Abstract Algebra II

Paul Melvin
Bryn Mawr College
Spring 2020

lecture notes loosely based on Dummit and Foote's text

Abstract Algebra (3rd edition)

Prerequisite: Algebra I (Math 303)

Rough course outline:

I Rings (Ch 7–9)
II Vector Spaces (Ch 11)
III Modules (Ch 10, 12)
IV Fields (Ch 13–14)

I Rings

§1. Basics

Assume familiarity with

Definition ring R, ring morphism $f : R \to S$ (and all refinements epi/mono/iso/endo/automorphism), $\ker(f)$ and $\mathrm{im}(f) = f(R)$, subring $S < R$, ideal $J \lhd R$, quotient ring R/J

Examples of rings: $\mathbb{Z}$, $\mathbb{Z}_n$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}, \mathbb{H}$, $R[x]$, $M_n(R)$, $R \times S$

morphisms: "evaluation" $R[x] \to R, f \to f(a)$ (for fixed $a \in R$)

the "natural projection" $R \to R/J, r \mapsto r + J$

ideals: the "trivial" ones 0 and R in any ring R, $\ker(f) \lhd R$ for $f : R \to S$

Remarks ① $\ker(f) = 0 \Longleftrightarrow f$ is 1-1 (i.e. f is a monomorphism)

② If $f : R \to S$ is a ring morphism, then

a) (First Isomorphism Theorem) $R/\ker(f) \cong f(R)$ (via the map sending $r + \ker(f)$ to $f(r)$)

b) (Correspondence Theorem) $J \leftrightarrow f(J)$ is a 1-1 corresp between the ideals in R that contain $\ker(f)$, and the ideals in $\mathrm{Im}(f)$.

Proof ① If $\ker(f) = 0$, then $f(x) = f(y) \Longrightarrow f(x-y) = f(x) - f(y) = 0 \Longrightarrow x - y = 0$, i.e. $x = y$. Thus f is 1-1. The converse is obvious.

② a) Exercise. b) The inverse of $J \mapsto f(J)$ is the map sending any $L \lhd f(R)$ to $f^{-1}(L)$, the full preimage of L under f:

$\boxed{ff^{-1}(L) = L}$ $\subset$ holds in general, and $\supset$ holds since $L \subset f(R)$

($s \in L \Longrightarrow s = f(r),\ r \in R$, i.e. $r \in f^{-1}(s) \Longrightarrow s = f(r) \in ff^{-1}(L)$)

$\boxed{f^{-1}f(J) = J}$ The inclusion $\supset$ holds in general (for *any* function f), and $\subset$ holds since $\ker(f) \subset J$ ($r \in f^{-1}f(J) \Longrightarrow f(r) = f(j)$ for some $j \in J \Longrightarrow f(r-j) = 0$, i.e. $r - j \in \ker(f) \subset J \Longrightarrow r \in J$).

It remains to show $J \lhd R \Longleftrightarrow f(J) \lhd f(R)$: ($\Longrightarrow$) $f(r) \in f(R), f(j) \in f(J) \Longrightarrow f(r)f(j) = f(rj) \in f(J)$, ($\Longleftarrow$) $r \in R, j \in J \Longrightarrow f(rj) = f(r)f(j) \in f(J) \Longrightarrow rj \in f^{-1}f(J) = J$. □

Further notions in a ring R (treated in the first homework assignment)

a) sum and product of ideals $A, B \lhd R$:

$$A + B = \{a + b \mid a \in A,\ b \in B\}$$
$$AB = \{a_1b_1 + \cdots + a_nb_n \mid a_i \in A,\ b_i \in B\}$$

Both are ideals in R (exercise).

Example If $R = \mathbb{Z}$, $A = a\mathbb{Z}$ and $B = b\mathbb{Z}$, then $A + B = \gcd(a, b)\mathbb{Z}$ and $AB = ab\mathbb{Z}$.

b) special proper ideals $J \underset{\neq}{\trianglelefteq} R$ <u>maximal</u>: $J \subset K \underset{\neq}{\trianglelefteq} R \Longrightarrow J = K$

<u>prime</u>: (R commutative) $ab \in J \Longrightarrow a \in J$ or $b \in J$ (for $a, b \in R$)
(in general) $AB \subset J \Longrightarrow A \subset J$ or $B \subset J$ (for $A, B \lhd R$)

c) special elements $r \in R$ <u>zero divisors</u>: $r \neq 0$ and $\exists\ s \neq 0$ with $rs = 0$ or $sr = 0$

<u>nilpotent elements</u>: $r^n = 0$ for some $n \in \mathbb{N}$

<u>units</u> (if R has 1): $\exists\ s \in R$ with $rs = sr = 1$ (a <u>2-sided</u> <u>inverse</u> of r)
Two sided inverses are unique ($s = srs' = s'$) so usually denoted r^{-1}.

<u>Remarks</u> ① As will be seen in the homework, $A + B$ is the smallest ideal containing both A and B, $A \cap B$ is largest contained in both, and $AB \subset A \cap B$ (but not necessarily equal).

② In a commutative ring R with 1, every maximal ideal is prime,[†] but not conversely (e.g. $0 \lhd \mathbb{Z}$ is prime but not maximal; also see homework).

③ Any nonzero nilpotent element is a zero divisor, but not conversely. For example in $\mathbb{Z}_{12}$ we have $2 \cdot 6 = 0$, so both 2 and 6 are zero divisors. In this ring, 2 is not nilpotent (no power of 2 is divisible by 6) but 6 is ($6^2 = 0$). The set of all nilpotent elements in a ring is explored in the homework.

④ In a ring with 1, the sets R° of all zero divisors and $R^\bullet$ of all units are disjoint (exercise). $R^\bullet$ is a group under multiplication (exercise). For example, $M_n(R)^\bullet$ (R a commutative ring) is the group of matrices whose determinants are units in R; we explore this group in the homework, where you are asked to show (among other things) that $M_2(\mathbb{Z}_2)^\bullet \cong S_3$.

<u>Definition</u> Let R be a commutative ring with $1 \neq 0$. R is called an <u>integral</u> <u>domain</u> (or simply <u>domain</u>) if it has no zero divisors (which means $ab = 0 \Longrightarrow a = 0$ or $b = 0$, or equivalently $ab = ac$ and $a \neq 0 \Longrightarrow b = c$). R is called a <u>field</u> if every nonzero element is a unit. Thus field $\Longrightarrow$ domain, but not conversely (e.g. $\mathbb{Z}$ is a domain that is not a field).

<u>Remarks</u> ① R is a domain $\Longleftrightarrow \{0\} \lhd R$ is prime (exercise)

② R is a field $\Longleftrightarrow$ R has no nontrivial proper ideals (exercise)

③ An ideal $J \underset{\neq}{\trianglelefteq} R$ is prime $\Longleftrightarrow R/J$ is a domain, and is maximal $\Longleftrightarrow R/J$ is a field (homework, using ② and the correspondence theorem)

$\boxed{\text{HW\#1}}$ (Arithmetic of Ideals) Let A and B be ideals in a ring R. You may assume that it has been shown already that $A + B$ and AB are ideals.

(a) Prove that $A + B$ is the smallest ideal of R containing both A and B.

(b) Prove that $AB \subset A \cap B$.

(c) Prove that if R is commutative with 1 and $A + B = R$, then $AB = A \cap B$, but that in general equality may fail (give an example).

$\boxed{\text{HW\#2}}$ (Prime/Maximal Ideals) Let R be a commutative ring with $1 \neq 0$.

[†] If $J \lhd R$ is maximal and $ab \in J$ with $a \notin J$, then $K = \{j + ra \mid j \in J, r \in R\}$ is an ideal (verify this) and so by maximality must be all of R. Thus $1 = j + ra$ for suitable j, r, so $b = jb + rab \in J$. Note: without $1 \in R$, this may fail; e.g. $4\mathbb{Z} \lhd 2\mathbb{Z}$ is maximal but not prime.

(a) Prove that a proper ideal $J \subsetneqq R$ is prime $\Longleftrightarrow$ R/J is a domain, and is maximal $\Longleftrightarrow$ R/J is a field.

(b) Give another proof (different from the notes) that any maximal ideal in R is prime, using (a).

(c) Prove using (a) that $(x) \lhd \mathbb{Z}[x]$ is prime but not maximal.

$\boxed{\text{HW\#3}}$ (Nilpotence) Let $\mathcal{N}$ be the set of nilpotent elements in a ring R.

(a) Prove that if R is commutative, then $\mathcal{N}$ is an ideal in R which is contained in *every* prime ideal in R. Hint: Use the binomial theorem to show closure under addition. (It can be shown that $\mathcal{N}$ is in fact the intersection of all prime ideals in R; it is called the *nilradical* of R.)

(b) Show by example that $\mathcal{N}$ need not be an ideal if R is not commutative. (There is an easy example in the ring of integer 2×2-matrices.)

(c) Show that if R is commutative with identity 1, then any sum $u + x$, where u is a unit and $x \in \mathcal{N}$, is a unit. (Hint: first show that $1 - x$ is a unit by factoring $1 - x^n$ for suitable n.)

$\boxed{\text{HW\#4}}$ (Units) Recall that if R is a commutative ring with $1 \neq 0$, the group $M_n(R)^\bullet$ consists of all matrices $A \in M_n(R)$ with $\det(A) \in R^\bullet$.

(a) List the elements in $M_2(\mathbb{Z}_2)^\bullet$, and give an explicit isomorphism $M_2(\mathbb{Z}_2)^\bullet \to S_3$ (where S_3 is the symmetric group of degree 3).

(b) Find the order of the group $M_2(\mathbb{Z}_4)^\bullet$.

$\boxed{\text{HW\#5}}$ Prove that every finite integral domain R is a field. (Hint: For any $r \neq 0$ in R, consider the map $R \to R,\ x \mapsto rx$)

Quadratic Integer Rings

Fix a "square free" integer d (i.e. d is not divisible by the square of any prime) and consider the integral domain

$$\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} \mid a, b \in \mathbb{Z}\}$$

which equals $\mathbb{Z}$ if $d = 1$, is a dense subset of $\mathbb{R}$ if $d > 1$,[†] and is a "lattice" in $\mathbb{C}$ if $d < 0$ (picture).

Remarks ① $\mathbb{Z}[\sqrt{d}]$ is closely related to a classically studied ring $\mathcal{O}_d$ of "quadratic integers" (see page 229 in the text for the definition). In fact

$$\mathcal{O}_d = \begin{cases} \mathbb{Z}[\sqrt{d}] & \text{when } d \equiv 2 \text{ or } 3 \pmod 4 \\ \mathbb{Z}[(1+\sqrt{d})/2] & \text{when } d \equiv 1 \pmod 4 \end{cases}$$

(Note that $d \not\equiv 0 \pmod 4$ since d is square free.) In particular, $\mathcal{O}_{-1} = \mathbb{Z}[i]$, the Gaussian integers.

② the expression $a + b\sqrt{d}$ for elts in $\mathbb{Z}[\sqrt{d}]$ is unique, i.e.

$$a + b\sqrt{d} = a' + b'\sqrt{d} \Longrightarrow a = a',\ b = b'$$

[†] This follows from a classical result of Oresme from the 14th century (proved using the "pigeon-hole" principle) that for any irrational number r (in our case $r = \sqrt{d}$), the orbit of a point on the circle under all rotations by integer multiples of $2\pi r$ is dense in the circle.

since $\sqrt{d}$ is irrational (because d is square free).

Define the algebraic norm $N : \mathbb{Z}[\sqrt{d}] \to \mathbb{Z}$ by

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

or equivalently $N(x) = |x\overline{x}|$ where by definition

$$\overline{a + b\sqrt{d}} = a - b\sqrt{d}.$$

(Note: the absolute values are not needed when $d < 0$.)

If $d < 0$ then $\overline{x}$ is just the complex conjugate of x, so in this case

$$N(x) = |x|^2,$$

the square of the length of x. This observation is useful for determining the units in $\mathbb{Z}[\sqrt{d}]$, and more generally for factoring in $\mathbb{Z}[\sqrt{d}]$ (see below).

In general the map $x \mapsto \overline{x}$ is an involution (i.e. an endomorphism of order two) on $\mathbb{Z}[\sqrt{d}]$ and so N is "multiplicative"

$$N(xy) = N(x)N(y)$$

(proof: $N(xy) = |xy\overline{xy}| = |xy\overline{x}\,\overline{y}| = |x\overline{x}||y\overline{y}| = N(x)N(y)$).

Beware: N is not additive, i.e. $N(x + y)$ need not equal $N(x) + N(y)$.

Application Find $\mathbb{Z}[\sqrt{d}]^\bullet$. Key observation:

$$x \text{ is a unit in } \mathbb{Z}[\sqrt{d}] \iff N(x) = 1$$

Proof: If x is a unit, then by hypothesis the complex number $1/x$ is in $\mathbb{Z}[\sqrt{d}] \Longrightarrow N(x)N(1/x) = N(1) = 1 \Longrightarrow N(x) = 1$. Conversely, $N(x) = 1 \Longrightarrow 1/x = \overline{x}/(x\overline{x}) = \pm\overline{x}/N(x) = \pm\overline{x} \in \mathbb{Z}[\sqrt{d}]$, so x is a unit. □

So the units $a + b\sqrt{d}$ correspond to the solutions a, b to Pell's Equation:

$$a^2 - b^2 d = \pm 1$$

It follows that there are only finitely many units when $d < 0$ (homework).

In contrast, this equation can be used to find infinitely many units when $d > 0$. For example, $u = 1 + \sqrt{2}$ is a unit in $\mathbb{Z}[\sqrt{2}]$ since $1^2 - 1^2 \cdot 2 = -1$, and so the powers of u form an infinite family of units (these are distinct since $|u| \neq 1$). You are asked in the homework to carry out an analogous argument for $d = 3, 5, 6$ and 7.

$\boxed{\text{HW\#6}}$ ⓐ Show that $\mathbb{Z}[\sqrt{d}]$ has only finitely many units for each $d < 0$, and find them all. (Hint: use the geometry of $\mathbb{C}$)

ⓑ Find infinitely many units in $\mathbb{Z}[\sqrt{d}]$ for $d = 3, 5, 6$ and 7.

§2. Principal Ideal Domains and Unique Factorization

For any subset S of a ring R, there is a unique *smallest* ideal in R containing S, namely the intersection of all ideals containing S. This ideal is called the ideal generated by S, and is denoted by (S). If S has only one element s, then $(S) = (s)$ is called a principal ideal.

If R is commutative with 1, then can describe (S) explicitly by

$$(S) = \{r_1 s_1 + \cdots + r_n s_n \mid r_i \in R,\ s_i \in S\}$$

(exercise; for $\subset$ observe that any $s \in S$ can be written as $1s$). In particular the principal ideal $(s) = Rs = \{rs \mid r \in R\}$.

Definition A principal ideal domain (PID) is a domain in which all ideals are principal.

Examples ① Any field F is a PID: the only ideals are (0) and $F = (1)$.

② $\mathbb{Z}$ is a PID. Proof: Given $J \lhd \mathbb{Z},\ J \neq 0$, choose a smallest positive element $m \in J$. Then m divides every $j \in J$. (If not, some $j = qm + r$ with $0 < r < m$. But then $r = j - qm \in J \Longrightarrow\!\Longleftarrow$). Thus $J = (m)$.

③ $F[x]$ is a PID for any field F. Proof. Given $J \lhd F[x],\ J \neq 0$, choose a non-zero polynomial $f \in J$ of smallest degree. Then any $g \in J$ is a multiple of f, i.e. $J = (f)$. (If not, some $g = qf + r$ with $r \neq 0$, $\deg(r) < \deg(f)$, by the "division algorithm" – see below. But then $r = g - qf \in J \Longrightarrow\!\Longleftarrow$.)

④ Not all quadratic integer rings are PID's (for $d < 0$, $\mathcal{O}_d = \text{PID} \iff |d| = 1, 2, 3, 7, 11, 19, 43, 67$ or 163: very hard result of H.Stark 1967; unknown for $d > 0$ – even whether $\exists$ finitely many such d's). One approach is via "Euclidean norms", generalizing ② and ③.

Definition A norm on a domain R is a function

$$N : R \to \mathbb{Z}$$

satisfying $N(0) = 0$ and $N(r) \geq 0$ for all $r \in R$; it is positive if $N(r) > 0$ for $r \neq 0$. The norm N on R is Euclidean if it satisfies the "division algorithm" below, in which case the pair (R, N) is called a Eucidean domain (ED).

Division Algorithm $\forall\ a, b \in R - 0$, $\exists\, q, r \in R$ with $r = 0$ or $N(r) < N(b)$ such that $a = bq + r$. (Note: q, r need not be unique, cf. HW #7a below.)

Theorem 2.1 *ED $\Longrightarrow$ PID*

Proof Suppose (R, N) is a ED. Given $0 \neq J \lhd R$, choose $m \in J - 0$ with $N(m)$ minimal. Then every $j \in J$ is divisible by m. (If not, some $j = qm + r$ with $r \neq 0$ and $N(r) < N(m)$. But then $r = j - qm \in J \Longrightarrow\!\Longleftarrow$). Thus $J = (m)$. $\square$

Examples ① $(\mathbb{Z}, |\ |)$ and $(F[x], \deg)$ are ED's.

② $(\mathcal{O}_d, N)$[†] $= ED$ (and thus a PID) if and only if d is one of the numbers $-11, -7, -3, -2, -1, 2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73$ (Inkeri 1949, Chatland-Davenport 1950). $\exists$ examples, e.g. $\mathcal{O}_{69}$, that are Euclidean but not wrt N (Clark 1994).

[†] Here N is the algebraic norm, $N(a + b\sqrt{d}) = |a^2 - b^2 d|$.

The case $d < 0$ uses the geometry of $N(x) = |x|^2$ (which is defined and multiplicative on all of $\mathbb{C}$). For example, if $d = -1$ (the Gaussian integers), then for $a, b \in \mathbb{Z}[i]$, let $q =$ closest pt in $\mathbb{Z}[i]$ to a/b. Then $a/b = q + s$ where $s := a/b - q$, $N(s) \le 1/2$ (by geometry). Multiplying by b get

$$a = qb + r$$

where $r = sb$, and $N(r) = N(s)N(b) \le N(b)/2 < N(b)$. So $(\mathbb{Z}[i], N)$ is Euclidean (and thus a PID).

$\boxed{\text{HW\#7}}$ Show that (a) $(\mathcal{O}_{-2}, N)$ is Euclidean, while (b) $(\mathcal{O}_{-5}, N)$ is not (hint: try to apply the division algorithm to $a = 1 + \sqrt{-5}$, $b = 2$)

Challenge Show that $\mathcal{O}_{-5}$ is not a Euclidean domain (wrt *any* norm)

Remark (J. Green 1997) There is a characterization of PID's in terms of norms: A domain R is a PID $\iff \exists$ a positive Dedekind-Hasse norm N on R, meaning that $\forall a, b \in R$, either

$$(1)\ \ b|a \quad \text{or} \quad (2)\ \ \exists\ r \neq 0 \text{ in the ideal } (a,b) \text{ with } N(r) < N(b).$$

Divisor Theory in commutative rings R

Definition Let R be commutative with $1 \neq 0$, and a, b be non-zero elements of R. We say

ⓐ b divides a, written $b|a$, if $a = rb$ for some $r \in R$, or equivalently in terms of ideals, $(a) \subset (b)$.

