Theorem 10.2.5. *Let $K \subset L$ be a finite extension. Every embedding of K can be extended to exactly $[L : K]$ embeddings of L.*

Proof. We go by induction on $[L : K]$. If it's 1, then $L = K$ and there is nothing to show. Suppose $[L : K] = n > 1$. Pick an embedding σ of K. Let K' be the image of K under σ, so that σ is an isomorphism from K to K'. We need to show that there are n ways to extend this to an embedding of L, meaning that there are exactly n embeddings of L which do the same thing to K as σ does. Pick $\alpha \in L \setminus K$ (possible since $[L : K] > 1$). We will consider the tower of extensions $K \subset K(\alpha) \subset L$, and work out the extensions explicitly to $K(\alpha)$, and finally apply the inductive assumption to the other extension $K(\alpha) \subset L$.

First we try to extend σ to $K(\alpha)$. Let $f \in K[x]$ be the minimal polynomial of α say it has degree m so $[K(\alpha) : K] = m$. If we apply σ to the coefficients of f, we get a polynomial $g \in K'[x]$, which is also irreducible because σ is an isomorphism (so in particular, it sends irreducibles to irreducibles). Now let $\beta_1, \ldots, \beta_m$ be the roots of this g. We can use these to define m maps $\tau_i \colon K(\alpha) \to K'(\beta_i)$, by sending α_i to β_i and sending any $a \in K$ to $\sigma(a) \in K'$. So these maps each extend σ to $K(\alpha)$. Moreover, these maps are isomorphisms, because they're injective (maps of fields always are), and $K(\alpha) \cong K[x]/(f) \cong K'[x]/(g) \cong K'(\beta_i)$[1]. So far we have shown that there are m embeddings of $K(\alpha)$ which extend σ. Note that they're all distinct maps because the β_i are distinct by 10.3.3. There are no others because an embedding extending σ is determined by what it does to α, and (by the same argument in 10.1.5), it must send α to one of the β_i.

Now we consider the problem of extending an embedding of $K(\alpha)$ to an embedding of L. But the degree of this extension is less than n, since $\alpha \notin L$. Therefore by inductive hypothesis there are $[L : K(\alpha)]$ possible such extensions, for each embedding of $K(\alpha)$. So in total there are $[L : K(\alpha)]m$ ways to extend σ. But by 8.1.7 we have

$$[L : K(\alpha)]m = [L : K(\alpha)][K(\alpha) : K] = [L : K],$$

which proves that there are $[L : K]$ extensions of σ.

□

Taking σ to be the identity embedding in the above, we get:

Corollary 10.2.6. *There are exactly $[L : K]$ embeddings of L which fix K pointwise.*

10.3 Proof of Primitive Element Theorem

Finally we can prove the Primitive Element Theorem, stated at the end of lecture 9:

Theorem 10.3.1 (Theorem of the Primitive Element)**.** *Every finite algebraic extension of K is a simple extension. In other words, given a finite algebraic extension L of K, there is an element $\alpha \in L$ such that $L = K(\alpha)$.*

Definition 10.3.2. The element α is a **primitive element** for the extension.

Proof. By induction it suffices to consider the case where $L = K(\alpha, \beta)$. Let $n = [L : K]$. For any $a \in K$, we choose the element $\alpha + a\beta \in L$ as a candidate for the primitive element. We will show that there are only finitely many choices for a such that this element *does not* generate L; we will call these "bad" choices of a. Thus "almost every" choice of a gives a primitive element.

Suppose that a is one of those choices; namely $K(\alpha + a\beta) \neq L$. Then $\alpha + a\beta$ has a minimal polynomial f of degree less than n, hence has fewer than n roots. But we also have, by Corollary 10.3.5,

[1] Any injective map between fields which are already known to be isomorphic must itself be an isomorphism.

n embeddings σ_i of L which fix K pointwise. Each of these σ_i takes $\alpha + a\beta$ to one of the other roots of f. Since there are more σ_i's than roots, at least two of the σ_i's must take $\alpha + a\beta$ to the same root. Say these two are σ_1 and σ_2. Then we have

$$\sigma_1(\alpha) + a\sigma_1(\beta) = \sigma_2(\alpha) + a\sigma_2(\beta)$$

Solving for a gives us

$$a = \frac{\sigma_1(\alpha) - \sigma_2(\alpha)}{\sigma_2(\beta) - \sigma_1(\beta)}$$

This is the form[2] any such "bad" choice of a must take. But since there are only finitely many possible σ, there are therefore only finitely many "bad" choices of a. Thus since K is an infinite field (a subfield of $\mathbb{C}$ must always contain $\mathbb{Q}$) there are infinitely many values of a for which $K(\alpha + a\beta) = L$. So L is simple.

□

The theorem gives a sort of algorithm for producing a primitive element. If L is generated over K by α and β, then you can start trying elements of the form $\alpha \pm \beta$, $\alpha \pm 2\beta$, etc as primitive elements and you are guaranteed success in finite time. This is essentially what we did in Example 9.2.4, in which case $a = 1$ worked.

Example 10.3.3.

1. Consider the extension $\mathbb{Q} \subset \mathbb{Q}(\sqrt{2}, \sqrt{3})$, which does not at first appear simple. We will show that actually $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2}+\sqrt{3})$. It's clear that $\mathbb{Q}(\sqrt{2}+\sqrt{3}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3})$. For the other inclusion, letting $\alpha = \sqrt{2}+\sqrt{3}$, we have to show that both $\sqrt{2}$ and $\sqrt{3}$ can be written as polynomials in α (with rational coefficients). We observe that since $\alpha^2 = 2\sqrt{6}+5$, $\sqrt{6} \in \mathbb{Q}(\alpha)$. So is $\sqrt{6}\alpha = \sqrt{12} + \sqrt{18} = 2\sqrt{3} + 3\sqrt{2}$. So $\sqrt{6}\alpha - 2\alpha = \sqrt{3} \in \mathbb{Q}(\alpha)$ and $-(\sqrt{6}\alpha - 3\alpha) = \sqrt{2}$. Once we know that $\sqrt{2}$ and $\sqrt{3}$ are in $\mathbb{Q}(\alpha)$, we get that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) \subseteq \mathbb{Q}(\alpha)$.

