Introductory Algebra

Course No. 100 321

Fall 2005

Michael Stoll

Contents

1. Monoids and Groups

1.1. **Definition.** A *monoid* is a set M, together with a binary operation

$$\cdot : M \times M \longrightarrow M$$

and an element $e \in M$ (called the *neutral* or *identity* element) satisfiying the following axioms.

(M1) $\forall a, b, c \in M : (a \cdot b) \cdot c = a \cdot (b \cdot c)$ (associativity);
(M2) $\forall a \in M : a \cdot e = e \cdot a = a$.

M is called *commutative* if in addition, we have

(M3) $\forall a, b \in M : a \cdot b = b \cdot a$.

Usually, the dot is dropped, and we simply write ab for $a \cdot b$. In the commutative case, the operation is often denoted by $+$ (and the identity element e by 0).

Note that the identity element e is uniquely determined: if e and e' both are identities, then

$$e = ee' = e' .$$

Therefore, e is usually suppressed in the notation: instead of $(M, \cdot, e)$, we just write $(M, \cdot)$ or even M, when the operation is clear from the context.

1.2. **Definition.** Let M be a monoid. $a \in M$ is called *left invertible* if there exists $b \in M$ such that $ba = e$. a is called *right invertible* if there exists $c \in M$ such that $ac = e$. a is called *invertible* if a is both left and right invertible.

Note that if a is invertible, then left and right inverse coincide: if $ba = e$ and $ac = e$, then

$$b = be = b(ac) = (ba)c = ec = c .$$

We write a^{-1} for the inverse of a. The above also shows that the inverse is uniquely determined (every left inverse equals every right inverse if at least one of each kind exists).

1.3. **Definition.** A monoid such that each of its elements is invertible is called a *group.*

If G is a group with operation written multiplicatively, we usually denote the identity element by 1. If G is an *abelian* group, i.e., a group that is a commutative monoid, the operation is usually written $+$, and the identity element denoted 0.

1.4. **Examples.**

(1) The *trivial monoid* or *trivial group* is $\{e\}$ with the unique choice of binary operation on it.
(2) $(\mathbb{Z}, +)$ is an abelian group, $(\mathbb{N}, +)$ is a commutative monoid that is not a group (note: we use $\mathbb{N} = \{0, 1, 2, \dots\}$!). $(\mathbb{N} \setminus \{0\}, +)$ is not even a monoid (it is a *semigroup*; it only satisfies (M1)).
(3) $(\mathbb{Z}, \cdot)$, $(\mathbb{Z} \setminus \{0\}, \cdot)$, $\mathbb{N}, \cdot)$, $(\mathbb{N} \setminus \{0\}, \cdot)$ all are monoids, but not groups.
(4) Let $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$ with operations $+$ and $\cdot$ "mod n" (i.e., add or multiply as usual, then replace by the remainder of division by n).

    $(\mathbb{Z}_n, +)$ is an abelian group, $(\mathbb{Z}_n, \cdot)$ is a commutative monoid. $(\mathbb{Z}_5 \setminus \{0\}, \cdot)$ is a monoid, even a group, but $(\mathbb{Z}_4 \setminus \{0\}, \cdot)$ does not even make sense, since $\cdot$ is not a binary operation on $\mathbb{Z}_4 \setminus \{0\}$.

(5) Let F be a field. Then (by definition) $(F,+)$ and $(F \setminus \{0\}, \cdot)$ are abelian groups. Let $\mathrm{Mat}(n,F)$ be the set of all $n \times n$ matrices over F, with the ususal addition and multiplication of matrices. Then $(\mathrm{Mat}(n,F),+)$ is an abelian group and $(\mathrm{Mat}(n,F),\cdot)$ is a monoid that is not commutative when $n \geq 2$.

(6) Let X be any set, and let M be the set of all maps $X \to X$. Then $(M, \circ)$ is a monoid (where $\circ$ denotes composition of maps), which is not commutative when X has at least two elements.