ⓑ a, b are associates, written $a \sim b$, if $a|b$ and $b|a$, or equivalently, $(a) = (b)$. For example the associates of 1 are the units in R: $(u) = R \iff u \in R^\bullet$. If R is a domain, then $a \sim b \iff a = ub$ for some $u \in R^\bullet$ (exercise).

ⓒ d is a greatest common divisor of a and b, written $d = \gcd(a, b)$, if d is a common divisor of a and b, and any common divisor of a and b divides d, or equivalently $(a), (b) \subset (d)$ and $(a), (b) \subset (c) \implies (d) \subset (c)$, i.e. (d) *is the smallest principal ideal containing* (a, b). Such an ideal need not exist[†] and so gcd's need not exist. Note however that if they exist, then any two are associates, by ⓑ.

Examples of gcd's ① In $\mathbb{Z}$, the Euclidean Algorithm (repeated appl of the division algorithm) finds $d =$gcd(a, b) quickly: Start with $a > b$. Then

$$\begin{aligned} a &= \Box\, b + r_1 \\ b &= \Box\, r_1 + r_2 \\ r_1 &= \Box\, r_2 + r_3 \\ &\ \vdots \\ r_{n-2} &= \Box\, r_{n-1} + r_n \\ r_{n-1} &= \Box\, r_n \end{aligned}$$

where the $\Box$'s indicate appropriate quotients, with remainders $r_1 > \cdots > r_n$.

Claim: $d = r_n$. Proof: Working up from the bottom we have

$$r_n | r_{n-1} \implies r_n | r_{n-2} \implies \cdots \implies r_n | r_1 \implies r_n | b \implies r_n | a$$

[†] e.g. there is no smallest principal ideal in $\mathbb{Z}[\sqrt{-5}]$ containing $(2, 1 + \sqrt{-5})$ (exercise).

and working down from the top $c|a, b \Longrightarrow c|r_1 \Longrightarrow c|r_2 \Longrightarrow \cdots \Longrightarrow c|r_n$.

Moreover, working up from the bottom we can efficiently find $r, s \in \mathbb{Z}$ such that $\gcd(a, b) = ra + sb$. Indeed the penultimate equation gives $d = r_n$ as a linear combination of r_{n-1}, r_{n-2}, and the equation before then gives r_{n-1}, and thus d, as a linear combination of r_{n-2}, r_{n-3}, etc. For example if $a = 28$ and $b = 10$, then

$$\begin{aligned} 28 &= 2 \cdot 10 + 8 \\ 10 &= 1 \cdot 8 + 2 \\ 8 &= 4 \cdot 2 \end{aligned}$$

so $\gcd(28, 10) = 2 = 10 - 8 = 10 - (28 - 2 \cdot 10) = -1 \cdot 28 + 3 \cdot 10$.

② Similarly in $F[x]$, or any Euclidean domain, can use the division algorithm repeatedly to find the gcd of any two elements, and to express it as a linear combination (with coefficients in the domain) of the two elements. More generally:

Lemma 2.2 (gcd's in PID's) *Let R be a PID. Then any two nonzero elements a, b have a gcd which can be expressed as a linear combination of a and b with coefficients in R.*

Proof The ideal $(a, b) = (a) + (b)$ is principal, since R is a PID, so $(a, b) = (d)$ for some d. Thus (d) is the smallest principal ideal containing (a, b), i.e. $d = \gcd(a, b)$, and since $d \in (a, b)$, it follows that d is a linear combination of a and b. □

Exercise Define the notion of least common multiple (lcm) and show that any two nonzero elements a, b in a PID have an lcm.

Lemma 2.3 *Let R be a PID and J be a nonzero ideal in R. Then J prime $\Longleftrightarrow$ J maximal.*

Proof ($\Longrightarrow$) By hypothesis $J = (p)$ for some $p \neq 0$. Choose an arbitrary ideal $(m) \supset (p)$. Claim $(m) = (p)$ or R. Well $p \in (m) \Longrightarrow p = rm$ for some nonzero $r \in R \Longrightarrow r$ or $m \in (p)$ (since (p) is prime). Case ① $r \in (p)$. Then $r = ps$ some $s \in R \Longrightarrow r = rms \Longrightarrow sm = 1$ (since R is a domain) $\Longrightarrow m \in R^\bullet$, so $(m) = R$. Case ② $m \in (p)$. Then $(m) \subset (p)$, so $(m) = (p)$. Thus (p) is maximal.

($\Longleftarrow$) If (m) maximal, then $R/(m)$ is a field $\Longrightarrow R/(m)$ is a domain $\Longrightarrow (m)$ prime (by HW2b). □

Definition A nonzero, nonunit element p in a domain R is called

① prime if $p|ab \Longrightarrow p|a$ or $p|b$.
Equivalently, in terms of ideals, $ab \in (p) \Longrightarrow a$ or $b \in (p)$, i.e. *the ideal (p) is prime*

② irreducible if $p = ab \Longrightarrow a$ or b is a unit in R.
Since R is a domain, this means $a|p \Longrightarrow a$ is a unit or an associate of p, or in terms of ideals, $(p) \subset (a) \Longrightarrow (a) = R$ or $(a) = (p)$, i.e. *the ideal (p) is maximal among proper principal ideals*

Corollary 2.4 *In a PID (e.g. in $\mathbb{Z}$), a nonzero nonunit is prime $\Longleftrightarrow$ it is irreducible*

Proof This is immediate from 2.3 and the ideal characterization of the definitions.[†] □

More generally for elements in a domain, prime $\Longrightarrow$ irreducible (HW8) but $\Longleftarrow$ may fail (HW9).[†]

[†] Here's another proof that does not use the ideal characterization of the definitions: Suppose $p|ab$. Since p is irreducible, $\gcd(p, a) \sim p$ or a unit $\Longrightarrow$ either $p|a$, and we're done, or $rp + sa = 1$ for some r, s (by Lemma 2.2), which implies $rpb + sab = b$, and so $p|b$.

[†] The situation is more complicated for rings that are not domains. For example in $\mathbb{Z}_n$, irreducible $\Longrightarrow$ prime *always*, but prime $\Longrightarrow$ irreducible only *sometimes* (when n is a square?)

$\boxed{\text{HW\#8}}$ Show that in any domain, prime elements are always irreducible (cf. the proof of 2.3).

$\boxed{\text{HW\#9}}$ Show $3 \in \mathbb{Z}[\sqrt{-5}]$ is ⓐ irreducible, but ⓑ not prime.
(Hint: use the norm, and note that $9 = 3^2 = (2+\sqrt{-5})(2-\sqrt{-5})$).

Remarks ① In $\mathbb{Z}[\sqrt{d}]$, the multiplicativity of the norm N, and the fact that $N(x) = 1 \iff x$ is a unit, make N a useful tool in investigating questions of irreducibility and primality. In particular these properties show that *if $N(x)$ is not the product of two smaller norms of elements in $\mathbb{Z}[\sqrt{d}]$* (e.g. if $N(x)$ is prime in $\mathbb{Z}$), *then x is irreducible in $\mathbb{Z}[\sqrt{d}]$*.

② Primes in $\mathbb{Z}$ need not remain prime in $\mathbb{Z}[\sqrt{d}]$. For example 2 and 5 are *not* prime in $\mathbb{Z}[i]$, since $2 = (1+i)(1-i)$ and $5 = (2+1)(2-i)$, but 3 is, e.g. because $N(3)$ is not the product of smaller norms of elements in $\mathbb{Z}[i]$ (to see this, record some norms on the lattice to see a pattern).

Definition A unique factorization domain (UFD) is a domain R for which every element r which is nonzero and not a unit can be "uniquely" factored into irreducibles:

($\exists$) $r = p_1 \cdots p_n$ for suitable irreducible p_i's

(!) this decomposition is unique up to associates, i.e. $r = q_1 \cdots q_k$ with q_i irreducible $\implies n = k$ and (after renumbering) $p_i \sim q_i$ for all i.

Theorem 2.5 *PID $\implies$ UFD*

Proof Let R be a PID and r be a nonzero, nonunit element of R. Note that by Corollary 2.4, irreducible = prime in R.

($\exists$) Suppose some r cannot be factored into primes. In particular r is not prime, so it is reducible: $r = a_1 b_1$ with $a_1, b_1 \notin R^\bullet$. Now at least one of a_1 or b_1 has no prime factorization (since otherwise, r would have one), say a_1. In the same way $a_1 = a_2 b_2$ where a_2 has no prime factorization and $b_2 \notin R^\bullet$, etc. This gives a strictly ascending sequence of ideals

$$(a_1) \underset{\neq}{\subsetneq} (a_2) \underset{\neq}{\subsetneq} (a_3) \underset{\neq}{\subsetneq} \cdots$$

But this cannot be: The union $\cup(a_i)$ is an ideal so $= (a)$ for some a, since R is a PID. Thus a lies in some (a_n), which implies $(a_{n+1}) \subset (a) \subset (a_n)$, so $(a_{n+1}) = (a_n)$, a contradiction.

(!) If $p_1 \cdots p_n = q_1 \cdots q_k$ then $p_1 \,|\, q_1 \cdots q_k \implies p_1 \,|$ some q_i (say q_1 after reordering if necessary) since p_1 is prime, $\implies p_1 \sim q_1$ since q_1 is irreducible. We now cancel p_1 and q_1 (up to a unit multiplier) and then proceed by induction. □

Remarks ① In this proof, showed PID satisfies the ascending chain condition (ACC) for ideals:

$$J_1 \subset J_2 \subset \cdots \implies \exists\, n,\ J_n = J_{n+1} = \cdots\ ,$$

which is equivalent to the condition that every ideal is finitely generated (exercise). Any domain with this property is called a Noetherian domain (named after Emmy Noether, who taught at Bryn Mawr in the 1940's and is buried in the Cloisters).

② UFD $\not\Longrightarrow$ PID, e.g. $\mathbb{Z}[x]$ is a UFD (see §3) but not a PID, e.g. $J = (2, x)$ is not principal. (Proof: If $J = (f)$ for some polynomial f, then 2 would be a multiple of $f \Longrightarrow f \equiv \pm 1$ or ± 2, by degree considerations. But $f = \pm 1 \Longrightarrow (f) = \mathbb{Z}[x] \Longrightarrow\!\Longleftarrow$, since all polys in J have even constant term, and $f = \pm 2 \Longrightarrow 2|x \Longrightarrow\!\Longleftarrow$.)

Theorem 2.5 $\Longrightarrow$

<u>Fundamental Theorem of Arithmetic</u> $\mathbb{Z}$ *is a UFD*

<u>Corollary</u> (Euclid) $\exists$ *infinitely many primes in* $\mathbb{Z}$

<u>Proof</u> If not, let $p =$ largest prime. then $p! + 1$ leaves remainder of 1 upon division by any prime, hence has no prime factorization $\Longrightarrow\!\Longleftarrow$ $\square$

<u>Remarks</u> ① UFD's have gcd's and lcm's, namely

$$\gcd(\prod p_i^{r_i}, \prod p_i^{s_i}) = \prod p_i^{\min(r_i, s_i)} \qquad \operatorname{lcm}(\prod p_i^{r_i}, \prod p_i^{s_i}) = \prod p_i^{\max(r_i, s_i)}$$

(allowing $r_i, s_i = 0$).

② prime $\Longleftrightarrow$ irreducible for elements in a UFD

Proof: ($\Longrightarrow$) HW#8 ($\Longleftarrow$) p irred, $p|ab \Longrightarrow ab = pc$, some c. Writing a, b, c as products of irreducibles, uniqueness shows $p|a$ or $p|b$. $\square$

However $\exists$ nonzero prime ideals which are not maximal in some UFD's (e.g. $(x) \lhd \mathbb{Z}[x]$ is not maximal, since $(x) \subsetneq (2, x) \lhd \mathbb{Z}[x]$, or alternatively, since $\mathbb{Z}[x]/(x) \cong \mathbb{Z}$ is not a field)

③ Many interesting rings (e.g. imaginary quadratic rings $\mathbb{Z}[\sqrt{-d}]$ and cyclotomic rings $\mathbb{Z}[e^{2\pi i/n}]$ for large d, n) are not UFD's (Cauchy-Lamé's mistake in attempts on Fermat's Last theorem was to assume so) but most have unique prime factorization *for ideals* (commutative rings with 1 having this property are called <u>Dedekind</u> <u>domains</u>)

$\boxed{\text{HW\#10}}$ Show that $\mathbb{Z}[\sqrt{-5}]$ is not a unique factorization domain. (Hint: use homework #9)

$\boxed{\text{HW\#11}}$ ⓐ Show that $\mathbb{Z}[i]$ is a unique factorization domain.
ⓑ Explain why $10 = (3 + i)(3 - i) = 2 \cdot 5$ does not contradict unique factorization in $\mathbb{Z}[i]$.
ⓒ (Library research) Describe (without proof) all the primes in $\mathbb{Z}[i]$.

§3. Polynomial Rings

<u>Definition</u> R be a commutative ring with $1 \neq 0$. The <u>polynomial ring</u> over R in one variable x is defined by

$$R[x] = \{\textstyle\sum_{i=0}^{n} a_i x^n \mid a_i \in R\}$$

with addition and multiplication of polynomials defined in the usual way.[†] More generally, define polynomial rings $R[x_1, \ldots, x_n]$ in more variables in the "obvious" way, or inductively by $R[x_1, \ldots, x_{n-1}][x_n]$.

For $f(x) = \sum_{i=0}^{n} a_i x^i$ with $a_n \neq 0$, define the <u>degree</u> $\deg(f) = n$ and call a_n the <u>leading coefficient</u> of f. The <u>value</u> of f at any $c \in R$ is $f(c) := \sum a_i c^i \in R$.

[†] $\sum_i a_i x^i + \sum_i b_i x^i = \sum_i (a_i + b_i) x^i$ and $\sum_i a_i x^i \cdot \sum_j b_j x^j = \sum_{i,j} a_i b_j x^{i+j}$.

<u>Remarks</u> (exercises) ① If R is a domain, then

ⓐ $\deg(fg) = \deg(f) + \deg(g)$ ⓑ $R[x]^\bullet = R^\bullet$ ⓒ $R[x]$ is a domain.

② In general the map $R[x] \to R,\ f \mapsto f(c)$ (for fixed c) is a ring morphism

<u>Theorem 3.1</u> *If F is a field then $F[x]$ is a ED (and thus a PID and UFD by previous results)*

Immediate from:

<u>Division Algorithm</u> *If f, g are polynomials in $R[x]$, where g is nonzero with invertible leading coefficient (which is automatic if R is a field), then $\exists\, q, r \in R[x]$ with $f = qg + r$, where $r = 0$ or $\deg(r) < \deg(g)$.*

<u>Proof</u> Suppose f and g have degrees n and k and leading coefficients a_n and $b_k \in R^\bullet$, resp. Can assume $n \geq k$ (else take $q = 0$ and $r = f$).

Induct on n. If $n = 0$, then $f = a_0$, $g = b_0$ so take $q = a_0 b_0^{-1}$ and $r = 0$. For $n > 0$, consider $f' = f - cg$, where $c = a_n b_k^{-1} x^{n-k}$, of degree $< n$. By the induction assumption $f' = q'g + r$ for suitable $q', r \in R[x]$ with $r = 0$ or $\deg r < \deg g$. Thus for $q = q' + c$ have $f = qg + r$. □

$\boxed{\text{HW\#12}}$ Let F be a field. ⓐ Show that "irreducible" and "prime" are the same for polynomials in $F[x]$. ⓑ Prove that for $f \in F[x]$, the quotient $F[x]/(f)$ is a field if and only if f is irreducible.

$\boxed{\text{HW\#13}}$ (quotients of $\mathbb{Z}[x]$) Let $f = x^3 - 2x + 1 \in \mathbb{Z}[x]$, and for any $g \in \mathbb{Z}[x]$, let $\overline{g}$ denote the image of g in the quotient ring $\mathbb{Z}[x]/(f)$ (i.e. $\overline{g} = g + (f)$). For $p = 2x^7 - 7x^5 + 4x^3 - 9x + 1$ and $q = (x-1)^4$, express each of the elements $\overline{p}$, $\overline{q}$, $\overline{p+q}$ and $\overline{pq}$ in the form $\overline{g}$ for some polynomial g of degree ≤ 2 (the existence of such a g follows from the division algorithm).

Another consequence of the division algorithm:

<u>Theorem 3.2</u> *Let R be a commutative ring with 1, and f be a nonzero polynomial in $R[x]$ of degree n. Then*

ⓐ $c \in R$ is a root of f (i.e. $f(c) = 0$) $\iff (x - c) | f$, and

ⓑ If R is a domain, then f has at most n roots in R.

<u>Proof</u> ⓐ Div alg $\Longrightarrow f(x) = q(x)(x - c) + r$ with $r \in F \Longrightarrow f(c) = r$. So $f(c) = 0 \Longrightarrow r = 0 \Longrightarrow (x - c)|f$. Conversely $(x - c)|f \Longrightarrow f(x) = q(x)(x - c) \Longrightarrow f(c) = 0$.

ⓑ Induct on n. If $n = 1$, then $f(x) = ax + b$ (with $a \neq 0$) has at most one root since R is a domain. (Indeed, c, d are roots $\Longrightarrow ac = ad \Longrightarrow c = d$.) If $n > 1$ and f has a root c then $f(x) = (x - c)q(x)$ where $\deg(q) = n - 1$, by ⓐ, and q has at most $n - 1$ roots (by the inductive assumption). Since R is a domain, any root of f other than c must also be a root of q, which completes the proof. □

<u>Corollary</u> *Any finite subgroup G of the multiplicative group $R^\bullet$ of units in a domain R is cyclic.*

<u>Proof</u> If G is not cyclic, then it is a product $C_n \times C_m \times \cdots$ of cyclic groups with $n|m|\cdots$. Now all n elements of C_n have order dividing n, as do some of the elements of C_m, so G has more than n elements x satisfying the equation $x^n = 1$. But this contradicts part ⓑ of the theorem. □

Unique factorization in polynomial rings

Theorem 3.3 *If R is a UFD, then so is $R[x]$.*[†]

Idea: Work in the UFD $F[x]$, where F is the "field of fractions" of R:

Definition For any domain R , define the field of fractions of R to be

$$F(R) = \{a/b \mid a, b \neq 0 \ \in \ R\}/\sim$$

where $a/b \sim c/d \iff ad = bc$, with operations

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

$F(R)$ is a field containing R (via $r \mapsto r/1$) to which any ring morphism $f : R \to K$, where K is a field, extends uniquely (draw the diagram). For example $F(R) = R$ if R is a field (such as $\mathbb{Z}_p$ for prime p) and $F(\mathbb{Z}) = \mathbb{Q}$.