2. Consider the extension $\mathbb{Q}(i\sqrt[3]{2})$ of $\mathbb{Q}$. We have seen in prior computations that $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt[3]{2})$ are subfields of $\mathbb{Q}(i\sqrt[3]{2})$. This is because, letting $\alpha = i\sqrt[3]{2}$, we have $-\frac{1}{2}\alpha^3 = i$ and $\frac{1}{2}\alpha^4 = \sqrt[3]{2}$. Thus $\mathbb{Q}(i, \sqrt[3]{2}) \subseteq \mathbb{Q}(i\sqrt[3]{2})$. But the reverse inclusion is clear (if i and $\sqrt[3]{2}$ are in a field, then so is their product), so in fact $i\sqrt[3]{2}$ is a primitive element for the extension, even though it does not have the form prescribed in the proof of the theorem. This shows that elements $\alpha + a\beta$ are not *the only* ways to produce a primitive element. Indeed, you may prefer to try guesses of the form $\alpha\beta$, as in this example[3].

[2] Note that this makes sense: the denominator is nonzero, because if $\sigma_1(\beta) = \sigma_2(\beta)$, then we would also have $\sigma_1(\alpha) = \sigma_2(\alpha)$, which implies that $\sigma_1 = \sigma_2$ since $\alpha + a\beta$ generates the extension. But σ_1 and σ_2 are distinct.

[3] These types of guesses don't always work: for example, $i\sqrt{2}$ is not a primitive element for $\mathbb{Q}(i, \sqrt{2})$.

11 The Galois Correspondence

Previous lecture - Next lecture

Recall our standing assumptions: all field extensions are finite, and take place inside of $\mathbb{C}$.

Let us summarize the state of affairs: all extensions L of K (being finite) are algebraic and simple and thus correspond to a unique monic irreducible polynomial. Conversely, every monic irreducible polynomial corresponds to a unique extension L of K, namely $L = K[x]/(f)$. Given such an extension $K \subset L$, with corresponding polynomial f, every automorphism of L fixing K pointwise (i.e., every element of $\mathrm{Gal}(L : K)$) permutes the roots of f. But unfortunately, **not every permutation of the roots of f gives rise to an automorphism of L fixing K pointwise[1].** The fundamental question we wish to address is: how can we decide which permutations of the roots correspond to elements of $\mathrm{Gal}(L : K)$?

To answer this question, we must go back to a study of the Galois group, together with its subgroup structure, and see how these subgroups relate to **intermediate subfields of the extension $K \subset L$**, by which we mean subfields F with $K \subseteq F \subseteq L$.

Example 11.0.1 (Vague)**.** We motivate this with a brief example: take $K = \mathbb{Q}$ and $L = \mathbb{Q}(i, \sqrt[4]{2})$; the extension $K \subset L$ has degree 8. Consider an element $\sigma \in \mathrm{Gal}(L : K)$. Since $i^2 = -1$ any such σ, since it fixes $\mathbb{Q}$, must send i to $\pm i$; similarly $\sqrt[4]{2}^4 = 2$ forces $\sigma(\sqrt[4]{2})^4 = 2$, so there are four choices for what σ does to $\sqrt[4]{2}$. Thus you might expect a generator of order 2 and one of order 4 in $\mathrm{Gal}(\mathbb{Q}(\sqrt[4]{2} : \mathbb{Q})$. This intuition is validated when one computes that the Galois group $\mathrm{Gal}(\mathbb{Q}(i, \sqrt[4]{2}) : \mathbb{Q})$ is in fact D_8, the dihedral group! By investigating the behavior of σ on i and $\sqrt[4]{2}$, we are essentially looking at the behavior of automorphisms on the intermediate subfields $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt[4]{2})$.

11.1 The Galois Connection

Here we introduce a relationship between subgroups of $\mathrm{Gal}(L : K)$ and intermediate subfields of $K \subset L$.

Definition 11.1.1. Let $K \subset L$ be an extension. For each intermediate subfield F of $K \subset L$, we define

$$\Gamma(F) = \{\sigma \in \mathrm{Gal}(L : K) \,|\, \sigma(f) = f \text{ for all } f \in F\}.$$

For each subgroup H of $\mathrm{Gal}(L : K)$, we define

$$\Phi(H) = \{x \in L \,|\, \sigma(x) = x \text{ for all } \sigma \in H\}$$

We usually call $\Phi(H)$ the **fixed field** of the subgroup H, and call $\Gamma(F)$ the **fixing group**[2] of F. The maps Φ and Γ together are called the **Galois connection**.

You can remember which is which as follows: Φ is greek for "F", so stands for "field" (or "fixed"); whereas Γ is greek for "G", and stands for "group". So Φ turns (sub)groups into (intermediate) fields, and Γ turns (intermediate) fields into (sub)groups.

Of course, we have to check that these things are really subfields and subgroups:

Proposition 11.1.2. *In the situation of the definition above, $\Gamma(F)$ is a subgroup of $\mathrm{Gal}(L : K)$ and $\Phi(H)$ is an intermediate subfield of $K \subset L$.*

Proof. Exercise. Note that $\Gamma(F) = \mathrm{Gal}(L : F)$. □

Example 11.1.3. Consider the extension $\mathbb{Q}(i, \sqrt{3})$ of $\mathbb{Q}$. The Galois group has order four[3], and is generated by σ, which sends i to $-i$ and does nothing to $\sqrt{3}$ and τ, which sends $\sqrt{3} \mapsto -\sqrt{3}$ and does nothing to i. The other two elements are the identity map id, and the composite $\sigma \circ \tau$. Now let H be the subgroup of order 2 generated by the element $\sigma \circ \tau$. Its fixed field is $\Phi(H) = \mathbb{Q}(i\sqrt{3})$, because $\sigma \circ \tau$ applies a negative sign to both i and $\sqrt{3}$, so it does nothing to $i\sqrt{3}$. Consider the intermediate field $F = \mathbb{Q}(\sqrt{3})$, it is a degree two subextension. It's fixing group $\Gamma(F)$ is just the group generated by σ, since both τ and $\sigma \circ \tau$ negate $\sqrt{3}$, so they cannot be in $\Gamma(F)$.