Gauss' Lemma *Let R be a* UFD *and F be its field of fractions. If $f \in R[x]$ is reducible over F, then it is reducible over R.* (Here, saying f is "reducible over F" means "reducible as an element of $F[x]$, and likewise for "reducible over R".) *Conversely, if f is reducible over R and is primitive* (meaning the only common factors of its coefficients are units) *then it is reducible over F. Thus the notions of reducibility over R and F are equivalent for primitive polynomials in $R[x]$.*

Example $f = 2x^2 + x - 6 = (\frac{8}{5}x - \frac{12}{5})(\frac{5}{4}x + \frac{5}{2}) = (2x-3)(x+2)$

HW#14 Show that if R is not a field, then you can always find an example of a reducible polynomial in $R[x]$ that is irreducible in $F[x]$. Can you describe *all* such examples?

Prove Gauss' Lemma using

Lemma *If p is a prime element in a domain R, then p (viewed as a constant polynomial) is also prime in $R[x]$*

Proof Suppose $p|fg$ with $f = a_0 + a_1x + \cdots$, $g = b_0 + b_1x + \cdots$. Suppose $p \nmid f$ and $p \nmid g$. Choose smallest i, j with $p \nmid a_i$, $p \nmid b_j$. The coefficient of x^{i+j} in fg is $(\cdots + a_{i-1}b_{j+1}) + a_ib_j + (a_{i+1}b_{j-1} + \cdots)$. But p divides the first and last terms $\implies p|a_ib_j \implies p|a_i$ or $p|b_j \implies\impliedby$. $\therefore p|f$ or $p|g$. □

Proof of Gauss First suppose that $f = GH$ with $G, H \in F[x]$. If G, H are both already in $R[x]$, then just take $g = G$ and $h = H$. Otherwise multiply by rs, where r and s are common multiples of all the denominators in the coefficients of G and H, respectively, to get

$$rsf = g'h'$$

where $g' = rG$ and $h' = sH$ are in $R[x]$. Now any prime factor p of rs divides $rsf = g'h'$ and so divides g' or h' (by the Lemma). An inductive argument then gives a factorization $rs = r's'$ in R for which $r' \mid g'$ and $s' \mid h'$, and so setting $g = g'/r'$ and $h = h'/s'$, both in $R[x]$, we have $f = gh$.

Conversely, any reducible primitive $f \in R[x]$ can be factored over R into a product polynomials of positive degree (the only constant factors are units) $\implies f$ reducible over F as well. □

[†] False for UFD replaced by PID, e.g. $R = \mathbb{Z}$

Proof of 3.3 Any $f \in R[x]$ factors into irreducibles over F by Theorem 3.1, giving a factorization over R using Gauss' Lemma, although the factors need not be irreducible over R. But pulling out the gcd's of the coefficients of each of these factors, we can write $f = rf_1 \cdots f_n$ where $r \in R$ and the f_i are primitive irreducible polynomials in $R[x]$. Now factor $r = p_1 \cdots p_k$ into primes in R. Then

$$f = p_1 \cdots p_k \, f_1 \cdots f_n$$

is a factorization of f into irreducibles in $R[x]$. The uniqueness follows from uniqueness in $F[x]$ and in R. $\square$

Irreducibility Criteria

Cubic Criterion *If R is a domain and f is a primitive polynomial in $R[x]$ of degree ≤ 3, then f is reducible over $R \iff$ it has a root in $F(R)$*

Indeed f reducible/$R \iff$ reducible/$F(R)$ (by Gauss' Lemma, since f is primitive) $\iff f$ has a linear factor (since $n \leq 3$) $\iff f$ has a root in $F(R)$ (by Theorem 3.2).

If $R = \mathbb{Z}_p$ for prime p, then $F(R) = R$ so the roots of f can be found by simply plugging in all the elements of R. If there are none, then f is irreducible, and as an added bonus f can then be used to produce a finite field with $p^{\deg(f)}$ elements, namely $\mathbb{Z}_p[x]/(f)$ (see example ① below).

More generally, if R is any UFD with $R^\bullet$ finite (such as $\mathbb{Z}$), then there are only finitely many elements to check. Indeed the only possible roots are of the form r/s where $r|a_0$ and $s|a_n$ (this is called the rational root test).[†]

Examples ① The polynomial $f(x) = x^2 + x + 1$ is irreducible in $\mathbb{Z}_2[x]$ (since $f(0) = f(1) = 1 \neq 0$). Thus $\mathbb{F}_4 := \mathbb{Z}_2[x]/(f)$ is a field (see HW #12) with four elements, namely

$$\mathbb{F}_4 = \{0,\ 1,\ x,\ x+1\}$$

(all polys of deg $<$ deg(f)) with the usual addition, but multiplication defined mod f. Thus for example $(x+1)^2 = x^2 + 2x + 1 = x + (x^2 + x + 1) = x$. Another way to say this: working mod f means $f = 0$, i.e. $x^2 = x + 1$, so $x^2 + 2x + 1 = x^2 + 1$ (since $2 = 0$) $= (x+1) + 1 = x$.

Note: it is convenient to use "base 2" notation $a_n \cdots a_0$ (where the $a_i = 0$ or 1) for the element $a_n x^n + \cdots + a_0$ of $\mathbb{Z}_2[x]$ or its residue class in $\mathbb{F}_4$. In particular $x = 10$ and $x + 1 = 11$. Addition and multiplication are carried out in the usual way, but working mod 2 and at the end reducing mod $x^2 + x + 1 = 111$. For example the calculation $(x+1)^2 = x$ above becomes $11^2 = 101 = 10$ (since $101 \equiv 10 + 111 \pmod 2$).

② The polynomial $f(x) = 2x^3 + x + 1$ is irreducible in $\mathbb{Z}[x]$ since $f(\pm 1/2) \neq 0$. We only need to look at $\pm 1/2$ by the rational root test since these are the ratios of the divisors ± 1 of $a_0 = 1$ by the divisors ± 2 of $a_n = 2$.

$\boxed{\text{HW\#15}}$ Use the cubic criterion to show

(a) $x^2 + 1$ is irreducible in $\mathbb{Z}_3[x]$. Then use this to explicitly construct a field $\mathbb{F}_9$ with 9 elements and to write down its multiplication table. Use base 3 notation $a_n \cdots a_0$ (where the $a_i = 0$, 1 or 2) for the residue class of $a_n x^n + \cdots + a_0$ in $\mathbb{F}_9$, as above.

(b) $x^3 + 6x + 12$ is irreducible in $\mathbb{Z}[x]$.

[†] Proof: $f(r/s) = 0 \implies 0 = s^n f(r/s) = a_n r^n + \cdots + a_0 s^n \equiv_r a_0 s^n$ and $\equiv_s a_n r^n \implies r|a_0,\ s|a_n$.

Eisenstein Criterion *Let R be a domain and $f(x) = a_n x^n + \cdots + a_0 \in R[x]$ be primitive. If $\exists$ prime ideal P in R such that $a_n \notin P$, $a_i \in P$ for $i < n$, and $a_0 \notin P^2$, then f is irreducible over R.* [Special case: $R = \mathbb{Z}$. If some prime p divides $a_0, \ldots, a_{n-1}$, but $p \nmid a_n$ and $p^2 \nmid a_0$, then f irred/$\mathbb{Z}$.]

Examples ① $x^3 - 9x^2 + 6x - 3$ is irred/$\mathbb{Z}$ $(p = 3)$

② The p^{th}-cyclotomic polynomial (for p prime)

$$\phi_p(x) := \prod_{\substack{\zeta \in \mathbb{C}:\ \zeta^p = 1 \\ \zeta^k \neq 1 \text{ for } 0<k<p}} (x - \zeta) \;=\; \frac{x^p - 1}{x - 1} \;=\; x^{p-1} + x^{p-2} + \cdots + 1$$

is irreducible/$\mathbb{Q}$, and $\therefore$ /$\mathbb{Z}$ by Gauss' lemma. (In fact it's irreducible for nonprime p as well, cf. §13.6 in Dummit and Foote.) Can't apply Eisenstein directly, but the trick is to substitute $y = x - 1$ to get a related polynomial

$$\psi_p(y) := \phi_p(y + 1) = \frac{(y + 1)^p - 1}{y} = y^{p-1} + p y^{p-2} + \binom{p}{2} y^{p-3} + \cdots + p$$

which is irreducible (by Eisenstein) $\Longrightarrow$ ϕ_p is irred (any factorization of ϕ_p gives one for ψ_p).

Proof of Eisenstein The natural morphism $R \to R/P$, $a \to \bar{a} := a + P$, extends to a morphism $R[x] \to (R/P)[x]$,

$$f = a_n x^n + \cdots + a_1 x + a_0 \;\mapsto\; \bar{f} = \bar{a}_n x^n + \cdots + \bar{a}_1 x + \bar{a}_0.$$

Note that R/P is a domain since P is prime.

Now suppose f is reducible over R, i.e. $\exists$ nonconstant (since f is primitive) polynomials $h, g \in R[x]$ with $f = gh$. But then $\bar{f} = \bar{g}\bar{h} = \bar{a}_n x^n$ in $(R/P)[x] \Longrightarrow \bar{g}$ and $\bar{h}$ are monomials (since R/P is a domain) $\Longrightarrow$ constant terms in $\bar{g}$, $\bar{h}$ are $\bar{0} \Longrightarrow$ const terms in g, h are in $P \Longrightarrow a_0 \in P^2 \Longrightarrow\!\Longleftarrow$. $\therefore f$ is irreducible. □

$\boxed{\text{HW\#16}}$ Prove using Eisenstein's criterion that (a) $x^3 + 6x + 12$ and (b) $x^4 + x^3 + x^2 + x + 1$ are irreducible in $\mathbb{Z}[x]$.

Kronecker's algorithm for factoring over $\mathbb{Z}$ (or equivalently over $\mathbb{Q}$)

Fix a primitive polynomial $f \in \mathbb{Z}[x]$. We describe a slow but foolproof algorithm due to Kronecker for factoring f into irreducible polynomials. Note that if f is not primitive, then $f = d \cdot h$ where $d > 1$ is an integer and h is primitive, and so f is factored by factoring d into primes (which can be slow) and applying the algorithm to h.

The idea is to search for a factor g of f of smallest possible degree s, which evidently must be less than or equal to $\deg(f)/2$. Once g is found, we have $f = g \cdot f/g$ and f/g can be factored by reapplying the algorithm.

- $(s = 1)$ Linear factors are found using the rational root test.
- $(s > 1)$ We assume f has no factors of degree $< s$, and look for a factor g of degree s.

Key observations:

1) $g|f \Longrightarrow g(k)|f(k)$ for all $k \in \mathbb{Z}$

2) (Lagrange interpolation) g is determined by its values on any $s+1$ integers $a_0, \ldots a_s$.

So use 2) to find the finitely many polynomials whose values on each a_i is a divisor of $b_i = f(a_i)$. By 1), these are the only candidates for factors of degree s; check if any of them divide f.

We elaborate on 2):

Lagrange interpolation To arrange that $g(a_i) = b_i$ (for $0 \le i \le s$) set $g = \sum_{i=0}^{s} b_i g_i$ where

$$g_j(x) = \prod_{i \neq j} \frac{x - a_i}{a_j - a_i}.$$

(Note that $\deg(g_j) = s$ and $g_j(a_i) = \delta_{ij}$.) To see that g is unique, suppose that $h(a_i) = b_i$ for some h of degree $\le s$. Then $g - h$ has $s+1$ roots (the a_i's) so $g = h$ by Theorem 3.2a.

In practice, finding $g(x) = c_s x^s + \cdots + c_0$ amounts to solving the system of $s+1$ equations $g(a_i) = b_i$ in the $s+1$ unknowns $c_0, \ldots, c_s$, or equivalently the matrix equation $Ac = b$ where $A = (a_i^j)$ (numbering the rows and columns from 0 to s). The solution is then $c = A^{-1}b$. (Note that A is a Vandermonde matrix, cf. Exercise 27 in §14.6 in Dummit-Foote, which is always invertible.) Thus computing A^{-1} allows one to quickly find the g's for many different choices of b's, as is required in Kronecker's algorithm.

Example Factor the polynomial $f(x) = x^5 + x^3 - x^2 - 1$.

- ($s = 1$) We find the linear factor $x - 1$ using the rational root test. Thus $f(x) = (x-1)h(x)$, where $h(x) = x^4 + x^3 + 2x^2 + x + 1$, and using the rational root test again we find that $h(x)$ has no linear factors.

- ($s = 2$) We look for quadratic factors $g(x) = c_2x^2 + c_1x + c_0$ of h by interpolation with inputs $a_0 = -1$, $a_1 = 0$ and $a_2 = 1$. Thus we must solve $Ac = b$ for c, where

$$A = (a_i^j) = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \qquad c = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} \qquad b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$

  and $b_0 = \pm 1$ or 2, $b_1 = \pm 1$, and $b_2 = \pm 1, 2, 3$ or 6, the divisors of $h(-1) = 2$, $h(0) = 1$ and $h(1) = 6$, respectively. We compute A^{-1}, and then find (possibly after some false starts, as there are 64 possibilities for c to check) a solution $c = (1,0,1)^t = A^{-1}(2,1,2)^t$ that corresponds to a factor $x^2 + 1$ of h with $h(x) = (x^2+1)(x^2+x+1)$.

Thus

$$f(x) = (x-1)(x^2+1)(x^2+x+1)$$

is the factorization of f into irreducibles.

$\boxed{\text{HW\#17}}$ Factor (a) $x^5 + x^2 - x - 1$ and (b) $2x^5 - 6x^4 + 5x^3 + 3x^2 - 8x + 2$ as products of irreducible polynomials in $\mathbb{Z}[x]$.

II Vector Spaces

§1. Basics

Definition Let F be a field. A vector space over F (or F-vector space) consists of an additive abelian group V and an operation $F \times V \to V$, $(\alpha, v) \mapsto \alpha v$ (called scalar multiplication) satisfying

V1) $(\alpha + \beta)v = \alpha v + \beta v$ and $\alpha(v + w) = \alpha v + \alpha w$

V2) $(\alpha\beta)v = \alpha(\beta v)$ **V3)** $1v = v$ (where 1 is the multiplicative identity in F)

for all $\alpha, \beta \in F$ and $v, w \in V$. The elements of F and V are called scalars and vectors, respectively. Elementary properties, such as

$$\alpha 0 = 0 = 0v \quad \text{for all } \alpha \in F,\ v \in V,$$

where the first two 0's are the zero vector, and the last is the zero scalar, are easily verified (exercise).

Examples ① $\mathbb{R}^n$ is a real vector space; similarly F^n is an F-vector space for any field F. Also $M_n(F) = \{n \times n\text{-matrices}/F\}$ (verify this).

② The set $C[0,1]$ of continuous functions $[0,1] \to \mathbb{R}$ is a real vector space using the usual addition and scalar multiplication of functions.

③ If F is a subfield of E, then E is a vector space over F.

Definition A function $T : V \to W$ between F-vector spaces is called a homomorphism (or an F-linear transformation or linear map) if

$$\textbf{T1)}\ T(v + w) \ = \ T(v) + T(w) \qquad \textbf{T2)}\ T(\alpha v) \ = \ \alpha T(v)$$

for all $\alpha \in F$ and $v, w \in V$.[†] Compositions of linear maps are linear (exercise). Have the usual mono, epi, iso, endo and auto refinements. A monomorphism is also called an embedding. Spaces V, W are isomorphic, written $V \cong W$, if $\exists$ an isomorphism $V \to W$. For example $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$.

Definition A nonempty subset U of a vector space V is called a subspace, denoted $U < V$, if it is closed under addition and scalar multiplication. Clearly U is then a vector space w.r.t. the induced operations. It is easy to show that intersections of subspaces are subspaces.

Two important subspaces associated to any linear map $T : V \to W$ are its kernel and image:

$$\ker T = \{v \in V : T(v) = 0\} < V \qquad \text{and} \qquad \mathrm{Im}T = \{T(v) : v \in V\} < W.$$

(The reader should verify that these are indeed subspaces.) Easily prove T monic $\iff \ker T = \{0\}$ and T epic $\iff \mathrm{Im}T = W$.

Also of great importance are "eigenspaces" associated with "eigenvalues" of an endomorphism $T : V \to V$: A scalar λ is called an eigenvalue of T if $\exists\, v \neq 0$ in V with $T(v) = \lambda v$, and the set $V_\lambda = \{v \in V \mid T(v) = \lambda v\}$ (which includes the 0 vector) is called the λ-eigenspace of T. Vectors in V_λ are called eigenvectors for λ.

$\boxed{\text{HW\#18}}$ Show that the map $T\colon M_n(\mathbb{R}) \to M_n(\mathbb{R})$ that sends a matrix A to its transpose A^T is linear. Find all its eigenvalues (assuming $n > 1$) and identify their associated eigenspaces.

$\boxed{\text{HW\#19}}$ Show that any eigenspace V_λ of an endomorphism $T\colon V \to V$ is a subspace of V, and that any two distinct eigenspaces V_λ and V_μ of T intersect trivially (i.e. only in the zero vector).

[†] Note that (1) and (2) $\iff$ the single condition $T(\alpha v + w) = \alpha T(v) + T(w)$.

Ways to construct new vector spaces from old:

Quotient Spaces   Start with $U < V$. Then the additive quotient group

$$V/U = \{v + U : v \in V\}$$

(with addition $(v+U)+(w+U) = (v+w)+U$) can be made into a vector space with scalar multiplication defined by $\alpha(v+U) = (\alpha v)+U$. (Verify that this is well defined and that the vector space axioms hold.) This is called the quotient space of V by U. The natural projection

$$p\colon V \to V/U, \qquad v \mapsto v+U$$

is linear (check this). It satisfies the usual universal property[†] which can be used to prove the usual isomorphism theorems (as in group and ring theory).

HW#20 Give a direct proof (without appealing to the universal property of quotient spaces) of the First Isomorphism Theorem: *If $T\colon V \to W$ is linear, then* $V/\ker(T) \cong \mathrm{Im}(T)$. In other words, write down an explicit map $V/\ker(T) \to \mathrm{Im}(T)$ (using the notation $\overline{v}$ for $v+\ker(T)$ to simplify your notation) and then prove that it is a well-defined isomorphism.

Spaces of Homomorphisms   Start with two F-vector spaces V, W. Define

$$\mathrm{Hom}(V,W) = \{\text{all } F\text{-linear maps } V \to W\}.$$

This becomes an F-vector space in its own right with respect to the operations $(S+T)(v) := S(v)+T(v)$ and $(\alpha T)(v) := \alpha(T(v))$ (check this).

Avery important special case (when $W = F$) is the dual space of V

$$V^* = \mathrm{Hom}(V,F) = \{\text{all } F\text{-linear maps } V \to F\}.$$

The elements of V^* are often called linear functionals on V. Note that any linear map $T : V \to W$ induces a linear map $T^* : W^* \to V^*$, called the dual or adjoint of T, given by $T^*(f) = f \circ T$. Much more about this later; see for example HW#21, 23, 25, 26 and 28.

Examples ① The projection $\mathbb{R}^2 \to \mathbb{R}$, $(x,y) \mapsto x$ is an element of $(\mathbb{R}^2)^*$, as are any maps of the form $(x,y) \mapsto ax+by$ for $a, b \in \mathbb{R}$.

② The trace and determinant of matrices can be viewed as maps

$$\mathrm{tr},\ \det : M_n(\mathbb{R}) \to \mathbb{R}.$$

Then tr is in $M_n(\mathbb{R})^*$, while det is not. (Do you see why?)

③ The map

$$I : C[0,1] \to \mathbb{R}\ , \quad f \mapsto \int_0^1 f(x)dx$$

is in $C[0,1]^*$, since $I(\lambda f + g) = \lambda I(f) + I(g)$.

[†] For any linear $T\colon V \to W$ with $U \subset \ker(T)$, $\exists!$ linear $S\colon V/U \to W$ satisfying $T = S \circ p$ (namely $S(v+U) = T(v)$). Furthermore $\ker(S) = \ker(T)/U$ and $\mathrm{Im}(S) = \mathrm{Im}(T)$.

Remark We will see below that any vector space V can be embedded in its dual space V^* (in fact $V \cong V^*$ when V is "finite dimensional"), but there is generally no natural way to find such an embedding. However there is a natural embedding $V \to V^{**}$ given by $v \mapsto e_v$, where $e_v(f) := f(v)$. Think of e_v as "evaluation at v".

$\boxed{\text{HW\#21}}$ Show that the map $e : V \to V^{**}$, $v \mapsto e_v$ where $e_v(f) = f(v)$, does indeed map V to V^{**} (i.e. that e_v is a linear map $V^* \to F$), and that it is an embedding (i.e. an injective linear map).

Direct Sums The direct sum of two F-vector spaces U, V is the vector space

$$U \oplus V = \{(u, v) : u \in U,\ v \in V\}$$

with componentwise operations. If U and V happen to be subspaces of another space W, then can also consider their sum,

$$U + V = \{u + v : u \in U,\ v \in V\}$$

which is a subspace of W.

$\boxed{\text{HW\#22}}$ Show that if U and V are subspaces of a vector space W satisfying (a) $U \cap V = \{0\}$ and (b) $U + V = W$, then $U \oplus V \cong W$.

§2. Linear combinations, bases and dimension

Fix a vector space V, and let S be an arbitrary set of vectors in V.

Definition The span of S, denoted $\langle S \rangle$, is the smallest subspace of V containing S, i.e. the intersection of all subspaces of V containing S. More constructively $\langle S \rangle$ can be described as the set of all (finite) linear combinations of elements in S, meaning vectors of the form $\alpha_1 v_1 + \cdots \alpha_k v_k$ where the α_i's are scalars and the v_i's are vectors in S. We say that V is finitely generated if $V = \langle S \rangle$ for some *finite* subset S of V.

We say that an ordered list $(v_1, \ldots, v_n)$ of finitely many vectors in V, possibly with repetitions,[†]

① spans V ② is (linearly) independent in V ③ is a basis for V

according to whether each vector in V can be expressed as a linear combination of the vectors in the list in ① *at least* one way, *at most* one way, or ③ *exactly* one way, respectively. Thus to say that the list *spans* V means that for each $v \in V$, there must exist scalars α_i such that $v = \sum \alpha_i v_i$. To say that the list is *independent* means that *if* such scalars exist for any given v, then they are unique; this is easily seen to be equivalent to the statement that the assumption $\alpha_1 v_1 + \cdots \alpha_n v_n = 0$ forces all the α_i's to be zero, which is how one usually proves that a given list of vectors is independent. To say that the list is a *basis* means that it *spans* V *and is independent*, that is, each $v \in V$ is *uniquely* expressible in the form $v = \sum \alpha_i v_i$.

Remarks ① These notions generalize to *infinite* lists of vectors, still only allowing *finite* linear combinations (or equivalently infinite linear combinations with all but finitely many coefficients equal to zero). We will show below that every vector space V has a (possibly infinite) basis, that all bases for V have the same (possibly infinite) size, which we call the dimension of V.

[†] The importance of the ordering will become clear when we discuss coordinates below.

② Why are bases important? By definition, they provide unique expressions for the vectors in V, thus giving these vectors "coordinates" (see below), but they can also be used to describe linear maps $T : V \to W$. Indeed T is *uniquely determined by its values on a basis* $(v_1, v_2, \dots)$ *for* V since

$$T(\sum \alpha_i v_i) = \sum \alpha_i T(v_i)$$

by linearity. Furthermore, this equality can be used to *define* T if it is a priori only defined on the basis. In other words, for any list $(w_1, w_2, \dots)$ of vectors in W, there exists a unique linear map $T : V \to W$ for which $T(v_i) = w_i$, given by $T(\sum \alpha_i v_i) = \sum \alpha_i w_i$. This bears repeating:

$$\boxed{\text{Linear maps are uniquely determined by their values on a basis.}}$$

More on this below.

Our next order of business is to prove that bases exist. But before doing so, here is a homework problem to illustrate how one checks that a list of vectors forms a basis. Let V be a vector space that has a finite basis $B = (v_1, \dots, v_n)$. Then there is an associated dual basis $B^* = (v^1, \dots, v^n)$ (note the superscripts) for its dual space V^* characterized by

$$v^i(v_j) \;=\; \delta^i_j \;:=\; \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \quad .$$

$\boxed{\text{HW\#23}}$ Prove that the functionals v^i do indeed form a basis for V^*, and that any $f \in V^*$ can be expressed uniquely in this basis as $f = \sum_i f(v_i) v^i$.

Now for some theory. First observe that bases can be characterized as either "minimal" spanning or "maximal" independent lists, and that this will allow us to establish their existence :

Lemma 2.1 *Let V be a vector space.*

ⓐ *Any minimal spanning list*[1] *or maximal independent list*[2] *of vectors in V forms a basis for V.*

ⓑ *If V is finitely generated, then any finitespanning list S of vectors in V can be shrunk to a basis, and any finite independent list I can be extended to a basis* (necessarily finite by 2.3 below). *In particular, any finitely generated vector space has a finite basis.*

Proof ⓐ Let $(v_1, v_2, \dots)$ be a *minimal* spanning list of vectors in V, and suppose some (finite) linear combination $\sum \alpha_i v_i = 0$. If some $\alpha_j \neq 0$, then v_j can be written in terms of the other v_i's, indeed $v_j = \sum_{i \neq j}(-\alpha_j^{-1}\alpha_i)v_i$, so dropping v_j from the list leaves a *smaller* spanning list $\Longrightarrow\Longleftarrow$. Thus all the α_i's $= 0$, so $(v_1, v_2, \dots)$ is independent, therefore a basis.

If $(v_1, v_2, \dots)$ is a maximal independent list, then prepending any $v \in V$ to this list gives a dependent list $(v, v_1, v_2, \dots)$. Thus $\exists$ scalars $\alpha, \alpha_1, \dots,$ not all zero, with $\alpha v + \alpha_1 v_1 + \cdots = 0$. In particular $\alpha \neq 0$ since $(v_1, v_2, \dots)$ is independent, so $v = \sum(-\alpha^{-1}\alpha_i)v_i$ lies in the span of $(v_1, v_2, \dots)$. Thus $(v_1, v_2, \dots)$ is both independent and spans V, so is a basis.

ⓑ For the first assertion, shrink S to a minimal spanning list, and apply ⓐ. For the second, choose any finite basis B for V, and let J be a maximal independent extension of I inside the concatenated list IB. Then for any vector v in B that is not in J, the list Jv is dependent, so v is in the span of J. Thus J spans V, so is a (finite) basis extending I. □

i.e. no sublist spans

[2] i.e. no superlist is independent

Remark The conclusion of 2.1b with “finite” omitted holds for arbitrary vector spaces; this requires delicate work using Zorn's Lemma – a form of “transfinite” induction – which we omit. Thus in fact every vector space has a basis (proved above for finitely generated vector spaces).

Now that we know that bases always exist, we would like to show that they all have the same size. There are various ways to do this; we choose one that focuses on an interplay between the notions above and linear maps between vector spaces. In particular, spanning and independent lists can be used to characterize when such maps are onto or one-to-one, as follows :

Lemma 2.2 *A linear map $T : V \to W$ of vector spaces is onto if and only if it carries spanning lists in V to spanning lists in W, and is one-to-one if and only if it carries independent lists in V to independent lists in W. Thus T is an isomorphism if and only if it carries bases to bases. Furthermore, if $T : V \to V$ with V is finitely generated* (this last hypothesis is necessary) *then T is onto if and only if it is one-to-one.*

Proof Let $B = (v_1, v_2, \dots)$ be any list of vectors in V, with image $TB = (Tv_1, Tv_2, \dots)$ in W. Then T onto $\iff \forall\, w \in W$, $\exists$ scalars α_i such that $T(\sum \alpha_i v_i) = w \iff \forall\, w \in W$, $\exists$ scalars α_i such that $\sum \alpha_i T(v_i) = w$ (since T is linear) $\iff TB$ spans V. Similarly T one-to-one $\iff \ker(T) = \{0\}$ $\iff [T(\sum \alpha_i v_i) = 0 \implies \sum \alpha_i v_i = 0] \iff [\sum \alpha_i T(v_i) = 0 \implies$ all $\alpha_i = 0] \iff TB$ is independent.

To prove last statement, recall that V has *at least* one finite basis, by 2.1b. Let B be a smallest such, with say n vectors. If T is onto, then TB spans V, so is in fact a basis (otherwise it could be shrunk to one with $< n$ vectors). Thus TB is independent, so T is one-to-one. Conversely, if T is one-to-one then TB is independent. If TB were in fact a basis, then it would span W, so T would be onto and we'd be done. So we must show TB is indeed a basis. But if its not, then it could be extended to a basis C with $> n$ vectors by 2.1b, and one could then construct an linear map $V \to V$ that was onto but not one-to-one (sending the first n vectors in C to their corresponding vectors in B, and the rest of the vectors in C to zero). Thus TB *is* a basis, and we're done. $\square$

The next result is arguably the most important theoretical result in linear algebra :

Theorem 2.3 *All bases for a vector space V have the same size, called the dimension of V and denoted* $\dim(V)$*. In particular V cannot have both a finite and infinite basis.*

Proof We first prove the last statement. If V had a finite basis B, say with n vectors, and an infinite basis C, then as in the previous proof, we could construct a linear map $V \to V$ that is onto but not one-to-one (sending the first n vectors in C to their corresponding vectors in B, and the rest of the vectors in C to zero). But this contradicts Lemma 2.2. In fact this same argument for finite bases B and C with $|B| < |C|$ proves the first statement for finitely generated vector spaces; the nonfinitely generated case again requires Zorn's Lemma. $\square$

Corollary 2.4 *Let V be a finite dimensional vector space.* ⓐ $U \lneq V \implies \dim(U) < \dim(V)$.

ⓑ *Any spanning list or independent list of exactly* $\dim(V)$ *vectors in V is a basis for V.*

Proof By 2.1b, any basis B for U extends it to a (larger since $U \neq V$) basis C for V. Then $\dim(U) = |B| < |C| = \dim(V)$, proving ⓐ. ⓑ is immediate from Lemma 2.1b. $\square$

Next we prove one of the most quoted theorems in elementary linear algebra. It concerns the nullity and rank of a linear transformation $T : V \to W$, defined by

$$\mathrm{null}(T) := \dim(\ker(T)) \qquad \text{and} \qquad \mathrm{rk}(T) := \dim(\mathrm{Im}(T))\,.$$

Theorem 2.5 (rank+nullity theorem) *If $T : V \to W$ is linear, then* $\dim(V) = \mathrm{rk}(T) + \mathrm{null}(T)$.

Proof Choose a basis $B = (v_1, \ldots, v_n)$ for V that extends one $(v_1, \ldots, v_k)$ for $\ker(T)$ (so $k \leq n$). Then we claim $C = (T(v_{k+1}), \ldots, T(v_n))$ is a basis for $\mathrm{Im}(T)$; this would prove the theorem, since $n = (n-k) + k$. To see this we must show two things: ① C is independent and ② C spans W. For ① we have $\sum_{i>k} \alpha_i T(v_i) = 0 \implies T(\sum_{i>k} \alpha_i v_i) = 0 \implies \sum_{i>k} \alpha_i v_i = \sum_{i \leq k} \alpha_i v_i$ (for some α_i for $i \leq k$) $\implies$ all $\alpha_i = 0$, since B is independent. For ②, any $w \in \mathrm{Im}(T)$ is $T(v)$ for some $v = \sum \alpha_i v_i \in V$ (since B spans V). Thus $w = T(\sum \alpha_i v_i) = \sum \alpha_i T(v_i) = \sum_{i>k} \alpha_i T(v_i)$ since $T(v_i) = 0$ for $i \leq k$. This completes the proof. □

Corollaries *Let U and V be finite dimensional F-vector spaces.*

① *If $U < V$, then* $\dim(V/U) = \dim(V) - \dim(U)$.

② *In general we have* $\dim(U \oplus V) = \dim(U) + \dim(V)$.

③ *If $U, V < W$, then* $\dim(U + V) = \dim(U) + \dim(V) - \dim(U \cap V)$.

④ *Any n homogeneous linear equations/F in $k > n$ variables has nontrivial solutions.*

Proof For ①, apply the theorem to the natural projection $V \to V/U$. ② and ③ are homework. For ④, rewrite any system of linear equations

$$\begin{gathered} a_{11}x_1 + \cdots + a_{1k}x_k = 0 \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nk}x_k = 0 \end{gathered}$$

as a single matrix equation $Ax = 0$, where $A = (a_{ij})$, $x = (x_1, \ldots, x_k)^T \in F^k$ and $0 = (0, \ldots, 0)^T \in F^n$. Now view the solutions to the system as the kernel of the linear map $A : F^k \to F^n$ sending x to Ax. We must show $\mathrm{null}(A) > 0$. But $\mathrm{rk}(A) \leq n$ (by the corollary to theorem 2.1) and so $\mathrm{null}(A) = k - \mathrm{rk}(A) \geq k - n > 0$. □

$\boxed{\text{HW\#24}}$ Prove ② and ③

As another application we revisit the dual space V^* of a finite dimensional vector space V. Recall from HW #23 that if $B = (v_1, \ldots, v_n)$ is any basis for V, then there is an associated dual basis $B^* = (v^1, \ldots, v^n)$ for V^*, characterized by $v^i(v_j) = \delta^i_j$. It follows that (for finite dimensional spaces) $V \cong V^*$ (via the map $v_i \mapsto v^i$) and in particular

$$\dim(V) = \dim(V^*).$$

Now for any subspace U of V, define the annihilator U° of U to be the subspace of V^* consisting of all functionals which vanish on U. You should verify that this is a subspace.

$\boxed{\text{HW\#25}}$ Prove that $\dim(V) = \dim(U) + \dim(U^\circ)$. (Hint: Extend a basis A for U to a basis B for V, and use this to define an linear map $V \to V^*$ that sends the vectors in A to their duals in B^*, and the vectors in $B - A$ to zero.)

$\boxed{\text{HW\#26}}$ Let $T : V \to W$ be a linear map between finite dimensional vector spaces. Prove that $\ker(T^*) = (\mathrm{Im}T)^\circ$ and deduce that $\mathrm{rk}(T) = \mathrm{rk}(T^*)$ (this is in disguise the "row rank = column rank" theorem from linear algebra (see HW#28 below).

§3. A brief introduction to homological algebra

The rank+nullity theorem (Theorem 2.2 above) can be recast in a more abstract setting, as follows: Let

$$0 \longrightarrow U \xrightarrow{S} V \xrightarrow{T} W \longrightarrow 0$$

be a short exact sequence, meaning a sequence of linear transformations in which S is 1-1, T is onto, and $\mathrm{Im}(S) = \ker(T)$. Then

$$\dim(V) = \dim(U) + \dim(W).$$

To prove this result, apply the rank+nullity theorem to T, noting that $U \cong \mathrm{Im}(S) = \ker(T)$ and $W = \mathrm{Im}(T)$. Alternatively, one can construct an isomorphism $V \cong U \oplus W$ and then deduce the result from HW23 (exercise).

Conversely, Theorem 2.2 follows from this result since $T : V \to W$ gives rise to a short exact sequence $0 \to \ker(T) \hookrightarrow V \to \mathrm{Im}(T) \to 0$ where the map $V \to \mathrm{Im}(T)$ "is" T (i.e. sends x to $T(x)$).

More generally consider any sequence of linear transformations

$$\cdots \longrightarrow V_{k-1} \xrightarrow{T_{k-1}} V_k \xrightarrow{T_k} V_{k+1} \longrightarrow \cdots .$$

This is called a chain complex if $T_k \circ T_{k-1} = 0$ (i.e. $\mathrm{Im}(T_{k-1}) \subset \ker(T_k)$) for all k. If for some k we have $\ker(T_k) = \mathrm{Im}(T_{k-1}$, then the sequence is said to be exact at V_k, and it is an exact sequence if it is exact at each V_k.

HW#27 Show that a sequence

$$\cdots \to 0 \to U \xrightarrow{S} V \to \cdots$$

is exact at U if and only if S is 1-1, and

$$\cdots \to V \xrightarrow{T} W \to 0 \to \cdots$$

is exact at W if and only if T is onto (so if $\cdots \to 0 \to V \xrightarrow{T} W \to 0 \to \cdots$ is exact, then T is an isomorphism). This explains the terminology "short exact sequence" above.

Digression Chain complexes in simplicial and smooth topology

An important integer invariant of any finite chain complex $\mathcal{V}\colon 0 \to V_1 \to V_2 \to \cdots \to V_n \to 0$ is its Euler characteristic

$$\chi(\mathcal{V}) = \sum_{k=1}^{n} (-1)^k \dim(V_k).$$

It is an easy consequence of the rank+nullity theorem that the Euler characteristic of any exact chain complex is zero.

Remark There is a subtler measure of inexactness associated with any chain complex $\mathcal{V}$, namely the quotient spaces

$$H_k(\mathcal{V}) := \ker(T_k)/\mathrm{Im}(T_{k-1}).$$

These are generally called the homology groups of the complex (where the word "group" refers to the additive structure).

§4. Coordinates

Coordinate vectors

Let V be a finite dimensional vector space. For any chosen basis $B = (b_1, \dots, b_n)$, each vector $v \in V$ can be written uniquely as $v = \sum_{i=1}^{n} \alpha_i b_i$ (with $\alpha_i \in F$, the field of scalars) so record this as a column vector

$$v_B = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

called the coordinate vector for v with respect to B. The map $v \to v_B$ is then an isomorphism $V \to F^n$. Note that the i^{th} coordinate of v_B is $b^i(v)$, where $B* = b^1, \dots, b^n$ is the basis of V dual to B, i.e. $v_B^T = (b^1(v), \dots, b^n(v))$.