[1] Recall the fundamental example of this: the roots of $x^3 - 2$. Sending the real root to one of the imaginary roots induces only an embedding of $\mathbb{Q}(\sqrt[3]{2})$, not an automorphism of it.

[2] The term "fixed field" is widely used; whereas the term "fixing group" is for some reason very rarely used.

[3] This Galois group is isomorphic to the Klein four group.

Proposition 11.1.4. *1. The Galois connection is order reversing, in the following sense: If $H_1 \subseteq H_2$ as subgroups of $\mathrm{Gal}(L:K)$, then $\Phi(H_!) \supseteq \Phi(H_2)$ as subfields of L.*

2. $\Phi(\Gamma(F)) \supseteq F$ and $\Gamma(\Phi(H)) \supseteq H$

Part 2 of the proposition suggests that the maps Φ and Γ might be inverses to one another, but sadly they are not always so. We are interested, then, in the following question: under what conditions on the extension $K \subset L$ are the maps Γ and Φ mutually inverse? I.e., for which extensions do we have a bijection

$$\{\text{intermediate subfields } K \subseteq F \subseteq L\} \underset{\Phi}{\overset{\Gamma}{\rightleftharpoons}} \{ \text{ subgroups } H \subseteq \mathrm{Gal}(L:K)\}$$

11.2 Normality

The answer to this question is suggested by our example showing that permutations of roots need not lead to elements of $\mathrm{Gal}(L:K)$ - the extension $\mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2})$. Here Φ and Γ are not mutually inverse, because the Galois group is trivial (check!) but there are 2 intermediate subfields, so we cannot have a bijection. In this case, the problem is essentially due to the fact that the other two roots of the minimal polynomial $x^3 - 2$, namely $\sqrt[3]{2}\omega$ and $\sqrt[3]{2}\omega$, did not lie in the field $\mathbb{Q}(\sqrt[3]{2})$. So this extension was bad because it contained some, but not all, of the roots of the minimal polynomial. Normal extensions are ones which do not have this problem:

Definition 11.2.1. An extension $K \subset L$ is called **normal** if all the roots of the minimal polynomial for L over K lie in L.

Theorem 11.2.2 (Equivalent Characterizations of Normality)**.** *The following statements about a field extension $K \subset L$ are equivalent.*

1. *$K \subset L$ is a normal extension.*
2. *Every embedding of L which fixes K pointwise is an automorphism of L.*
3. *$|\mathrm{Gal}(L:K)| = [L:K]$*

Proof. Let $n = [L:K]$, and let f be the minimal polynomial for the extension $K \subset L$. The equivalence of 2 and 3 is the easier one: the set $\mathrm{Gal}(L:K)$ of automorphisms of L fixing K is a subset of the set of embeddings of L fixing K, so the sets are equal if and only if they have the same size; the size of the latter set is $[L:K]$, by Corollary 10.2.6.

Now assume 1 holds, and let σ be an embedding of L which fixes K. Since σ is determined by what it does to the roots of f, and these all lie in L the image of σ must be contained in L, hence equal to L because the image has degree n over K also (σ is an isomorphism onto its image).

Conversely, assuming 2, first write $L = K(\alpha)$; this α will be a root of f, and we aim to show that all the other roots of f are in L also. Given any such root β, there is an embedding of L which sends α to β. But by our assumption, this embedding is actually an automorphism, so that β lies in L.

□

As a first clue that normal extensions are the ones that make the Galois connection Φ and Γ inverse to one another, we have

Proposition 11.2.3. *If $K \subset L$ is normal, then $\Phi(\mathrm{Gal}(L:K)) = K$, and for any proper subgroup H of $\mathrm{Gal}(L:K)$, $\Phi(H) \neq K$.*

It says that for normal extensions, "the largest subgroup corresponds uniquely to the smallest intermediate field."

Proof. Let $n = [L:K] = |\mathrm{Gal}(L:K)|$ (since $K \subset L$ is normal). Set $F = \Phi(\mathrm{Gal}(L:K))$, and suppose for contradiction that F is strictly larger than K. Then since $[L:K] = [L:F][F:K]$, and by assumption $[F:K] > 1$, we have that $[L:F] < n$, so by 11.2.2, there are less than n automorphisms of L fixing F pointwise. But on the other hand, F is the fixed field of $\mathrm{Gal}(L:K)$, so all of the n automorphisms in $\mathrm{Gal}(L:K)$ fix F pointwise. This is a contradiction.

For the second statement, begin with a subgroup H of $\mathrm{Gal}(L:K)$ whose fixed field is K (i.e., $\Phi(H) = K$). We will show that H must be all of $\mathrm{Gal}(L:K)$. Write $m = |H|$ (so we want to show $m = n$). Since L is finite over K we can write it as $K(\alpha)$ for some $\alpha \in L$. Then the minimal polynomial f for α has degree n (the degree of the extension). Now look at the polynomial

$$g(x) = \prod_{\sigma \in H} (x - \sigma(\alpha)).$$

What are the properties of this g? First of all, its coefficients certainly live in L, so it is a polynomial in $L[x]$. Moreover it has degree m, and has α as a root, since one of the $\sigma \in H$ must be the identity map so there is a term $(x-\alpha)$ in g, which will cancel α. But in fact by a worksheet problem, the coefficients of g actually live in $\Phi(H)$, which by our assumption is equal to K. So $g \in K[x]$ and cancels α which means the minimal polynomial f for α divides g, so $n = \deg f \leq \deg g = m$. But we also know $m \leq n$ since H is a subgroup of size m in the group $\mathrm{Gal}(L:K)$ of size n. Thus $m = n$, so $H = \mathrm{Gal}(L:K)$.