Coordinate matrices

Let $T : V \to W$ be linear. For any bases $B = (b_1, \dots, b_n)$ and $C = (c_1, \dots, c_k)$ for V and W, respectively, the coordinate matrix for T with respect to B and C is the $k \times n$ matrix whose j^{th} column is $(Tb_j)_C$ (writing Tv for $T(v)$)

$$T_{CB} = ((Tb_1)_C, \dots, (Tb_n)_C),$$

or equivalently, using the dual basis $C^* = (c^1, \dots, c^k)$, whose ij^{th} entry is $c^i(Tb_j)$. This matrix represents T in the following sense

$$T_{CB}\, v_B = (Tv)_C.$$

(Check on a basis.) The map $T \mapsto T_{CB}$ is an isomorphism $\text{Hom}(V, W) \to M_{k \times n}(F)$. When $V = W$ and $B = C$, we write T_B for T_{BB}, the coordinate matrix for $T : V \to V$ w.r.t. the basis B.

Composition $U \xrightarrow{S} V \xrightarrow{T} W$ becomes matrix multiplication

$$(T \circ S)_{CA} = T_{CB} S_{BA}$$

w.r.t. bases A, B, C for U, V, W.

Change of basis

The coordinate matrix for the identity map $I : V \to V,\ I(v) = v$ w.r.t. two bases B, B' satisfies

$$I_{B'B} v_B = v_{B'}.$$

Thus multiplication by $I_{B'B}$ effects the change of basis for coordinate vectors from B to B' in V. Clearly $I_{B'B}^{-1} = I_{BB'}$.

Now if $T : V \to W$ is linear, and C, C' are bases for W, then the change of basis formula for coordinate matrices is

$$T_{C'B'} = I_{C'C} T_{CB} I_{BB'}.$$

When $V = W$, $B = C$ and $B' = C'$, that is when T is an endomorphism of a vector space V with basis B, this becomes

$$T_{B'} = P T_B P^{-1},$$

where $P = I_{B'B}$. In this case we say T_B and $T_{B'}$ are similar, or to emphasize that P should have entries in the field F of scalars, that they are similar over F.

Remark In general, two square matrices A, B are similar, written $A \sim B$, if $\exists$ invertible matrix P such that

$$A = PBP^{-1}.$$

The discussion above shows that if A is a coordinate matrix for an endomorphism T of a finite dimensional vector space (with respect to some basis), then the similarity class of A is the exactly the set of all possible coordinate matrices for T.

One of the central problems in linear algebra is to find the "simplest" possible matrix similar to a given one, or equivalently, to find bases w.r.t. which a given transformation is in the simplest possible form. More on this below.

$\boxed{\text{HW\#28}}$ Show that if $T : V \to W$ is a linear transformation between finite dimensional vector spaces, then for any bases B and C for V and W, with dual bases B^* and C^* for V^* and W^*,

$$T^t_{CB} = T^*_{B^*C^*}.$$

That is, the coordinate matrix of the dual of a transformation is the *transpose* of its coordinate matrix, w.r.t. dual bases. Then deduce using HW#26 that the "row rank" (= dimension of the row space) and "column rank" (= dimension of the column space) of any matrix are equal.

III Modules

§1. Basics

The notion of an R-module is a generalization of the notion of a vector space in which the field of scalars is replaced by any ring R with 1.

Definition A left R-module is an abelian group $(M, +)$ with scalar multiplication $R \times M \to M$ satisfying

① $(r+s)m = rm + sm$ and $r(m+n) = rm + rn$

② $(rs)m = r(sm)$

③ $1m = m$

for all $r, s \in R$ and $m, n \in M$.[†] Similarly define right R-module.

An R-module homomorphism (or R-linear map) is a map $f\colon M \to N$ of R-modules satisfying $f(rm+n) = rf(m) + f(n)$. As usual it follows that $f(0) = 0$. The kernel of f is the subset $\ker(f) = f^{-1}(0)$ of M, which equals $\{0\} \iff f$ is 1-1 (exercise).

Examples ① $R =$ field F. R-modules$=$ F-vector spaces, R-linear maps $=$ F-linear maps, submodules $=$ subspaces. So this is just vector space theory.

② $R = \mathbb{Z}$. $\mathbb{Z}$-modules $=$ abelian groups (scalar mult $=$ repeated $+$), $\mathbb{Z}$-linear maps $=$ group homomorphisms, submodules $=$ subgroups. So this is just abelian group theory.

③ $R = F[t]$. Let $T : V \to V$ be an endomorphism of an F-vector space V. Want to understand T. Concoct an $F[t]$-module V_T whose structure tells us all about T. In particular, V_T is additively the same as V, but the scalars are enlarged from F to include all of $F[t]$, acting by $f(t) \cdot v := f(T)v$, where the polynomial $f(T)$ in T represents an element of $\operatorname{End}(V)$. For example if $f(t) = t^2 - 3t + 1$, then $f(T) = T^2 - 3T + I$, which maps any $v \in V$ to $T^2v - 3Tv + Iv = T(T(v)) - 3T(v) + v$.

The set of all linear maps $M \to N$ is denoted $\operatorname{Hom}_R(M, N)$. It is an abelian group ($(f+g)(m) := f(m) + g(m)$), and in fact an R-module ($(rf)(m) := r(f(m))$) if R is commutative. If $M = N$, write $\operatorname{End}_R(M)$ for $\operatorname{Hom}_R(M, M)$.

$\boxed{\text{HW\#29}}$ It is a fact that $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Z}_m, \mathbb{Z}_n) \cong \mathbb{Z}_k$ for some k. Find k in terms of m and n, and prove it. (If you want to get a feel for this, experiment with some small values of m and n.)

Other basic notions

① Submodules K is a submodule of M, written $K < M$, means K is a subset of M that is closed under all the operations: $0 \in K$ and $m, n \in K$, $r \in R \Longrightarrow rm + n \in K$.

Examples: (1) $M, \{0\} < M$; say M is simple if these are its only submodules, (2) kernels and images of homomorphisms, and (3) ideals in R (noting that R itself is an R module)

② Quotient modules $M/K =$ set of cosets $m+K$ of K in M with operations $(m+K)+(n+K) = (m+n)+K$ and $r(m+K) = (rm)+K$. Get the usual universal property for the canonical projection $M \to M/K$, and all the isomorphism theorems that follow from it.

[†] Alternatively think of scalar multiplication as a "ring action" on M, i.e. a ring homomorphism $R \to \operatorname{End}(M)$, where $\operatorname{End}(M)$ is a ring (under $+$ and composition $\circ$) of endomorphisms of the additive group M.

③ Direct sums $M \oplus N = \{(m,n) : m \in M, n \in N\}$, and more generally

$$M_1 \oplus M_2 \oplus \cdots \;=\; \{(m_1, m_2, \dots) : m_i \in M_i \text{ with only finitely many nonzero } m_i\text{'s}\}.$$

Write M^n for $M \oplus \cdots \oplus M$ (n-copies) and M^∞ for $M \oplus M \oplus \cdots$ (countably infinitely many copies).

§2. Special kinds of modules

Free modules

Definition A subset B of a module M is a basis for M if every element in M can be written uniquely as a finite linear combination of elements in B; we then say that M is free on B. A module is called a free module if it has a basis. (The zero module $M = \{0\}$ is considered to be free on $\varnothing$.)

It is not hard to show that an R-module M is free (on some $B \subset M$) iff either of the following conditions is satisfied:

① M is isomorphic to a direct sum of copies of R (e.g. $R, R^2, \dots$)

② (universal mapping property) Any function from B to an R-module N extends uniquely to an R-linear map $M \to N$.

Remarks ① All modules over a field (i.e. vector spaces) are free.

② (without proof) Free modules over any commutative ring R with 1 have a well-defined dimension (= size of basis), also called the rank of the module. This need not be true if either R has no 1 or is not commutative. For example for $R = \mathbb{Z}^\infty$ (the direct sum of a countable number of copies of $\mathbb{Z}$, commutative but with no 1), or $R = \mathrm{End}(\mathbb{Z}^\infty)$ (noncommutative with 1), we have $R \cong R \oplus R$.

Torsion modules

Definition An element m in an R-module M is called a torsion element if $rm = 0$ for some nonzero $r \in M$, and M is called a torsion module if all of its elements are torsion elements. (Thus the zero module is torsion, so both torsion and free).

Examples ① If R is a domain, then torsion R-modules are never free.

② If R is a field, then there are no non-zero torsion R-modules.

③ $\mathbb{Z}_n$ is a torsion $\mathbb{Z}$-module for all $n > 0$, since $\overline{k} \in \mathbb{Z}_n \implies n\overline{k} = \overline{0}$. In fact any finite abelian group A of order n is a torsion module: $na = 0 \ \forall a \in A$. So are some infinite abelian groups, for example $\mathbb{Z}_{n_1} \oplus \mathbb{Z}_{n_1} \oplus \cdots$ for any infinite sequence $n_1, n_2, \dots$ of positive integers.

④ The $F[t]$-module V_T (defined above) is torsion if $\dim(V) = n < \infty$. To see this, one must prove that for all $v \in V$, there is a nonzero "scalar" $f(t) \in F[t]$ such that $f(t)v = f(T)v = 0$. In fact, there is a *single* polynomial $m(t) \in F[t]$ with $m(T)v = 0$ for all $v \in V$. This is because $\mathrm{End}(V) \cong M_n(F)$ is n^2-dimensional, and so the powers $I, T, T^2, \dots, T^{n^2}$ are dependent. Thus $\exists\, \alpha_i \in F$ with $\sum_{i=0}^{n^2} \alpha_i T^i = 0$, so take $f(t) = \sum_{i=0}^{n^2} \alpha_i t^i$.

Remark The set

$$\{f \in F[t] : f(T) = 0\}$$

is an ideal in $F[t]$, and so it has a unique monic generator $m_T(t)$ since $F[t]$ is a PID (where "monic" means the leading coefficient is 1). This is called the minimal polynomial of T, and finding it is one of the most useful computations in linear algebra. The famous Cayley-Hamilton Theorem asserts that the characteristic polynomial

$$c_T(t) = \det(tI - T)\ ^{\dagger}$$

(which is easy to compute) is a multiple of m_T, which often gives a starting point for finding m_T. It is also a fact that every root of c_T is a root of m_T. This will be proved in §3 below, but for now we use it to calculate m_T where T is the endomorphism of $\mathbb{R}^3$ given by multiplying by the 3×3 matrix with rows $(-1, 10, -2)$, $(-1, 6, -1)$ and $(-2, 10, -1)$. (Answer: $t^2 - 3t + 2$).

$\boxed{\text{HW\#30}}$ Compute the characteristic and minimal polynomials of the endomorphism of $\mathbb{R}^3$ given by multiplying by the 3×3 matrix which has 1's in the four corners and 0's elsewhere.

Cyclic Modules

$\boxed{\textit{Henceforth, assume } R \textit{ is a commutative ring with } 1 \neq 0.}$

Definition Let $M = R$-module, $S =$ subset of M. The submodule $\langle S \rangle$ generated by S is the smallest submodule of M containing S, or constructively,

$$\langle S \rangle = \{\textstyle\sum r_i s_i : r_i \in R,\ s_i \in S\}.$$

For example, if M is free on B, then $M = \langle B \rangle$. Say M is finitely generated if $M = \langle S \rangle$ for some finite $S \subset M$, and cyclic if $M = \langle m \rangle$ (i.e. $\langle \{m\} \rangle$) for some $m \in M$.

Proposition *M is a cyclic R-module $\Longleftrightarrow$ $M \cong R/J$ for some $J \lhd R$.*

Proof ($\Longrightarrow$) Given $M = \langle m \rangle$, the linear map

$$R \to M, \quad r \mapsto rm$$

is onto with kernel some ideal $J \lhd R$, so $M \cong R/J$.

($\Longleftarrow$) Given $M \cong R/J$, note that $R/J = \langle 1 + J \rangle$, so M is also cyclic. □

Remark It follows that there is (up to isomorphism) only one nonzero free cyclic R-module, namely to R viewed as a module over itself. All other cyclic R-modules are isomorphic to R/J for some nonzero ideal J in R, so are torsion (annihilated by any nonzero r in J).

$\boxed{\text{HW\#31}}$ (Schur's Lemma) Let M and N be simple R-modules (recall that this means they have no nontrivial proper submodules) and $f : M \to N$ be a nonzero R-linear map. Show that (a) f is an isomorphism, and (b) if $M = N$ and R is commutative, then f is multiplication by a scalar, i.e. $\exists\, r \in R$ such that $f(x) = rx$ for all $x \in M$. (Hint for (b): Consider the cyclic submodule $\langle x_0 \rangle$ for any nonzero $x_0 \in M$.)

[†] Recall from basic linear algebra that this is the polynomial whose roots are the eigenvalues of T. Here the determinant of a transformation is defined to be the determinant of its coordinate matrix w.r.t. *any* basis. (Exercise: this is independent of the choice of basis.)

§3. Modules over a PID

Theorem 3.1 (Structure of Finitely Generated Modules over a principal ideal domain R) *If M is a finitely generated R-module, then $\exists$ a unique nonnegative integer r (called the rank of M) and a unique decreasing sequence of nonzero proper ideals $J_1 \supset J_2 \supset \cdots \supset J_k$ in R (called the invariant factors of M) such that*

$$M \cong R/J_1 \oplus \cdots \oplus R/J_k \oplus R^r.$$

The proof relies on the following technical but very useful result, whose proof will be given at the end of this section:

Lemma 3.2 *If U is a submodule of a free R-module V of rank n, where R is a* PID, *then*

ⓐ *U is free of rank $\ell \leq n$.*

ⓑ *There exists bases $(u_1, \dots, u_\ell)$ for U and $(v_1 \dots, v_n)$ for V, and nonzero scalars $r_1, \dots, r_\ell$ with each r_i dividing r_{i+1}, such that each $v_i = r_i v_i$.*

Uniqueness (not proved here)*: The ideals $\langle r_1 \rangle \supset \cdots \supset \langle r_m \rangle$ are uniquely determined by U and V.*

Proof (of 3.1 from 3.2) Choose generators $x_1, \dots, x_n$ for M. Then there is an epimorphism $R^n \to M$ sending e_i to x_i for $i = 1, \dots, n$, where $(e_1, \dots, e_n)$ is the standard basis for R^n.

Write V for R^n and U the kernel of this map, so $M \cong V/U$. By Lemma 3.2, there exists another basis $B = (v_1, \dots, v_n)$ for V and nonzero scalars $r_1, \dots, r_\ell$ for some $\ell \leq n$ with $r_1 | r_2 | \cdots | r_\ell$ such that $(r_1 v_1, \dots, r_\ell v_\ell)$ is a basis for U. Thus

$$M \;\cong\; V/U \;=\; R^n \,/\, \langle r_1 v_1 \rangle + \cdots + \langle r_\ell v_\ell \rangle \;\cong\; R/J_1 \,\oplus \cdots \oplus\, R/J_\ell \oplus R^r$$

where $J_i = \langle r_i \rangle \lhd R$ and $r = n - \ell$.[†] Now throw out the r_i's that are units (at the beginning) and shift the numbering to get $M \cong R/J_1 \oplus \cdots \oplus R/J_k \oplus R^r$ with $R \supsetneq J_1 \supset J_2 \supset \cdots \supset J_k$, as desired. The uniqueness follows from uniqueness in 3.1. □

Examples ① ($R = \mathbb{Z}$) Every finitely generated abelian group is isomorphic to

$$\mathbb{Z}_{n_1} \oplus \cdots \oplus \mathbb{Z}_{n_k} \;\oplus\; \mathbb{Z}^r$$

for unique integers $n_i > 1$ with $n_1 | \cdots | n_k$ (the invariant factors) and $r \geq 0$ (the free rank).

② ($R = F[t]$, $F =$ field) Recall that $M = V_T$ ($=$ an F-vector space V with scalar multiplication $t \cdot v = T(v)$ for some given $T \in \mathrm{End}(V)$) is a torsion $F[t]$-module. Thus it has rank zero, and so there exist unique nonconstant monic polynomials $f_1 | \cdots | f_k$, each f_i dividing f_{i+1}, such that

$$V_T \cong F[t]/\langle f_1 \rangle \oplus \cdots \oplus F[t]/\langle f_k \rangle. \qquad (*)$$

The f_i's are called the invariant factors of T.

[†] The last isomorphism is induced via the first isomorphism theorem from the map $R^n \to R/J_1 \oplus \cdots \oplus R/J_\ell \oplus R^r$ sending x to $(\alpha_1 + J_1, \dots, \alpha_\ell + J_\ell, \alpha_{\ell+1} \dots, \alpha_n)$ where $\alpha_i = v^i(x)$, the ith coordinate of x_B.

Rational Canonical Form

What does the cyclic module $F[t]/\langle f\rangle$ for a given nonconstant polynomial

$$f(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$$

look like? As a vector space over F, it is finite dimensional with basis $B = (\bar{1}, \bar{t}, \bar{t}^2, \ldots, \bar{t}^{n-1})$. (Indeed, B spans by the division algorithm, and is independent since $\sum_{i=0}^{n-1} \alpha_i \bar{t}^i = \bar{0}$ in the quotient implies $f(t) \mid \sum_{i=0}^{n-1} \alpha_i t^i$, whence all $\alpha_i = 0$ since $\deg(f) = n$.) The coordinate matrix in this basis for the linear map corresponding to multiplication by t, called the companion matrix of f, is

$$C_f = \begin{pmatrix} 0 & 0 & & 0 & 0 & -a_0 \\ 1 & 0 & & 0 & 0 & -a_1 \\ 0 & 1 & & 0 & 0 & -a_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & & 1 & 0 & -a_{n-2} \\ 0 & 0 & & 0 & 1 & -a_{n-1} \end{pmatrix}.$$

Examples $C_{t^2-3t+1} = \begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}$, $C_{t^3-t^2-5t+2} = \begin{pmatrix} 0 & 0 & -2 \\ 1 & 0 & 5 \\ 0 & 1 & 1 \end{pmatrix}$.

Carrying these bases over to V via the isomorphism $(*)$ above shows that $\exists$ basis B for V with $T_B = C_{f_1} \oplus \cdots \oplus C_{f_k}$ where $\oplus$ denotes block sum.

Definition A matrix of the form $C_{f_1} \oplus \cdots \oplus C_{f_k}$ where the f_i are monic polynomials in $F[x]$ with $f_1 | \cdots | f_k$ is said to be in rational canonical form (RCF).

Example Since $t^2 - 3t + 1$ divides $t^3 - t^2 - 5t + 2$ (with quotient $t + 2$)

$$C_{t^2-3t+1} \oplus C_{t^3-t^2-5t+2} = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

is in rational canonical form.

Theorem 3.1 shows that any $T \in \mathrm{End}(V)$ has a coordinate matrix in rational canonical form, and that this matrix is *uniquely determined by* T. It is called the rational canonical form of T, and is denoted R_T. When $V = F^n$, so $\mathrm{End}(V)$ is the space $M_n(F)$ of $n \times n$ matrices over F, this says:

Corollary 3.3 *Any $A \in M_n(F)$ is similar to a unique matrix R_A in rational canonical form.*[†]*. Thus matrices $A, B \in M_n(F)$ are similar over F if and only if $R_a = R_B$.*

In fact R_A is independent of the field F, as long as F contains the entries of A (although not hard to prove, we do not do so here). This has a striking consequence: Any pair of square matrices with entries in a field are similar over that field if and only if they are similar over any larger field.