□

We close by summarizing the situation with some familiar examples, to illustrate the difference between normality and non-normality.

Example 11.2.4.

1. Let $K \subset L$ be the extension $\mathbb{Q} \subset \mathbb{Q}(\sqrt{2})$. It has degree 2 and its minimal polynomial is $x^2 - 2$. Since the roots of this polynomial are $\pm\sqrt{2}$, they both live in L, so we have a normal extension. The Galois group $\mathrm{Gal}(L:K)$ is isomorphic to $\mathbb{Z}/2\mathbb{Z}$, consisting of the identity automorphism and the automorphism τ which switches $\sqrt{2}$ and $-\sqrt{2}$. One could also check normality as follows: since there's only one automorphism of $\mathbb{Q}$, we just need to see that there are exactly two ways to extend this to an embedding of $\mathbb{Q}(\sqrt{2})$. Since any such embedding must send $\sqrt{2}$ to another root of $x^2 - 2$, we get precisely the automorphisms described above. Meanwhile, the fixed field of the trivial subgroup of $\mathrm{Gal}(L:K)$ is all of $\mathbb{Q}(\sqrt{2})$(the identity map fixes everything); whereas the fixed field of the entire group $\mathrm{Gal}(L:K)$ is just $\mathbb{Q}$, either by the preceding proposition or by noticing that once $\sqrt{2}$ is fixed, so is $-\sqrt{2}$. So the two subgroups correspond bijectively with the two subfields, with big subgroup corresponding to small subfield and vice versa.

2. Now let $\mathbb{Q} = K \subset \mathbb{Q}(\sqrt[3]{2}) = L$. This simple example, which we keep returning to, exhibits all the behavior that we want to avoid. The minimal polynomial of the extension is $f = x^3 - 2$. But the other two roots, $\omega\sqrt[3]{2}$ and $\omega^2\sqrt[3]{2}$, do not lie in $\mathbb{Q}(\sqrt[3]{2})$. Therefore by definition this extension is not normal. If we compute the Galois group, we need to decide where an automorphism sends $\sqrt[3]{2}$; it has to send it to another one of the cube roots of 2, but since $\sqrt[3]{2}$ is the only one in our field, our automorphism is forced to be the identity map. So $\mathrm{Gal}(K:\mathbb{Q}) = \{\mathrm{id}\}$. Then the fixed field of this group is all of $\mathbb{Q}(\sqrt[3]{2})$. But there is one another intermediate subfield, namely just $\mathbb{Q}$, and this intermediate subfield does not correspond to any subgroup of $\mathrm{Gal}(K:\mathbb{Q})$, since there's only one such subgroup. Thus the fact that there weren't enough roots of $x^3 - 2$ in our field restricted the size of the Galois group, and there was no bijection between subfields and subgroups.

3. **CAUTION:** A normal extension of a normal extension may not be normal! For example, consider the tower of extensions $\mathbb{Q} \subset \mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\sqrt[4]{2})$. The first extension is normal since both roots of the minimal polynomial $x^2 - 2$ lie in $\mathbb{Q}(\sqrt{2})$. The second extension is normal since the minimal polynomial of $\mathbb{Q}(\sqrt[4]{2})$ over $\mathbb{Q}(\sqrt{2})$ is $x^2 - \sqrt{2}$, both of whose roots $\pm\sqrt[4]{2}$ lie in $\mathbb{Q}(\sqrt[4]{2})$. But the overall extension $\mathbb{Q} \subset \mathbb{Q}(\sqrt[4]{2})$ is not normal since its minimal polynomial is $x^4 - 2$, with roots $\pm\sqrt[4]{2}, \pm i\sqrt[4]{2}$, only two of which lie in $\mathbb{Q}(\sqrt[4]{2})$.

12 Fundamental Theorem of Galois Theory

Previous lecture - Next lecture

Recall our standing assumptions: all field extensions are finite, and take place inside of $\mathbb{C}$.

12.1 The Theorem

Today's lecture is devoted to proving the following theorem:

Theorem 12.1.1 (Fundamental Theorem of Galois Theory)**.** *If $K \subset L$ is a normal[1] extension, then the maps Γ and Φ are mutually inverse, and thus induce bijections*

$$\{\textit{intermediate subfields } K \subseteq F \subseteq L\} \rightleftharpoons \{\textit{ subgroups } H \subseteq \mathrm{Gal}(L:K)\}.$$

Moreover, $H \subset \mathrm{Gal}(L:K)$ is a normal subgroup if and only if $F = \Phi(H)$ is a normal extension of K, in which case $G/H \cong \mathrm{Gal}(F:K)$.

We begin with a simple observation: if $G = \mathrm{Gal}(L:K)$, then G acts on both sets appearing in the Galois Correspondence. G acts on intermediate subfields by $\sigma \cdot F = \sigma F \stackrel{\text{Def}}{=} \{\sigma f \mid f \in F\}$, and $\sigma \cdot H = \sigma H \sigma^{-1}$ (the usual conjugation action on subgroups). The next theorem says that, with respect to these actions, the Galois connection is G-**equivariant**. Roughly speaking, this means that the two functions Φ and Γ of the correspondence "commute with the relevant groups actions".

Lemma 12.1.2. *Let $K \subset L$ be a normal extension, with Galois group G, and let σ be any element of G. If H is a subgroup of G, then $\Phi(\sigma \cdot H) = \sigma \cdot \Phi(H)$; similarly, if F is an intermediate field, then $\Gamma(\sigma \cdot F) = \sigma \cdot \Gamma(F)$*

Proof. To prove the first equality, pick a subgroup H of G. Then $\Phi(\sigma \cdot H)$ is the fixed field of the subgroup $\sigma H \sigma^{-1}$, which can be written $\{x \in L \mid (\sigma h \sigma^{-1})(x) = x \text{ for all } h \in H\}$. On the other hand, $\sigma \cdot \Phi(H)$ is the field obtained by taking the fixed field of H, and translating it by σ, i.e., $\{\sigma x \mid hx = x \text{ for all } h \in H\}$. To show these sets are equal, first pick σx such that $hx = x$ for all $h \in H$. Then $(\sigma h \sigma^{-1})(\sigma x) = \sigma h x = \sigma x$, so σx is in $\Phi(\sigma \cdot H)$. Now pick x such that $(\sigma h \sigma^{-1})(x) = x$ for all $h \in H$. Then we can write x as $\sigma(\sigma^{-1}x)$, where $\sigma^{-1}x$ satisfies $h(\sigma^{-1}x) = \sigma^{-1}x$ by our assumption on x. This shows that $x \in \sigma \cdot \Phi(H)$, so they're equal.

The proof of the statement for Γ is similar, and is omitted.

□

Proof of the Fundamental Theorem. Set $G = \mathrm{Gal}(L:K)$. Let F be an intermediate field (again, the cases $F = K$ and $F = L$ are allowed here), and for convenience denote by F' the field $\Phi(\Gamma(F)) = \Phi(\mathrm{Gal}(L:F))$, so we want to show that $F' = F$. For this, if we knew that L was normal over F, we could apply proposition 11.2.3 with K replaced by F, and we would be done. So why is $F \subset L$ normal? We can use part 2 of Theorem 11.2.2 (again with K replaced by F): If σ is an embedding of L which fixes F pointwise, then in particular σ fixes K pointwise (because $K \subset F$) and since we know that $K \subset L$ is normal, such a σ must be an automorphism of L. So Theorem 11.2.2 tells us that $F \subset L$ is normal. So we have shown that $\Phi(\Gamma(F)) = F$.

For the other direction, that $\Gamma(\Phi(H)) = H$, pick any subgroup H of G. Again for convenience, let $H' = \Gamma(\Phi(H))$, so we have to show that $H' = H$. We have seen (11.1.4) that $H \subseteq H'$. Now let $F = \Phi(H)$ and apply Proposition 11.2.3 to the extension $F \subset L$ (which is normal by the argument given in the preceding paragraph). It says that $\Phi(H') = \Phi(\mathrm{Gal}(L:F)) = F$, and for any proper subgroup of H', its fixed field is not F. Since H is a subgroup of H' whose fixed field is F (because we defined F to be $\Phi(H)$), this means that H cannot be proper, so $H' = H$. This completes the proof that when $K \subset L$ is normal, the Galois connection is a bijection.

[1] The hypotheses are comparatively mild, because we are working with finite extensions inside $\mathbb{C}$. In general, one requires the extension also to be **separable** in order for the Galois connection to be bijective. This is a technical condition which is automatically satisfied if your fields contain $\mathbb{Q}$, which ours, being subfields of $\mathbb{C}$, do.

The condition for normality of $K \subset F$ is that $\sigma F = F$ for all embeddings σ of F which fix K pointwise[2]. But each of these embeddings of F extends to an embedding of L fixing K, which is an automorphism since $K \subset L$ is normal. Conversely, any automorphism of L (fixing K) restricts to an embedding of F (fixing K). So we can restate the condition as follows: $K \subset F$ is normal if and only if $\sigma F = F$ for all $\sigma \in \mathrm{Gal}(L : K) = G$. In other words, **the normal subextensions are just the fixed points of the action of G on intermediate fields.** On the other hand, the normal subgroups of G are just fixed points of the action of G on its subgroups. Since the Galois Correspondence is G-equivariant, fixed points correspond to fixed points (check details? Nope!). Thus normal subgroups correspond to normal subextensions.

Finally, suppose we have a corresponding pair H and F, which are both normal (in their respective senses). We will define an isomorphism $G/H \to \mathrm{Gal}(F : K)$ using the isomorphism theorem. We first define a map $r\colon G \to \mathrm{Gal}(F : K)$ by sending an automorphism σ of L (fixing K) to its restriction to F. Since $K \subset F$ is normal, this is actually an automorphism of F (as opposed to just an embedding), so this is well-defined. The kernel of this map is just those elements σ of G which fix F pointwise, namely H. This gives an isomorphism $G/H \cong \mathrm{im}\ r$. But then $|\mathrm{im}\ r| = |G|/|H| = |\mathrm{Gal}(L : K)|/|\mathrm{Gal}(L : F)| = [L : K]/[L : F] = [F : K]$, which is precisely the size of $\mathrm{Gal}(F : K)$, so $\mathrm{im}\ r = \mathrm{Gal}(F : K)$. Thus r gives rise to an isomorphism $G/H \to \mathrm{Gal}(F : K)$, finishing the proof.

□

This corollary justifies our use of the notation $[K : F]$ for the degree of a field extension:

Corollary 12.1.3. *In the situation of the fundamental theorem, if F and G are normal and correspond under the Galois correspondence (meaning $H = \Gamma(F)$ and $F = \Phi(H)$), then the degree of F over K is equal to the index of H in G, i.e. $[F : K] = [G : H]$, or equivalently $[\Phi(H) : K] = [G : H]$, or equivalently $[F : K] = [G : \Gamma(F)]$.*