[†] This similarity is *over F*, meaning that there is an invertible matrix P *with entries in F* such that $PAP^{-1} = R_A$

Exercise Investigate how to compute R_A (see e.g. Dummit-Foote). For $n \leq 3$, the following two facts (assigned as homework below) show that R_A is determined by the minimal polynomial m_A and characteristic polynomials c_A:

ⓜ m_A is equal to the largest (last) invariant factor of A, and

ⓒ c_A is the product of all the invariant factors of A.

$\boxed{\text{HW\#32}}$ Prove ⓜ and ⓒ. (Hint: Show that the minimal and characteristic polynomials of a block sum are the least common multiple and product, respectively, of the corresponding polynomials of the summands, and also show that $m_{C_f} = c_{C_f} = f$ for any polynomial f. Then use 3.3.)

From this, we can deduce the following results stated in §2:

Corollary 3.4 *Fix a square matrix A over a field with minimal polynomial m_A and characteristic polynomial c_A* $= \det(tI - A)$ (whose roots are the eigenvalues of A)*. Then m_A divides c_A* (this is the Cayley-Hamilton theorem) *and c_A divides some power of m_A* (so each irreducible factor of c_A is a factor of m_A, and in particular *every root of c_A is a root of m_A*).

Examples ① Let $A = \begin{pmatrix} 2 & -2 & 14 \\ 0 & 3 & -7 \\ 0 & 0 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 0 & -4 & 85 \\ 1 & 4 & -30 \\ 0 & 0 & 3 \end{pmatrix}$, and $C = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{pmatrix}$. Direct computation shows that $c_A = c_B = c_C = (t-2)^2(t-3)$. It follows by Corollary 3.4 that the only possibilities for the minimal polynomials are $(t-2)(t-3)$ and $(t-2)^2(t-3)$. By evaluating these polynomials at A, B, C we find that

$$m_A = (t-2)(t-3) \qquad \text{and} \qquad m_B = m_C = (t-2)^2(t-3).$$

Thus the invariant factors of A are $t-2$ and $(t-2)(t-3) = t^2 - 5t + 6$, while B and C both have $(t-2)^2(t-3) = t^3 - 7t^2 + 16t - 12$ as their only invariant factor, and so

$$R_A = C_{t-2} \oplus C_{t^2-5t+6} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & -6 \\ 0 & 1 & 5 \end{pmatrix} \qquad R_B = R_C = C_{t^3-7t^2+16t-12} = \begin{pmatrix} 0 & 0 & 12 \\ 1 & 0 & -16 \\ 0 & 1 & 7 \end{pmatrix}.$$

In particular $A \not\sim B \sim C$.

② What are the possible RCF's over $\mathbb{Q}$ (resp. $\mathbb{C}$) for matrices with characteristic polynomial $p(t) = t^5 - t^4 + 2t^3 - 2t^2 + t - 1 = (t-1)(t^2+1)^2$? Note that these matrices must be 5×5 since $\deg(p) = 5$. Well, p factors into irreducibles over $\mathbb{Q}$ as ℓq^2 where $\ell = t - 1$ and $q = t^2 + 1$, and over $\mathbb{C}$ as $\ell \ell_+^2 \ell_-^2$ where $\ell_\pm = t \pm i$ (so $q = \ell_+ \ell_-$). For convenience set $c := \ell q = t^3 - t^2 + t - 1$ and $f_\pm := \ell_\pm c$, so the minimal polynomia/$\mathbb{Q}$ (resp. $\mathbb{C}$) must be either c or p (resp. c, $f_\pm$ or p). Thus the invariant factors/$\mathbb{Q}$ (resp. $\mathbb{C}$) must be either q, c or just p (resp. q, c, $\ell_\pm, f_\mp$ or just p) . Hence the possible RCF's/$\mathbb{Q}$ (resp. $\mathbb{C}$) are $C_q \oplus C_c$ or C_p (resp. $C_{\ell_\pm} \oplus C_{f_\mp}$ or C_p). For practice, write these out as 5×5 matrices.

$\boxed{\text{HW\#33}}$ Find all possible rational canonical forms for matrices/$\mathbb{Q}$, and then for matrices/$\mathbb{C}$, with characteristic polynomial $(t^4 - 1)(t^2 + 1)$.

Jordan Canonical Form

There is another useful canonical form for matrices which arises from the "primary" version of the Structure Theorem 3.1, which is derived from the following:

Chinese Remainder Theorem *If I and J are ideals in a commutative ring R with 1 such that $R = I + J$, then $R/(I \cap J) \cong R/I \oplus R/J$.*

Proof Since $R = I + J$, $1 = i + j$ for some $i \in I$ and $j \in J$. It follows that the natural projection $p\colon R \to R/I \oplus R/J$ sending r to $(r+I, r+J)$ is *surjective* (since any $(s+I, t+J) \in R/I \oplus R/J$ is $p(r)$ where $r = ti + sj$; indeed $r + I = sj + I = s(1-i) + I = s + I$ and $r + J = ti + J = t(1-j) + J = t + J$) with *kernel* $I \cap J$. Now apply the first isomorphism theorem. □

Remark When R is a PID, $I = (r)$ and $J = (s)$ for some $r, s \in R$, and $I + J = R$ means that r and s are coprime, so $I \cap J = (rs)$. The isomorphism $R/(rs) \to R/(r) \oplus R/(s)$ (from the theorem) is onto, so for any $a, b \in R$, there exists an $x \in R$ unique up to multiples of rs for which

$$x \equiv a \pmod r \qquad \text{and} \qquad x \equiv b \pmod s$$

i.e. x leaves remainders of a and b upon division by s and t respectively. Whence the name "remainder" theorem.

Using the Chinese Remainder Theorem, the structure theorem can be rephrased as follows:

Structure Theorem (Primary Form) *If V is a finitely generated module over a principal ideal domain R, then V is isomorphic to a direct sum of finitely many modules of the form R or $R/(p^n)$, where p is prime in R and $n > 0$.* (The p^n's are called the elementary divisors of V, and are unique, as is the number r of copies of R, called the rank. The rank is zero if and only if V is torsion.)

Applying this to the torsion $F[t]$-module V_T (see Ex. ③, p.24 and Ex. ②, p.27) we see that

$$V_T \cong F[t]/\langle p_1^{n_1}\rangle \oplus \cdots \oplus F[t]/\langle p_k^{n_k}\rangle. \tag{$**$}$$

where the p_i are irreducible monic polynomials and the $n_i > 0$.

An important special case arises when the p_i's are all linear, i.e. when c_T factors into linear factors (which is always the case for example when $F = \mathbb{C}$ by the fundamental theorem of algebra)

$$c_T = (t - \lambda_1)^{n_1} \cdots (t - \lambda_k)^{n_k}.$$

The λ_i are the eigenvalues of T. We can then analyze the action of t on each summand $F[t]/((t-\lambda)^n)$ with respect to the basis $(b_0, \ldots, b_{n-1})$, where $b_i = (t - \lambda)^i$. (This is even better suited to the situation at hand than the basis $1, t, \ldots, t^{n-1}$ we used to get the rational canonical form.) Clearly $(t - \lambda)b_i = b_{i+1}$ (where we set $b_n = 0$ for convenience) or equivalently $tb_i = \lambda b_i + b_{i+1}$ and so the corresponding coordinate matrix for t is the $n \times n$ matrix $J_{\lambda,n}$ with λ's on the diagonal, 1's right below the diagonal, and 0's elsewhere. For example

$$J_{2,3} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{pmatrix}.$$

This is called a Jordan block. Any matrix which is a block sum of Jordan blocks is said to be in Jordan canonical form.

It follows from $(**)$ that there exists a basis B for V such that

$$T_B = J_{\lambda_1,n_1} \oplus \cdots \oplus J_{\lambda_k,n_k}.$$

This is called a Jordan canonical form of T, also denoted J_T.

Corollary 3.6 *If A is a square matrix with entries in a field F whose characteristic polynomial factors into linear factors, then A is similar over F to a matrix J_A in Jordan canonical form (unique up to the order of its Jordan blocks). The total number of appearances in J_A of any given eigenvalue λ of A (i.e. root of c_A) is the multiplicity of λ as a root of c_A, and the size of the largest associated Jordan block is the multiplicity of λ as a root of m_A.* (To prove the last statement, note that the minimal and characteristic polynomials of a Jordan block $J_{\lambda,n}$ are both equal to $(t-\lambda)^n$.)

Examples The Jordan canonical forms of matrices A, B, C in the example above (page 29) are

$$J_A = J_{2,1} \oplus J_{2,1} \oplus J_{3,1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \qquad \text{and} \qquad J_B = J_C = J_{2,2} \oplus J_{3,1} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

$\boxed{\text{HW\#34}}$ Find all possible Jordan canonical forms for complex matrices whose characteristic polynomial is $(t^4-1)(t^2+1)$.

$\boxed{\text{HW\#35}}$ Show that any square matrix over any subfield of $\mathbb{C}$ is similar to its transpose. (This is true for matrices over any field, but we need to know more about fields before we can prove it.)

$\boxed{\text{HW\#36}}$ Let F be any field. Show $A \in M_n(F)$ is diagonalizable (meaning similar to a diagonal matrix) if and only if m_A is a product of distinct linear factors.[†]

We conclude this chapter with the proof of Lemma 3.2, from which the Structure Theorem 3.1 for finitely generated modules over a PID followed. Recall the statement: If U is a submodule of a free R-module V of finite rank n, where R is a PID, then (a) U is free of some rank $\ell \leq n$, and (b) there exists a basis $v_1, \ldots, v_n$ for V and nonzero scalars $r_1, \ldots, r_\ell$ (unique up to associates), each dividing the next, such that $r_1v_1, \ldots, r_\ell v_\ell$ is a basis for U.

Proof Assume $U \neq 0$; the case $U = 0$ is trivial. Fix any basis $(e_1 \ldots, e_n)$ for V with dual basis $(e^1, \ldots, e^n)$ for $V^* = \mathrm{Hom}_R(V, R)$ (the theory of dual vector spaces carries over to free modules). This identifies V with R^n in the usual way. Now view $V \subset F^n$, where F denotes the field of fractions of R. This allows us to define the rank of U to be the dimension of the subspace of F^n generated by U, which clearly agrees with the usual notion of rank if U were known to be free. Thus once we show U is free, it will follow that $\ell := \mathrm{rk}(U) \leq n$.

To analyze the structure of U, pick $f \in V^*$ so that the ideal $f(U) \lhd R$ is maximal among all ideals of the form $g(U)$ for $g \in V^*$ (so $f(U) \neq 0$ since $U \neq 0$). Such an f exists since R is a PID, and therefore Noetherian. Since R is a PID, we have $f(U) = (r_1)$ for some nonzero $r_1 \in R$. Choose $u_1 \in U$ such that $r_1 = f(u_1)$.

Remark The maximality of $(r_1) \Longrightarrow r_1 | h(u_1)$ for any $h \in V^*$ for which $h(U) \subset (r_1)$.

[†] Another interesting result along these lines is: two diagonalizable matrices A, B in $M_n(F)$ are simultaneously diagonalizable (meaning $\exists$ a single invertible matrix P such that both PAP^{-1} and PBP^{-1} are diagonal) if and only if $AB = BA$. This is left as an exercise for the ambitious reader!

Claim ① $u_1 = r_1 v_1$ for some $v_1 \in V$ (and so $f(v_1) = 1$ by the linearity of f)

② $V \cong Rv_1 \oplus \ker(f)$ and ③ $U \cong Ru_1 \oplus \ker(f|U)$

Proofs ① It suffices to show that r_1 divides each coordinate $e^i(u_1)$ of u_1 in $V = R^n$. If not, then $r_1 \nmid s_1 := e^j(u_1)$ for some j. Then $(r_1) \subsetneq (r_1, s_1)$, which can be written as (d) for some $d = ar_1 + bs_1$. But then setting $h = af + be^j \in V^*$ we have $h(u_1) = ar_1 + bs_1 = d$. Thus $(r_1) \subsetneq (d) \subset h(U)$, so $r_1 | d$ by the Remark above, i.e. $(d) \subset (r_1)$. Thus $(r_1) = (d)$, a contradiction.

② Let $v \in V$. Set $a = f(v)$. Then writing $v = av_1 + (v - av_1)$ shows that $v \in Rv_1 + \ker(f)$. Also if $v \in Rv_1 \cap \ker(f)$ then $v = rv_1$ for some $r \Longrightarrow 0 = f(rv_1) = rf(v_1) = r \cdot 1 = r \Longrightarrow r = 0$, so $v = 0$. Thus $Rv_1 + \ker(f) = V$ and $Rv_1 \cap \ker(f) = 0$, which proves $V \cong Rv_1 \oplus \ker(f)$.

③ Let $u \in U$. Then $f(u) = ar_1$ for some $a \in R$. Writing $u = au_1 + (u - au_1)$ shows that $Ru_1 + \ker(f|U) = U$. Also, if $u \in Ru_1 \cap \ker(f|U)$, then $u = ru_1$ for some $r \Longrightarrow 0 = f(ru_1) = rf(u_1) = r \cdot r_1 \Longrightarrow r = 0$ (since R is a domain and $r_1 \neq 0$) so $v = 0$. Thus $Ru_1 \cap \ker(f|U) = 0$, so $U \cong Ru_1 \oplus \ker(f|U)$.

We now use induction to complete the proof of Lemma 3.2, and thus Theorem 3.1.

ⓐ Induct on $\ell = \text{rk}(U)$. If $\ell = 0$, then $U = 0$ and we are done. For $\ell > 0$ we have $\text{rk}(\ker(f|U)) < \text{rk}(U)$ (since $u_1 \in U$ does not lie in the span of $\ker(f|U)$) and so by the inductive assumption $\ker(f|U)$ is free of rank $< \ell$. It follows by claim ③ that U is free of rank $\leq \ell$.

ⓑ Induct on $n = \text{rk}(V)$. The result is trivial if $n = 0$. For $n > 0$, part (a) implies that $\ker(f)$ is free of rank $n - 1$ (by claim ②). By the inductive assumption $\exists$ basis $v_2, \ldots, v_n$ for $\ker(f)$ and nonzero scalars $r_2, \ldots, r_k$, each dividing the next, such that $r_2 v_2, \ldots, r_k v_k$ is a basis for $\ker(f|U)$. It remains to show $r_1 | r_2$. But consider $g \in V^*$ defined by

$$\begin{aligned} g(v_1) = g(v_2) &= 1 \\ g(v_i) &= 0 \quad \text{for } i > 0. \end{aligned}$$

We have $r_1 = g(r_1 v_1)$, so $(r_1) \subset g(U)$. The maximality of (r_1) implies that $(r_1) = g(U)$, and so $r_1 | g(r_2 v_2) = r_2$, as desired.

For the uniqueness statement, see for example Lang's Algebra Chapter XV.2. □

For the reader's convenience, we reprint here the statement of the key

Lemma 3.2 *If U is a submodule of a free R-module V of rank n, where R is a* PID, *then*

ⓐ *U is free of rank $\ell \leq n$.*

ⓑ *There exists bases $(u_1, \ldots, u_\ell)$ for U and $(v_1 \ldots, v_n)$ for V, and nonzero scalars $r_1, \ldots, r_\ell$ with each r_i dividing r_{i+1}, such that each $v_i = r_i v_i$.*

Uniqueness (not proved here)*: The ideals $\langle r_1 \rangle \supset \cdots \supset \langle r_m \rangle$ are uniquely determined by U and V.*

IV Field Theory

§1. Basic Notions

Recall that a field F is a commutative ring with identity 1 $(\neq 0)$ for which each nonzero element has a multiplicative inverse. We say that F is algebraically closed if every polynomial in $F[x]$ has a root in F; for example $\mathbb{C}$ is algebraically closed,[†] but $\mathbb{Q}$ and $\mathbb{Z}_p$ are not (exercise: show the latter). The most basic invariant of a field is its "characteristic":

Definition The characteristic $\mathrm{ch}(F)$ of a field F is the *additive order* of $1 \in F$ if that order is finite, and *zero* otherwise (i.e. when 1 has infinite order). For example $\mathrm{ch}(F) = 0$ for any subfield F of $\mathbb{C}$, and for p prime, $\mathrm{ch}(\mathbb{Z}_p) = \mathrm{ch}(\mathbb{Z}_p[x]/(f)) = p$ for any irreducible polynomial $f \in \mathbb{Z}_p[x]$.

Lemma 1.0 $\mathrm{ch}(F)$ *is either* 0 *or a prime number.*

Proof If $\mathrm{ch}(F) = p \neq 0$ with $p = ab$, then $0 = p \cdot 1 = (a \cdot 1)(b \cdot 1)$ (by the distributive property) so either $a \cdot 1 = 0$ or $b \cdot 1 = 0$. The minimality of $p \implies$ either $a = p$ or $b = p$, so p is prime. □

The subfield P of F generated by 1 is called the prime subfield of F. It is contained in any other subfield of F, and is uniquely determined up to isomorphism by the characteristic of F, indeed isomorphic to $\mathbb{Z}_p$ when $\mathrm{ch}(F) = p \neq 0$ and to $\mathbb{Q}$ when $\mathrm{ch}(F) = 0$ (exercise).

Extension Fields

If F is a subfield of a larger field E, then we call E an extension field of F, denoted E/F (not to be confused with quotients objects) or by writing E right above F with a vertical line between them. The extension is proper means $E \supsetneq F$. The usual perspective in field theory is to analyze the structure of a field by studying its extension fields, in contrast to group theory where one analyzes the structure of a group by studying its subgroups.

Any extension of F inside E (including F or E itself) is called an intermediate field of the extension E/F. These play a central role in "Galois Theory".

An important positive integer associated with an extension E/F is its degree, denoted $\deg(E/F)$ (or sometimes $|E : F|$ or $|E/F|$). This is the dimension of E when viewed as a vector space over F:

$$\deg(E/F) := \dim_F(E).$$

We say that E/F is a finite extension if $\deg(E/F)$ is finite; our focus here will be on finite extensions inside the complex numbers $\mathbb{C}$. The following general result is fundamental and easy to prove:

Theorem 1.1 (Multiplicativity of Degree) *If K is an intermediate field of an extension E/F with E/F and F/K finite, then E/F is finite with* $\deg(E/F) = \deg(E/K)\deg(K/F)$.

Proof Let $e_1, \ldots, e_p$ and $k_1, \ldots, k_q$ be bases for E/K and K/F, respectively. Then the set of all products $e_i k_j$ (for $1 \leq i \leq p$ and $1 \leq j \leq q$) is a basis for E/F (straightforward verification). □

Algebraic and Transcendental Elements

Let E/F be a field extension. An element $\alpha \in E$ is said to be algebraic over F if it is the root of a nonzero polynomial $f \in F[x]$, and is otherwise said to be transcendental over F. Note that in the former case, α is also algebraic over any larger subfield K of E, since $F[x] \subset K[x]$.

[†] This is the *Fundamental Theorem of Algebra*, which surprisingly is most easily proved using *Analysis* and *Topology* (although see https://kconrad.math.uconn.edu/blurbs/fundthmalg/fundthmalglinear.pdf). We do not prove it here.

Examples ① Any $\alpha \in F$ is algebraic over F. Indeed it is the root of $x - \alpha \in F[x]$.
② The complex numbers $\sqrt{2}$, $\sqrt[3]{2}$ and i are all algebraic over $\mathbb{Q}$ (roots of $x^2 - 2$, $x^3 - 2$ and $x^2 + 1$ respectively), while e and π are transcendental over $\mathbb{Q}$.[†]

Notation For any $\alpha \in E$, write $F[\alpha]$ and $F(\alpha)$ for the smallest *subring* and *subfield* (respectively) of E containing F and α, obtained by *adjoining* α to F. Then $F[\alpha] = \{f(\alpha) \mid f \in F[x]\}$ is a subset $F(\alpha) = \{f(\alpha)/g(\alpha) \mid f, g \in F[x],\ g(\alpha) \neq 0\} = \{r(\alpha) : r \in F(x)\}$. In fact the inclusion $F[\alpha] \subset F(\alpha)$ is an equality if and only if α is algebraic :

Theorem 1.2 *Let α be an element in an extension field E of a field F.*

ⓐ *If α is algebraic over F, then it is a root of a unique monic irreducible polynomial $m_{\alpha/F} \in F[x]$ called the minimal polynomial of α over F. This polynomial divides any other polynomial in $F[x]$ that has α as a root. Furthermore $F(\alpha) = F[\alpha] \cong F[x]/\langle m_{\alpha/F}\rangle$. We define the degree of α over F, denoted* $\deg(\alpha/F)$*, to be the degree of its minimal polynomial over F.*

ⓑ *If α is transcendental over F then $F(\alpha)$ properly contains $F(\alpha)$. In fact $F(\alpha)$ is isomorphic to the field $F(x)$ of rational functions in one variable x, identifying α with x, and thus identifying $F[\alpha]$ with the polynomial ring $F[x]$. In this case we say that α has infinite degree over F.*

Proof Consider the "evaluation" ring homomorphism $e_\alpha : F[x] \to E$ mapping f to $f(\alpha)$.

If α is algebraic, then $\ker(e_\alpha) = \{f : f(\alpha) = 0\}$ is a nonzero ideal in the $F[x]$, so equal to $\langle m_\alpha \rangle$ for a unique monic polynomial m_α, since $F[x]$ is a PID. Furthermore, m_α is irreducible, since $m_\alpha = fg \implies m_\alpha(\alpha) = 0 = f(\alpha)g(\alpha) \implies f(\alpha) = 0$ or $g(\alpha) = 0 \implies f$ or g is an associate of m_α (by ② in Theorem 1.1a). Now recall that in a PID, irreducible = prime (for elements) and prime = maximal (for ideals). Thus (m_α) is maximal, so by the first isomorphism theorem $\mathrm{Im}(e_\alpha) = F[\alpha] \cong F[x]/\ker(e_\alpha) = F[x]/\langle m_{\alpha/F}\rangle$ is a field. Therefore $F[\alpha] = F(\alpha)$.

If α is transcendental, then $\ker(e_\alpha) = 0 \implies F[\alpha] \cong F[x]$ (by the first isomorphism theorem, which in this case carries α to x) and so $F(\alpha) \cong F(x)$, the field of fractions of $F[x]$. □

Remarks ① That $F[\alpha]$ is a field when α is algebraic (since it $= F(\alpha)$ in that case) is surprising. For example this implies that $\sqrt[3]{2}$ is invertible in the ring $\mathbb{Q}[\sqrt[3]{2}]$; see if you can find the inverse.

② The degree of the minimal polynomial of $\alpha \in E$ depends in general on the field $F \subset E$. For example $\sqrt{2} \in \mathbb{R}$ has degree 2 over $\mathbb{Q}$ (with minimal polynomial $m_{\sqrt{2}/\mathbb{Q}} = x^2 - 2$) and degree 1 over $\mathbb{R}$ (with $m_{\sqrt{2}/\mathbb{R}} = x - \sqrt{2}$). For another interesting example, let a and b be any pair of distinct cube roots of 2 (there are three of them). Then a has degree 3 over $\mathbb{Q}$ ($m_{a/\mathbb{Q}} = x^3 - 2$) and degree 2 over $\mathbb{Q}(b)$ (with $m_{a/\mathbb{Q}(b)} = x^2 + bx + b^2$); you should convince yourself that $a \notin \mathbb{Q}(b)$.

Corollary 1.3 *Let α be an element in an extension field E of a field F.*

ⓐ *If α is algebraic over F, then $F(\alpha)$ is a finite extension of F with basis $1, \alpha, \alpha^2, \cdots, \alpha^{n-1}$, where $n = \deg(\alpha/F)$. Thus* $\deg(F(\alpha)/F) = \deg(m_{\alpha/F}) = \deg(\alpha/F) = n$.

ⓑ *If E/F is finite, then* $\deg(\alpha/F)$ *divides* $\deg(E/F)$.

Proof ⓐ: By Theorem 1.2a, $F(\alpha) = F[\alpha] \cong F[\alpha]/(m_{\alpha/F})$, which has basis $1, \alpha, \cdots, \alpha^{n-1}$over F. For ⓑ, multiplicativity (Theorem 1.1) gives $\deg(E/F) = \deg(E/F(\alpha))\deg(F(\alpha)/F)$, and the last factor equals $\deg(\alpha/F)$ by part ⓐ. □

[†] For e this follows from elementary calculus (see e.g. Appendix 16 in Simmons' *Calculus with Analytic Geometry*) but for π the proof is much more difficult. Surprisingly, it is not known whether $e + \pi$ or $e\pi$ are algebraic or transcendental, or even whether either is irrational !

HW#37 Show that if $\alpha_1, \ldots, \alpha_k$ all lie in an extension field of F and are algebraic over F, then $d := \deg(F(\alpha_1, \ldots, \alpha_k)/F) \ \leq \ p := \prod_{i=1}^{k} \deg(\alpha_i/F)$, but d need not divide p.

Simple Extensions

Definition A proper field extension E/F is simple if E is obtained from F by adjoining a *single* element, i.e. $E = F(\alpha)$ for some $\alpha \in E - F$. We call α a primitive element for the extension.

Two types of simple extensions $F(\alpha)/F$ are of special interest: quadratic extensions, when $\alpha^2 \in F$, and cyclotomic extensions, when $\alpha^n = 1$ for some $n > 1$. That is, a quadratic extension of F is the result of adjoining the square root of some element $a \in F - F^2$ (where $F^2 := \{x^2 \mid x \in F\}$), written $F(\sqrt{a})$, while a cyclotomic extension is the result $F(u)$ of adjoining a root of unity $u \notin F$.

Example The fields $\mathbb{Q}(\sqrt{7})$, $\mathbb{Q}(e^{2\pi i/7})$), $\mathbb{Q}(i)$, $\mathbb{Q}(\sqrt{2} + \sqrt{3})$ are all simple extensions of $\mathbb{Q}$. The first is quadratic, the second is cyclotomic, the third is both quadratic and cyclotomic (the only such), and the fourth is neither (by Lemma 1.4a below and multiplicativity of degree).

Quadratic extensions, and more generally iterated quadratic extensions (the result of a sequence of extensions quadratic extensions $F = F_0 \subset F_1 \subset \cdots \subset F_n = E$) will feature prominently in the next section. Thus E/F is iterated quadratic means $E = F(\sqrt{a_1}, \ldots, \sqrt{a_n})$ where $a_1 \in F - F^2$, $a_2 \in F(\sqrt{a_1}) - F(\sqrt{a_1})^2$, and so forth.

Lemma 1.4 ⓐ *E/F is quadratic* $\Longleftrightarrow \deg(E/F) = 2$.
ⓑ *E/F is iterated quadratic* $\Longrightarrow \deg(E/F) = 2^n$ (but not conversely).

Proof ⓐ$\Longrightarrow$: If $E = F(\sqrt{a})$ where $a \in F - F^2$, then $m_{\sqrt{a}/F} = x^2 - a$, so $\deg(E/F) = 2$ by Corollary 1.3a. ⓐ$\Longleftarrow$: If $\deg(E/F) = 2$ then E has an F-basis of the form $(1, \alpha)$, so $\deg(\alpha/F) = 2$ by 1.3b. Thus $m_{\alpha/F} = x^2 + bx + c$ for some $b, c \in F$, so $E = F(\sqrt{b^2 - 4c})$. Part ⓑ$\Longrightarrow$ follows from ⓐ$\Longrightarrow$ and the multiplicativity of degree; we omit the proof ⓑ$\Longleftarrow$ fails , which is a bit harder. □

Algebraic Extensions

Definition A field extension E/F is algebraic if *every* α in E is algebraic over F, and otherwise (i.e. when E has *at least* one element transcendental over F) the extension is transcendental.

Theorem 1.5 *Every finite extension E/F is algebraic* (but not conversely[†]).

Proof For any $\alpha \in E$, the list $1, \alpha, \alpha^2, \ldots, \alpha^k$ is linearly dependent over F for any $k > \deg(E/F)$, i.e. $a_0 + a_1\alpha + \cdots + a_k\alpha^k = 0$ for suitable $a_i \in F$ not all zero. Thus $f(x) := a_0 + a_1x + \cdots + a_kx^k$ has α as a root, so α is algebraic over F. □

Corollary 1.6 ⓐ *If α is algebraic over F, then $F(\alpha)/F$ is algebraic.*
ⓑ *If α and β are algebraic over F, then $F(\alpha, \beta)/F$ is algebraic.*[†]

Proof For ⓐ, Corollary 1.3 shows $F(\alpha)/F$ is finite, so algebraic by Theorem 1.5. Similarly for ⓑ, it is enough by Theorem 1.4 to show $F(\alpha, \beta)/F$ is finite. But $F(\alpha, \beta) \supset F(\alpha) \supset F$, and each intermediate extension is finite by 1.3 (the first since β is algebraic over F, so certainly over $F(\alpha)$). Thus $F(\alpha, \beta)/F$ is finite by multiplicativity (Theorem 1.1). □

[†] Think about what Cor 1.5b says in simple terms: If α and β are algebraic over F, then so is any algebraic expression in α and β. For example, since $\sqrt{2}$ and $\sqrt{3}$ are both algebraic over $\mathbb{Q}$, so are $\sqrt{2} + \sqrt{3}$, $\sqrt{6}$, $\sqrt{6}(1 - \sqrt{8/3})$, etc. (although finding their minimal polynomials is a nontrivial matter). This $\Longrightarrow$ that the set $\overline{\mathbb{Q}}$ of all complex numbers algebraic over $\mathbb{Q}$ is an algebraic field extension of $\mathbb{Q}$, showing the failure of the converse of Theorem 1.4.

Quadratically Closed Fields

Before moving on, we define two more notions related to quadratic extensions. A field E is said to be quadratically closed (QC) if it has no quadratic extensions, i.e. $E = E^2$, so every element of E has a square root in E. For example $\mathbb{C}$ is QC : $\sqrt{re^{i\theta}} = \pm\sqrt{r}e^{i\theta/2}$. For any field F inside a QC field E, there is a unique *smallest* QC subfield $\sqrt[E]{F}$ of E containing F, namely the intersection of all QC intermediate fields of the extension E/F. This field, called the quadratic closure of F inside E, has the following useful characterization:

Lemma 1.7 *If E/F is an extension with E quadratically closed, then $\sqrt[E]{F}$ is the union of all iterated quadratic extensions of F inside E. In particular, $\sqrt[E]{F}$ is an algebraic over F, and each of its elements has degree a power of 2 over F.*

Proof The stated union is indeed a field (as the union of any two iterated quadratic extensions $F(\sqrt{a_1}, \ldots, \sqrt{a_n})$ and $F(\sqrt{b_1}, \ldots, \sqrt{b_k})$ lies in another one $F(\sqrt{a_1}, \ldots, \sqrt{a_n}, \sqrt{b_1}, \ldots, \sqrt{b_k})$), and it clearly lies inside any QC subfield of E containing F. Thus it equals $\sqrt[E]{F}$ by definition. The last statement follows from Lemma 1.4b and Corollary 1.3b □

Remark It can be proved (without too much difficulty) that the field $\sqrt[E]{F}$ is independent up to isomorphism of the choice of the QC extension field E of F, so can be denoted simply by $\sqrt{F}$. The special case $\sqrt{\mathbb{Q}} \subset \mathbb{C}$ is particularly nice (and relevant to the applications below). For example this field of $\mathbb{C}$ is invariant under conjugation – simply because the conjugate of a square root is the square root of the conjugate – a property not shared by all subfields of $\mathbb{C}$ (e.g. $\mathbb{Q}(\sqrt[3]{2}\,e^{2\pi i/3})$).

§2. Application: Geometric Constructions

In this section we prove the impossibility of performing certain classical constructions, where the only tools allowed are a *straightedge* and *compass*. These results, established in nineteenth century, settled a number of problems dating back thousands of years to the Greeks.

Fix a set S of two or more complex numbers, which we call S-points. Given two S-points a, b we call the line $L(a,b)$ through a and b an S-line, the circle $C(a,b)$ centered at a and through b an S-circle, and any S-line or S-circle an S-curve. Let S' denote the set of all intersection points between pairs of distinct S-curves; the S'-points and S'-curves are said to be "constructible in one step" from S. This gives a nested sequence $S = S_0 \subset S_1 \subset S_2 \subset \cdots$ of subsets of $\mathbb{C}$, where each $S_n = S'_{n-1}$, so consists of all points constructible in n steps from S. Thus the union $\mathbb{S} = \cup_{n=0}^{\infty} S_n$ is the set of all points constructible in a finite number of steps from S. Note that $\mathbb{S}' = \mathbb{S}$.

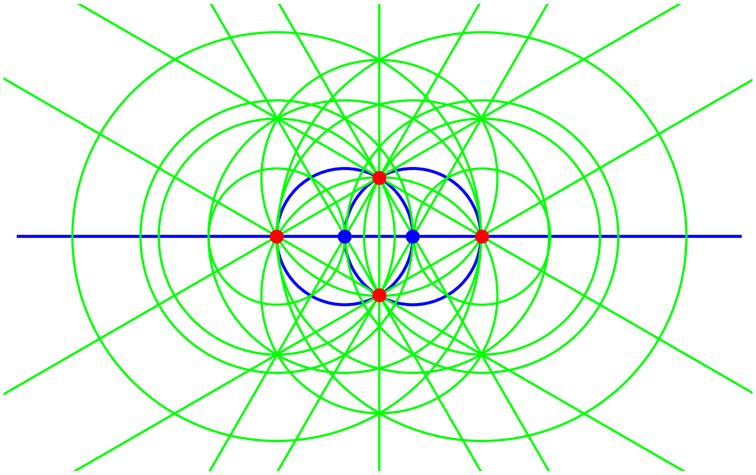
Remark It is assumed that the compass we are using is "collapsible", meaning it collapses when lifted from the page. However, the following argument, which we call "Euclid's trick", shows that this is as good as having a rigid compass. More precisely: *For any circle C centered at a point p, the circle C_q of the same size centered at any other point q* (the transfer of C to q) *can be constructed from $\{p, q\}$ using only a collapsible compass and straightedge.* For if q lies outside C, then for any $s \in C(p,q) \cap C(q,p)$ and $t \in C \cap C(q,p)$, we have $C_q = C(q,r)$ for any $r \in C(p,q) \cap C(s,t)$. If q lies on or inside C, then first transfer C to a point well outside of C (whose construction may require the straightedge and arbitrarily many uses of the collapsible compass) and then back to q.

HW#38 Show that the line $L_p^{\perp}$ perpendicular to any S-line L through any S-point p is an S_2-line, and hence that the line $L_p^{\parallel}$ parallel to L through p is an S_4-line. Also show that the line bisecting the angle formed by two S-lines is also an S_4-line.

Now specialize to the case when S is just the two-element set $K := \{0, 1\}$. Thus the set

$$\mathbb{K} := \cup_{n\geq 0} K_n$$

consists of all points in $\mathbb{C}$ constructible from 0 and 1, a.k.a. the constructible numbers. In the figure below the two blue dots form K_0, the four red dots are the additional points $-1, 2$ and $1/2 \pm (\sqrt{3}/2)i$ in K_1, and K_2 consists of all 179 points where the blue and green lines and circles intersect. The number of points in each K_n is of course finite, but (probably) grows doubly exponentially in n.



Note that each K_n is conjugation invariant (meaning $z \in K_n \implies \bar{z} \in K_n$) since K is, so $\mathbb{K}$ is as well. In fact $\mathbb{K}$ has many other wonderful properties, summarized by the remarkable:

Theorem 2.1 *The set $\mathbb{K}$ of constructible numbers is the quadratic closure $\sqrt{\mathbb{Q}}$ of the rationals inside $\mathbb{C}$, or equivalently* (*by* Lemma 1.7) *the union of all iterated quadratic extensions of $\mathbb{Q}$. In particular, $\mathbb{K}$ is algebraic over $\mathbb{Q}$, and each of its elements has degree a power of 2 over $\mathbb{Q}$.*

In practical terms, the first statement says that *a complex number is constructible if and only if it can be expressed by a formula using only integers, the algebraic operations $+, -, \cdot, \div$ and the extraction of square roots.* Thus for example the golden ratio $(1 + \sqrt{5})/2$ is constructible, as is its square root. The second statement gives a necessary (but not sufficient) condition for a complex number to be constructible, as we'll see in the applications below.

Proof of the Theorem It suffices to show that the set $\mathbb{K}$ has the following two properties:

ⓐ $\mathbb{K}$ is a quadratically closed field (so $\mathbb{K} \supset \sqrt{\mathbb{Q}}$).

ⓑ $\mathbb{K}$ lies in a union of iterated quadratic extensions of $\mathbb{Q}$ (so $\mathbb{K} \subset \sqrt{\mathbb{Q}}$).

The first property ⓐ is straightforward from "standard" constructions, and is left to the reader:

$\boxed{\text{HW\#39}}$ Show that $\mathbb{K}$ is closed under addition, multiplication, inversion (for nonzero elements), and taking square roots, i.e. if $a \neq 0$ and b are in $\mathbb{K}$ then so are $a + b$, ab, a^{-1} and $\sqrt{a}$.[†]

For ⓑ consider the fields $Q_n := \mathbb{Q}(K_n)$ obtained by adjoining all the numbers in K_n to $\mathbb{Q}$. Note that Q_n is conjugation invariant, since K_n is. We show by induction (starting with $Q_0 = \mathbb{Q}$)

[†] Hint: First recall how these operations are defined geometrically, e.g. to multiply two complex numbers, you multiply their lengths (a.k.a. norms) and add their angles (a.k.a. arguments). Then use HW#38 and Euclid's trick to help carry out the constructions. For $\sqrt{a}$ when a is a positive real number, note that the circle $C(i(a + 1)/2, -i)$ intersects the real axis at $\pm\sqrt{a}$.

that Q_n *is an iterated quadratic extension of* $\mathbb{Q}$, whence $\mathbb{K} = \cup K_n \subset \cup Q_n$ as desired. So assume this is true for some n. *It suffices to show that any* $z \in K_{n+1}$ *lies in a quadratic extension of* Q_n.