12.2 A Burly Example

Example 12.2.1. The Galois group of $\mathbb{Q}(i, \sqrt[4]{2})$ over $\mathbb{Q}$ is D_8. An automorphism σ of $L = \mathbb{Q}(i, \sqrt[4]{2})$ is determined by its action on i and $\sqrt[4]{2}$. Since i is a root of $x^2 + 1$, $\sigma(i)$ must also be, hence $\sigma(i) = \pm 1$. Similarly, since $\sqrt[4]{2}$ is a root of $x^4 - 1$, $\sigma(\sqrt[4]{2})$ must be also, and the four roots of $x^4 - 1$ are $\pm\sqrt[4]{2}, \pm i\sqrt[4]{2}$ (note that they all lie in L, as do both roots of $x^2 + 1$, so L is normal). Thus we obtain eight possibilities for σ, as follows:

$$\sigma_1\colon \begin{cases} \sqrt[4]{2} \mapsto \sqrt[4]{2} \\ i \mapsto i \end{cases} \quad \sigma_2\colon \begin{cases} \sqrt[4]{2} \mapsto i\sqrt[4]{2} \\ i \mapsto i \end{cases} \quad \sigma_3\colon \begin{cases} \sqrt[4]{2} \mapsto -\sqrt[4]{2} \\ i \mapsto i \end{cases} \quad \sigma_4\colon \begin{cases} \sqrt[4]{2} \mapsto -i\sqrt[4]{2} \\ i \mapsto i \end{cases}$$

$$\sigma_5\colon \begin{cases} \sqrt[4]{2} \mapsto \sqrt[4]{2} \\ i \mapsto -i \end{cases} \quad \sigma_6\colon \begin{cases} \sqrt[4]{2} \mapsto i\sqrt[4]{2} \\ i \mapsto -i \end{cases} \quad \sigma_7\colon \begin{cases} \sqrt[4]{2} \mapsto -\sqrt[4]{2} \\ i \mapsto -i \end{cases} \quad \sigma_8\colon \begin{cases} \sqrt[4]{2} \mapsto -i\sqrt[4]{2} \\ i \mapsto -i \end{cases}$$

So $\mathrm{Gal}(L : \mathbb{Q})$ is a group of order 8, and there are not many such groups. Check by composing each element with itself a few times that the orders of the automorphisms are, respectively, 1,4,2,4,2,4,2,4. This already rules out the quaternion group[3] (in which all non-identity elements have order 2). So in fact, as long as we check that it's non-abelian, we will be certain that this group is isomorphic to D_8. But let's take another approach, by finding generators and relations. Note that σ_1 is just the identity map. Let $\sigma = \sigma_2$ and $\tau = \sigma_5$. Draw a picture of square inscribed in circle, with vertices at the four roots of $x^4 - 2$, and indicate the actions of σ and τ. You will see that they exactly correspond to the generators r and s of the dihedral group.

Now draw a picture of the lattice of subextensions of $\mathbb{Q} \subset L$, and also the lattice of subgroups of D_8. Observe that they match up, and compute a few fixed fields for good measure. NOTE: I'm not typing up these diagrams, so you'd better come to class for this one...

[2] Because the issue of whether or not an embedding is actually an automorphism just comes down to what its image is, so forcing the image σF to be equal to F forces the embedding to be an automorphism.

[3] In fact, it rules out all the order 8 groups except for $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$.

13 Application: Solution by Radicals; Insolubility of the Quintic

Previous lecture

Recall our standing assumptions: all field extensions are finite, and take place inside of $\mathbb{C}$.

This material will not be on the exam. But read it anyway, for a laugh.

In this lecture we sketch an argument for the following famous and surprising result: there is no general formula which allows us to solve a quintic equation. This result in fact was the motivation for the development of Galois theory by Galois himself. At that time, the language of groups and rings was not available, and many of the ideas which went into the proof form some of the origins of modern abstract algebra. Thus it seems fitting to end the course here. Since there are quite a few technical details, I will omit some proofs and just give the main ideas.

13.1 Solvability by Radicals

Example 13.1.1. Say we wish to solve the equation $x^2 - x - 1 = 0$. We'd like to do so over the rationals, but that's not possible since the discriminant $b^2 - 4ac = 5$ is not a perfect square. The quadratic formula tells us that the roots are $\frac{1 \pm \sqrt{5}}{2}$ (the positive solution being the famous **golden ratio**), so to obtain the solution we must work not in $\mathbb{Q}$ but in the larger field $\mathbb{Q}(\sqrt{5})$.

Now suppose we wish to solve the cubic equation $x^3 - 6x + 2 = 0$. It turns out that the roots of this polynomial are

$$x = -\frac{1 + i\sqrt{3}}{\sqrt[3]{-1 + i\sqrt{7}}} - \frac{1}{2}(1 + i\sqrt{3})\sqrt[3]{-1 + i\sqrt{7}}$$
$$x = -\frac{1 - i\sqrt{3}}{\sqrt[3]{-1 + i\sqrt{7}}} - \frac{1}{2}(1 - i\sqrt{3})\sqrt[3]{-1 + i\sqrt{7}}$$
$$x = \frac{2}{\sqrt[3]{-1 + i\sqrt{7}}} + \sqrt[3]{-1 + i\sqrt{7}}$$

So in order to find the roots, we must first adjoin $i\sqrt{7}$, then adjoin $\sqrt[3]{-1 + i\sqrt{7}}$, and then adjoin $i\sqrt{3}$. This solution comes from a general formula, in which certain square and cube roots coming from the coefficients appear. So for this formula to make sense, most of the time we will have to work in a field obtained from $\mathbb{Q}$ by successively adjoining radicals.

More generally, suppose we have a polynomial f and a root α of f. Most of the time, α will not live in $\mathbb{Q}$, but in some extension field K of $\mathbb{Q}$. Continuing in the theme suggested by the above examples, it would be nice if we could obtain α by successively adjoining to $\mathbb{Q}$ various radicals. Surprisingly, this is not always possible! This motivates the following definition:

Definition 13.1.2. Let $f \in \mathbb{Q}[x]$, and let K_f be the subfield of $\mathbb{C}$ obtained by adjoining all the roots of f to $\mathbb{Q}$. We say that f is **solvable by radicals** if there is a chain of subfields $\mathbb{Q} \subset K_1 \subset \cdots \subset K_{m-1} \subset K_m$ such that

1. $K \subseteq K_m$, and
2. Each $K_{i+1} = K_i(\alpha_i)$, where α_i is a root of an *irreducible* polynomial of the form $x^{n_i} - b_i \in K_i[x]$.

The tower of extensions K_i (or sometimes just the top field K_m) is called a **radical extension**. The field K_f is called the **splitting field of** f. Finally, the **Galois group of the polynomial** f is the group $\mathrm{Gal}(f) = \mathrm{Gal}(K_f : \mathbb{Q})$.

Note that since every radical can be expressed in terms of prime radicals (e.g., $\sqrt[6]{x} = \sqrt[2]{\sqrt[3]{x}}$), in practice one usually takes the n_i to be primes p_i. This only has the effect of making the tower bigger.

13.2 Relation with Solvable Groups

The first two results, best seen by simply computing an example, tell us that in certain simple cases, splitting fields have cyclic Galois groups.

Theorem 13.2.1. *Let L be the splitting field of $x^m - 1$ over K. Then* $\mathrm{Gal}(L : K) \cong (\mathbb{Z}/m\mathbb{Z})^\times$*, the group of units*[1] *in $\mathbb{Z}/m\mathbb{Z}$. In particular, if $m = p$, a prime, then* $\mathrm{Gal}(x^p - 1)$ *is cyclic.*

Theorem 13.2.2. *Let $f = x^m - a$, for $a \in K$, with K an extension of $\mathbb{Q}$. Then the splitting field K_f contains $\omega = e^{2\pi i/m}$, so it contains all the mth roots of unity. Moreover,* $\mathrm{Gal}(K_f : \mathbb{Q}(\omega))$ *is cyclic . If f is irreducible,* $\mathrm{Gal}(K_f : \mathbb{Q}(\omega)) \cong \mathbb{Z}/m\mathbb{Z}$.

Definition 13.2.3. A finite group G is **solvable** if there exists a chain of subgroups

$$\{e\} = G_0 \subset G_1 \subset \cdots \subset G_m = G$$

such that each G_i is normal in G_{i+1} and G_{i+1}/G_i is cyclic.

Example 13.2.4.

1. Let $G = S_4$, and consider the chain of subgroups

$$\{\mathrm{id}\} \subset < (12)(34) > \subset H \subset A_4 \subset S_4,$$

where $H = \{(12)(34), (13)(24), (14)(23)\} \cong K_4$. Each is normal in the next, and the quotients are, reading from left to right, $\mathbb{Z}/2\mathbb{Z}$, $\mathbb{Z}/2\mathbb{Z}$, $\mathbb{Z}/3\mathbb{Z}$, and $\mathbb{Z}/2\mathbb{Z}$. So S_4 is solvable.

2. Any finite abelian group is solvable, by the classification theorem for finite abelian groups.

We often just say G is solvable if it has a chain of subgroups whose successive quotients are cyclic; just speaking of quotients at all implicitly assumes each is normal in the next.

Lemma 13.2.5. *(Solvability Lemma) Let G be a (finite) group and N a normal subgroup of G. Then*

1. *If G is solvable, every subgroup of G is solvable.*
2. *G is solvable if and only if both N and G/N are solvable.*

The main result, which amongst many other miracles, justifies the terminology "solvable group," is:

Theorem 13.2.6. *If f is solvable by radicals, then* $\mathrm{Gal}(f)$ *is a solvable group.*

Proof. Let $G = \mathrm{Gal}(f)$. The proof is inductive on the length of the radical extension for f. So we consider a radical extension of the form

$$\mathbb{Q} \subset \mathbb{Q}(\alpha) \subset L,$$

with $K_f \subseteq L$ and where α is a root of the irreducible $x^p - a$ ($a \in \mathbb{Q}$). Here p can be assumed prime, as in the comment after definition 13.1.2. We can, by adjoining more elements if necessary[2], assume that L is normal over $\mathbb{Q}$; let us do so (the argument will be the same, but with a longer tower). Our aim is to show that in this situation, $\mathrm{Gal}(f) = \mathrm{Gal}(K_f : \mathbb{Q})$ is a solvable group. But since $\mathbb{Q} \subset L$ is normal, so is $K_f \subset L$, and hence by the fundamental theorem $\mathrm{Gal}(K_f : \mathbb{Q}) \cong \mathrm{Gal}(L : \mathbb{Q}) \,/\, \mathrm{Gal}(L : K_f)$. By the solvability lemma, it's enough now to show that $\mathrm{Gal}(L : \mathbb{Q})$ is solvable.

First of all, we may need to adjoin pth roots of unity. For this, it is sufficient to just adjoin $e^{2\pi i/p}$, for the various powers of this number will produce all the pth roots. So let $\zeta = e^{2\pi i/p}$ and consider the tower

$$\mathbb{Q} \subset L \subset L(\zeta).$$

Since $L(\zeta)$ is a normal extension of L, $\mathrm{Gal}(L(\zeta) : L)$ is a normal subgroup of $\mathrm{Gal}(L(\zeta) : \mathbb{Q})$. By the fundamental theorem,

$$\mathrm{Gal}(L : \mathbb{Q}) \cong \mathrm{Gal}(L(\zeta) : \mathbb{Q}) \,/\, \mathrm{Gal}(L(\zeta) : L).$$

By the solvability lemma, it's enough to show that $\mathrm{Gal}(L(\zeta) : \mathbb{Q})$ is solvable.

[1] Recall that this group consists of those congruence classes $\overline{a}$ such that $\gcd(a, m) = 1$.

[2] This may require adjoining more elements not to L, but to $\mathbb{Q}$ on the bottom - the details are quite fussy.