By definition z lies in the intersection of two K_n-curves. The key observation is that any such curve is the zero set of a polynomial in $Q_n[z, \bar{z}]$ of degree 1 (for lines) or 2 (for circles), where $z\bar{z}$ is the unique quadratic term in the latter case. In particular

$$L(a,b) \;=\; \{z \in \mathbb{C} \mid (z-a)/(b-a) \in \mathbb{R}\} \;=\; \{z \in \mathbb{C} \mid \ell_{ab}(z,\bar{z}) = 0\}$$

and

$$C(a,b) \;=\; \{z \in \mathbb{C} \mid |z-a| = |b-a|\} \;=\; \{z \in \mathbb{C} \mid q_{ab}(z,\bar{z}) = 0\}$$

where $\ell_{ab}(z,\bar{z}) = (\bar{b}-\bar{a})z - (b-a)\bar{z} + (\bar{a}b - a\bar{b})$ and $q_{ab}(z,\bar{z}) = z\bar{z} - \bar{a}z - a\bar{z} + (a\bar{b} + \bar{a}b - b\bar{b})$, which both lie in $Q_n[z,\bar{z}]$ since Q_n is a conjugation invariant field.

It follows that the intersection point p of any two nonparallel K_n-lines lies in Q_n (not extended) by solving for $\bar{z}$ in one equation and substituting in the other to give a single linear equation in z with coefficients in Q_n for p. Also the intersection points p, q of a K_n-line and K_n-circle lies in a quadratic extension of Q_n, since solving for $\bar{z}$ in the linear equation and substituting this in the quadratic equation gives a single quadratic equation in z with coefficients in Q_n whose roots are p and q. And finally the intersection points of two K_n-circles lies in a quadratic extension of Q_n, since replacing one of the quadratic equations by its difference from the other reduces to the previous case. This completes the proof. □

Corollary 2.2 *The cube cannot be doubled* (with a straightedge and compass).

Proof If it could, then $\sqrt[3]{2} \in \mathbb{K}$, but $\deg(\sqrt[3]{2}/\mathbb{Q}) = 3 \;\Longrightarrow\!\Longleftarrow$. □

Corollary 2.3 *The circle cannot be squared.*

Proof If it could, then $\pi \in \mathbb{K}$, but $\deg(\pi/\mathbb{Q}) = \infty$ (i.e. π is transcendental) $\Longrightarrow\!\Longleftarrow$. □

Corollary 2.4 *Not all angles can be trisected. For example, a* 60° *angle cannot be trisected.*

Proof If it could, then $c = \cos 20^\circ \in \mathbb{K}$. But c is a root of the polynomial $8x^3 - 6x - 1$[†], which is irreducible (substituting $y = 2x$ gives $y^3 - 3y - 1$, which is irreducible by Eisenstein's criterion) and so equal to $m_c(x)$. Thus $\deg(c/\mathbb{Q}) = 3 \;\Longrightarrow\!\Longleftarrow$. □

While we're on the subject, the ancient Greek problem of identifying which regular n-gons are constructible remains open! For example there are only 31 known odd values of n for which constructions are known to exist (this follows from a remarkable theorem of Gauss and Wantzel that *the* n*-gon is constructible* $\Longleftrightarrow$ n *is a power of* 2 *times a product of distinct* Fermat primes, i.e. primes of the form $2^{2^n} + 1$, of which only 5 are known: $5, 3, 17, 257$ and 65537). Here is a ”simple” warm-up problem in this subject; you have the tools to solve it (maybe with a little help from the internet when $n = 5$, but without using the Gauss-Wantzel theorem).

HW#40 Show that a regular n-gon can be constructed for $3 \le n \le 6$, but not for $n = 7$.

[†] This follows from the triple angle formula $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$, which in turn follows by equating the real parts of the identity $\cos 3\theta + i \sin 3\theta = (\cos\theta + i\sin\theta)^3$

THE NOTES THAT FOLLOW ARE VERY ROUGH AND NEED REVISION ...

§3. The Isomorphism Extension Theorem

Recall that any nonzero field morphism $g : E \to E'$ is automatically one-to-one since its kernel is a proper ideal in E, and therefore trivial. When E and E' both lie in a possibly larger field (for example $\mathbb{C}$) and g fixes some field $F \subset E \cap E'$ pointwise (i.e. $g(x) = x$ for all $x \in F$), then g is called an F-embedding, and in particular an F-isomorphism if it is also onto. Any F-isomorphism $E \to E$ is called an F-automorphism of E. The following simple result is extraordinarily useful:

Lemma 3.1 (Root Lemma) *If F, E and E' are subfields of some larger field with $F \subset E \cap E'$, and $g : E \to E'$ is an F-embedding, then g maps the roots in E of any polynomial $f \in F[x]$ to roots of f in E'. Thus any F-automorphism of E permutes the roots in E of any polynomial in $F[x]$.*

Proof If $f(e) = 0$ for $e \in E$, then since g is an F-morphism, $f(g(e))) = g(f(e)) = g(0) = 0$. □

The Galois group of a field extension E/F is the group

$$\mathrm{Gal}(E/F) \;=\; \{F\text{-automorphisms of } E\}$$

under composition. This group will be featured in our study of polynomial equations below. A crucial ingredient in that study is the following:

Theorem 3.2 (Isomorphism Extension Theorem) *Fix an isomorphism $f\colon F \to F'$ between two subfields F and F' of a field E. Let $p \in F[x]$ be irreducible with p' the corresponding irreducible in $F'[x]$,[†] and a and a' be arbitrarily chosen roots in E of p and p', respectively. Then f can be extended uniquely to an isomorphism $F(a) \to F'(a')$ carrying a to a'. In particular, applying this result to $\mathrm{id}\colon F \to F$: For any two roots a and b in E of the same irreducible polynomial in $F[x]$, there is a unique F-isomorphism $F(a) \to F(b)$ mapping a to b.*

Proof Recall from Theorem 1.2a that each element of $F(a)$ can be written uniquely as $p(a)$ for some polynomial p of degree less than $\deg(a/F)$. It is now a routine exercise to show that the map sending $p(a)$ to $p'(a')$, where $a' = f(a)$, is the desired isomorphism. □

§4. Separability and the Primitive Element Theorem

For simplicity, we henceforth work entirely inside the field of complex numbers $\mathbb{C}$. Thus by the fundamental theorem of algebra, *every* (complex) polynomial of degree n has exactly n roots (counting multiplicities), i.e. every $f(x) \in \mathbb{C}[x]$ factors in $\mathbb{C}[x]$ into linear factors.

Definition A complex polynomial of degree n is said to be separable if it has n *distinct* complex roots, i.e. if it factors into *distinct* linear factors in $\mathbb{C}[x]$. For example $x^2 - 1 = (x + 1)(x - 1)$ is separable, whereas $x^2 - 2x + 1 = (x - 1)^2$ is not.

Lemma 4.1 *If F is a subfield of $\mathbb{C}$ and $f \in F[x]$ is irreducible, the f is separable.*

Proof Let a be any root of f in $\mathbb{C}$. Then by Theorem 1.1a, f is (up to a constant multiple) the minimal polynomial of a over F. If a was a multiple root of f, meaning $f(x)$ factors over $\mathbb{C}$ as

[†] If $p(x) = p_0 + p_1x + \cdots$, then by definition $p'(x) = p_0' + p_1'x + \cdots$ where $p_i' = f(p_i)$. The map $F[x] \to F'[x]$ sending p to p' is easily seen to be an isomorphism that carries irreducibles to irreducibles.

$(x-a)^2g(x)$, then a would also be a root of the derivative $f'(x) = (x-a)(2g(x) + (x-a)g'(x))$. By Theorem 1.1a, f divides f'. But this is impossible since f' is nonzero (this is where $F \subset \mathbb{C}$ is used) of degree less than $\deg f$. Thus a is a simple root of f. □

Definition A field extension E/F is simple if E is obtained from F by adjoining a *single* element, i.e. $E = F(e)$ for some $e \in E$. If this is the case then e is called a primitive element for the extension.

Theorem 4.2 (Primitive Element Theorem) *All finite extensions inside* $\mathbb{C}$ *are simple.*

Proof By induction, it suffices to show that any algebraic extension $E = F(a,b)$ of a field $F \subset \mathbb{C}$ can be written as $F(a+bc)$ for some $c \in F$. In fact this is true for any $c \neq (a'-a)/(b'-b)$, where a' and b' are roots of $m_{a/F}$ and $m_{b/F}$. For then we easily show $b \in F(a+bc) =: K$ (or equivalently $m_{b/K}$ is linear) whence $a \in K$ as well (since $c \in F$) as follows: Consider the polynomial $p(x) = m_{a/F}(a + (x-b)c) \in E[x]$. Clearly b is a common root of p and $m_{b/F}$, in fact their only common root by the choice of c, so $m_{b/K}$ divides both of these polynomials. But if $m_{b/K}$ were not linear, then it would have another root (by Lemma 4.1) which would provide another common root of p and $m_{b/F}$. Thus $m_{b/K}$ is linear, and we're done. □

Remark Theorem 4.2 can also be deduced from a more general theorem of Artin (by the same name) *characterizing* arbitrary finite simple extensions E/F as those having only finitely many intermediate fields between F and E. These intermediate fields will play a big role in what follows.

Corollary 4.3 *If* E/F *is a finite extension inside* $\mathbb{C}$, *then any embedding* $F \hookrightarrow \mathbb{C}$ *extends to an embedding* $E \hookrightarrow \mathbb{C}$.

Proof By the Primitive Element Theorem 4.2, $E = F(c)$ for some c, and so the result follows from the Isomorphism Extension Theorem 3.2. □

§5. Normal Extensions

Definition For any subfield F of $\mathbb{C}$ and polynomial $f \in F[x]$, the field obtained from F by adjoining all the roots $r_1, \ldots, r_n$ of f is called the splitting field of f over F, denoted F_f. It consists of all evaluations $h(r_1, \ldots, r_n)$ for rational functions h in n variables over F, and depends on both f and F.[†] A field extension E/F inside $\mathbb{C}$ is normal if E is the splitting field of some polynomial over F, that is, if $E = F_f$ for some $f \in F[x]$.

Example The extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is normal (since $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}_{x^2-2}$) whereas $\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$ is not (this will follow from Theorem 5.1c below, since $\mathbb{Q}(\sqrt[3]{2})$ does not contain $\sqrt[3]{2}e^{2\pi i/3}$).

Here are the properties that we will need about normal extensions:

Theorem 5.1 *Let* E/F *be normal* (*with both fields in* $\mathbb{C}$ *as usual*). *Then*

ⓐ *Any* F*-embedding* $g : E \hookrightarrow \mathbb{C}$ *is in fact an* F*-automorphism of* E.

ⓑ *If* K *is any "intermediate field"* K *lying between* F *and* E, *then* E/K *is normal.*

ⓒ *If* $p \in F[x]$ *is irreducible and has at least one root in* E, *then all of its roots lie in* E. *Furthermore, there exists an* F*-automorphism of* K *carrying any one root of* p *onto any other root of* p.

[†] For example if $f(x) = x^2 + 1$, then $\mathbb{Q}_f = \mathbb{Q}(i)$, whereas $\mathbb{R}_f = \mathbb{C}_f = \mathbb{C}$.

Proof By hypothesis $E = F(r_1, \ldots, r_n)$, where $r_1, \ldots, r_n$ are the roots of some polynomial $f \in F[x]$, or explicitly (as noted above)

$$E \;=\; \{p(r_1, \ldots, r_n) \;\mid\; p \in F[x_1, \ldots, x_n]\}.$$

By the Root Lemma 3.1, g permutes the r_i's, and so $p(E) = E$. This proves part ⓐ). Part ⓑ is obvious, since $E = E_{f/K}$.

For ⓒ, suppose p has a root $a \in E$, and let b be any other root of p. By Theorem 3.2, there is an F-isomorphism $F(a) \to F(b)$ carrying a to b, which extends using Corollary 4.3 to an F-embedding $g : E \hookrightarrow \mathbb{C}$. Since E/F is normal, $g(E) = E$ (by part a) and so $b = g(a) \in E$, which also proves the last statement in c). □

Corollary 5.2 *If E/F is normal of degree n, then the Galois group Gal(E/F) has order n, and can naturally be viewed as a subgroup of the symmetric group S_n.*

Proof By the Primitive Element Theorem 4.2, $E = F(c)$ for some $c \in E$. Let m denote the minimal polynomial of c over F. Thus $n = \deg(m)$. By Theorem 5.1c, E contains all the roots $c_1, \ldots, c_n$ of m (with say $c_1 = c$), and for each c_i there exists $g_i \in \mathrm{Gal}(E/F)$ with $g_i(c) = c_i$, which is unique since c generates E over F. By Lemma 3.1, there are no other elements in $\mathrm{Gal}(E/F)$. Thus $|\mathrm{Gal}(E/F)| = n$. We view $\mathrm{Gal}(E/F) < S_n$ by how its elements permute the roots of m. □

Examples ① $\mathrm{Gal}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) \cong C_2 \cong S_2$ generated by $\sqrt{2} \mapsto -\sqrt{2}$.

② $\mathrm{Gal}(\mathbb{C}/\mathbb{R}) \cong C_2 \cong S_2$ generated by complex conjugation.

③ $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}) \cong C_2 \times C_2 < S_4$, consisting of the four $\mathbb{Q}$-automorphisms $\sqrt{2} \mapsto \pm\sqrt{2}$, $\sqrt{3} \mapsto \pm\sqrt{3}$. (elaborate)

Remark If one works in fields other than $\mathbb{C}$ (in particular those of characteristic $p \neq 0$) then the extensions E/F that we call "Galois" are usually called "normal", and "Galois" then means normal *and* separable, the latter meaning $m_{e/F}$ is separable (i.e. has no multiple roots in any extension of F) for every $e \in E$. (Note that all extensions inside $\mathbb{C}$ are separable by Lemma 4.1.)

§6. Galois Theory

Fix a Galois extension E/F. Let $\mathcal{F}$ denote the set of all intermediate fields of the extension, and $\mathcal{G}$ the set of all subgroups of its Galois group.

Fundamental Theorem of Galois Theory (FTGT) *For any Galois extension E/F and with $\mathcal{F}$ and $\mathcal{G}$, there is a natural inclusion reversing bijection*

$$\begin{tikzcd}
\mathcal{F} = \mathcal{F}(E/F) \arrow[r, bend left, "\text{Fixer}"] & \mathcal{G} = \mathcal{G}(E/F) \arrow[l, bend left, "\text{FixedField}"]
\end{tikzcd}$$

defined as follows: For each $K \in \mathcal{F}$, let K' ...

§7. Galois Theory over $\mathbb{Q}$ (following John Stillwell)

Fix a monic polynomial $f(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1x + c_0$ in $\mathbb{Q}[x]$ of degree n. By the fundamental theorem of algebra, f has exactly n complex roots $r_1, \ldots, r_n$ counting multiplicities, i.e. $f(x)$ factors as $f(x) = (x - r_1)\cdots(x - r_n)$. Evidently the coefficients $c_0, \ldots, c_{n-1}$ all lie in the splitting field

$$\mathbb{Q}_f \ := \ \mathbb{Q}(r_1, \ldots, r_n)$$

of f, obtained from $\mathbb{Q}$ by adjoining all the roots of f.[†] Conversely, we seek a formula for the roots r_i built up from the coefficients c_i using only the field operations $+, -, \cdot$ and $\div$, and the extraction of roots. We call such a formula a radical formula for the roots of f. The quadratic formula is a general radical formula that works for all polynomials of degree $n = 2$, and there are similar formulas for $n = 3$ and 4. We will show below that there are no such formulas for $n \geq 5$.

Definition A field extension E/F is an elementary radical extension if $E = F(\sqrt[p]{b})$ for some $p \geq 2$ and $b \in F$ that is not a perfect pth power, i.e. $E = F(a)$ for some $a \in E - F$ satisfying $a^p \in F$. If p is prime and E contains either *all* of the pth roots of b (recall that there are exactly p of them in $\mathbb{C}$), or *exactly one* (namely a), then we call E/F a special elementary radical extension. In general, any extension obtained by a finite sequence of elementary radical extensions is called a radical extension.

Definition A polynomial $f \in \mathbb{Q}[x]$ is solvable if $\mathbb{Q}_f$ lies in some radical extension of $\mathbb{Q}$. Thus the solvability of f means the existence of a radical formula for its roots.

Our goal is to show:

Galois' Theorem *There exist nonsolvable polynomials $f \in \mathbb{Q}[x]$.*

The input from group theory is:

Definition A group G is solvable if there is a filtration $G = G_0 \supset G_1 \supset \cdots \supset G_n = 1$ by normal subgroups such that G_i/G_{i+1} is abelian for all i.

Definition The Galois group of an extension E/F is the group $\mathrm{Gal}(E/F)$ of all automorphisms of E that fix F pointwise. The Galois group $\mathrm{Gal}(f)$ of a polynomial $f \in \mathbb{Q}[x]$ is $\mathrm{Gal}(\mathbb{Q}_f/\mathbb{Q})$.

Here are the key facts needed to prove Galois' theorem:

- If f is an irreducible polynomial of degree n over $\mathbb{Q}$, then $\mathrm{Gal}(f) \cong S_n$.
- S_n is not solvable for $n \geq 5$.

§8. Computing Galois groups of a separable polynomial $f \in K[x]$

We view Gal_f as a group of permutations of the roots of f. For polynomials of small degree, it is therefore useful to recall what the proper nontrivial subgroups of S_n are for small n. In particular:

- For S_3, there are two C_2's and $A_3 = C_3$.
- For S_4, there are nine C_2's, four C_3's, three C_4's, four V_4's, four S_3 's, three D_8's, and A_4.

[†] Indeed they are ± the elementary symmetric polynomials in the roots: $c_k = (-1)^{n-k}\mathbf{e_k}(r_1, \ldots, r_n)$ where $\mathbf{e_k}(r_1, \ldots, r_n) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} r_{i_1} \cdots r_{i_k}$. Special cases are $c_0 = (-1)^n r_1 \cdots r_n$ and $c_{n-1} = -(r_1 + \cdots + r_n)$.

Two important notions: A subgroup of S_n is transitive if for all $1 \leq i, j \leq n$, some element in the subgroup maps i to j. The discriminant of a polynomial f, with roots $r_1, \ldots, r_n$, is

$$\Delta_f \;=\; \prod_{i \neq j} (r_i - r_j).$$

Note that Δ_f lies in K.

Here are some general results about a separable polynomial $f \in K[x]$ of degree n:

1. f is irreducible iff $\mathrm{Gal}(f)$ is a transitive subgroup of S_n.
2. $\mathrm{Gal}(f) \subset A_n$ iff Δ_f is a square in K.