Next, the extension $\mathbb{Q} \subset \mathbb{Q}(\zeta)$ is normal, with cyclic Galois group, by Theorem 13.2.1, so $\mathrm{Gal}(\mathbb{Q}(\zeta) : \mathbb{Q})$ is solvable (cyclic implies abelian implies solvable). The fundamental theorem applied to $\mathbb{Q} \subset \mathbb{Q}(\zeta) \subset L(\zeta)$ tells us that

$$\mathrm{Gal}(\mathbb{Q}(\zeta) : \mathbb{Q}) \cong \mathrm{Gal}(L(\zeta) : \mathbb{Q}) \,/\, \mathrm{Gal}(L(\zeta) : \mathbb{Q}(\zeta))$$

so by the solvability lemma, it's enough to show that $\mathrm{Gal}(L(\zeta) : \mathbb{Q}(\zeta))$ is solvable. For convenience, let us rewrite the situation as $K \subset K(\alpha) \subset L'$, where $K = \mathbb{Q}(\zeta)$ and $L' = L(\zeta)$. Thus we're looking at the tower

$$K \subset K(\alpha) \subset L',$$

which is just $\mathbb{Q} \subset \mathbb{Q}(\alpha) \subset L$ but with pth roots of unity adjoined all over the place. We know that α is a root of the polynomial $x^p - a$, where $a \in \mathbb{Q}$. But since K now contains all pth roots of 1, this is reducible over $K(\alpha)$:

$$x^p - a = (x - \alpha)(x - \zeta\alpha) \cdots (x - \zeta^{p-1}\alpha)$$

This means exactly that $K(\alpha)$ is the splitting field for $x^p - a$ over $K(\alpha)$, hence normal. Theorem 13.2.2 tells us that $\mathrm{Gal}(K(\alpha) : K)$ is cyclic. Since $K \subset K(\alpha)$ is normal, $H = \Gamma(K(\alpha)) = \mathrm{Gal}(L' : K(\alpha))$ is a normal subgroup of $G = \mathrm{Gal}(L' : K)$, and the fundamental theorem tells us that $G/H \cong \mathrm{Gal}(K(\alpha) : K)$, which is cyclic. So from the first extension, we've produced a normal subgroup with cyclic quotient G/H, which is the first step in proving G is solvable.

We now work inside H, and so are looking at the extension $K(\alpha) \subset L'$, whose Galois group is just H. By induction, we find that H is also solvable. Then since H is solvable, and G/H is cyclic, hence solvable, the solvability lemma (one final time) tells us that G is solvable.

□

13.3 Insolubility of the Quinitc

To put all this together and conclude that there is no general formula for solving the quintic, we need to know what's special about 5 (as opposed to, say, 2 or 3 or 4, for which degrees there *are* general formulae). It turns out to be this fact, mentioned in lecture (or discussion) ages ago: the alternating group[3] A_5 is simple (meaning it has no proper nontrivial normal subgroups). This means it has no "medium-sized" normal subgroup, and implies that S_5 is not solvable, because any chain of normal subgroups would have to have the form

$$\{1\} \subset A_5 \subset S_5,$$

but the first quotient $A_5/\{1\} \cong A_5$ is certainly not cyclic.

Now consider the contrapositive of theorem 13.2.6: If $\mathrm{Gal}(f)$ is not a solvable group, then f is not solvable by radicals. So if we can produce an f of degree 5, whose Galois group is S_5, then we will have exhibited a polynomial which is not solvable by radicals, and this will mean that it is **impossible to find a general formula for solving all quintics!**

The following result says that it is fairly easy to find a quintic with Galois group S_5:

Theorem 13.3.1. *Let $f \in \mathbb{Q}[x]$ be irreducible, and suppose that f has exactly two non-real roots. Then the Galois group* $\mathrm{Gal}(f)$ *is (isomorphic to) S_5.*

Proof. We've seen that elements of the Galois group permute the roots (which are distinct since f is irreducible), so $\mathrm{Gal}(f)$ can certainly be thought of as a subgroup of S_5. Now S_5 can be generated by a transposition of adjacent elements (for example (12)) and a 5-cycle, say (12345), so we'll be done if we can show that in this case $\mathrm{Gal}(f)$ contains one of each. Since the two non-real roots are conjugates of each other, there will be a permutation which interchanges the two and fixes the other roots: this is a transposition in S_5. On the other hand, we can write $\mathrm{Gal}(f)$ as the Galois group of the splitting field K_f. Picking any root α of f, we have a tower $\mathbb{Q} \subset \mathbb{Q}(\alpha) \subset K_f$, and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 5$ (since f is the minimal polynomial for α and f has degree 5). But on the other hand, since K_f is normal over $\mathbb{Q}$, it's normal over $\mathbb{Q}(\alpha)$, and the fundamental theorem says that $\mathrm{Gal}(\mathbb{Q}(\alpha) : \mathbb{Q}) \cong \mathrm{Gal}(K_f : \mathbb{Q}) \,/\, \mathrm{Gal}(K_f : \mathbb{Q}(\alpha)$,

[3] This is the subgroup of S_5 consisting of **even** permutations, i.e., those permutations which can be expressed as the product of an even number of transpositions.

which means that 5 divides $|\mathrm{Gal}(K_f : \mathbb{Q})$. By Cauchy's theorem (homework!), $\mathrm{Gal}(K_f : \mathbb{Q})$ must contain an element of order 5, which is a 5-cycle in S_5. After indexing the roots appropriately, we're done.

□

But can we find a quintic with exactly 3 real roots? Sure - here's how: start with $y = x^5$; its graph looks basically like a cubic, with just one real root, at 0. Now add an x^2 term, to give it a "wiggle". Looking from left to right, it goes up, then down, then up again - so we get a peak and a trough. If necessary, shift it up or down a bit to make sure the peak occurs above the x-axis, and the trough occurs below it. That's the quintic you want. I chose $x^5 + x^2 - 1/4$. It works...

13.4 Conclusion

This is the end of the course. Here's why algebra is important: human beings are only capable of a few things: adding, subtracting, multiplying, dividing, and drawing pictures and trying to discern their symmetries. We hope to have convinced you that these notions are somehow all related, and that algebra, with all its symbols and abstraction, is a language with which to perceive the unity of these activities.

References

The earlier material on rings was standard and followed no particular reference, with the exception of a few proofs more or less shamelessly lifted form **Algebra** by Michael Artin. The treatment of Galois theory, in which I tried to produce the shortest path to the fundamental theorem, used a strange and simplified blend of the treatments of the subject in the following texts:

1. **Fields and Galois Theory**, John M. Howie;
2. **Algebra**, Thomas W. Hungerford;
3. **Number Fields**, Daniel A. Marcus