2. Submonoids and Subgroups

As usual in mathematics (and in algebra in particular), with any structure (like groups, monoids, rings, fields, vector spaces, topological spaces, ...), we are interested in its *substructures.*

2.1. **Definition.** Let M be a monoid and $M' \subset M$ a subset. We call M' a *submonoid* of M if the binary operation on M restricts to a binary operation on M' and M' contains the identity element of M. We call M' a *subgroup* of M if it is a submonoid that is a group.

When H is a subgroup of a group G, we write $H \leq G$.

Every group G has two *trivial subgroups* (which may coincide), namely G itself and the trivial group $\{e\}$. A subgroup of G, which is not G itself, is called a *proper* subgroup.

2.2. **Examples.**

(1) $(\mathbb{N},+)$ is a submonoid of $(\mathbb{Z},+)$.

(2) $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ is a subgroup of $\mathbb{C}^\times = (\mathbb{C} \setminus \{0\}, \cdot)$ (the "unit circle").

(3) $(\{\pm 1\}, \cdot)$ is a subgroup of $(\mathbb{Z}, \cdot)$ (in fact, it is the largest subgroup).

This last example raises the question whether there is a largest subgroup in every monoid. The answer is yes.

2.3. **Proposition.** Let M be a monoid, and let

$$U(M) = \{a \in M : a \text{ is invertible}\}.$$

Then $U(M)$ is the largest subgroup of M.

Proof. If $G \subset M$ is a subgroup, then all its elements are invertible, hence $G \subset U(M)$. So we only have to show that $U(M)$ is a subgroup of M. We first check that $U(M)$ is closed under $\cdot$. Let a and b be invertible. Let $c = b^{-1}a^{-1}$. Then

$$c(ab) = b^{-1}(a^{-1}a)b = b^{-1}b = e \quad \text{and} \quad (ab)c = a(bb^{-1})a^{-1} = aa^{-1} = e\,,$$

so ab is invertible, with inverse $(ab)^{-1} = b^{-1}a^{-1}$. Now if a is invertible, then so is a^{-1} (its inverse is a), so all elements of $U(M)$ are invertible *within* $U(M)$. Finally, e is always invertible, so $e \in U(M)$. □

2.4. **Examples.**

(1) $U(\mathbb{Z}, \cdot) = \{\pm 1\}$.
(2) $U(F, \cdot) = F^\times = F \setminus \{0\}$, where F is a field.
(3) $U(\mathrm{Mat}(n, F), \cdot) = \mathrm{GL}(n, F)$, the *general linear group* of invertible $n \times n$ matrices.
(4) $U(\mathbb{Z}_n, \cdot) = \mathbb{Z}_n^\times$ is the set of invertible residue classes mod n. We will see soon that

$$\mathbb{Z}_n^\times = \{a \in \mathbb{Z}_n : \gcd(a, n) = 1\} .$$

(5) If $(M, \circ)$ is the monoid of maps $X \to X$ for a set X, then $U(M, \circ) = S(X)$ is the group of *permutations* of X. We write

$$S_n = S(\{1, 2, \ldots, n\}) ;$$

this is called the *symmetric group* on n objects. S_n is a non-abelian group as soon as $n \geq 3$.

Before we formulate an alternative and sometimes useful criterion for when a subset of a group is already a subgroup, let us prove a result that tells us that under certain conditions a monoid is already a group.

The result as such is maybe not so important (though useful), but it is a prototype of a class of results of the type "if a certain type of algebraic structure satisfies a finiteness property, then it actually is a more special structure". Examples of this principle are:

- A finite integral domain is a field.
- A finite skew field (or division ring) is a field.
- A compact complex Lie group is abelian.

But we need a definition first. In a group, we certainly have the implication

$$ab = ac \quad \Longrightarrow \quad b = c$$

(multiply from the left by the inverse of a).

2.5. **Definition.** A monoid M is *(left) cancellative* if for all $a, b, c \in M$, we have

$$ab = ac \quad \Longrightarrow \quad b = c .$$

It is certainly reasonable to ask a monoid to be cancellative, before we consider the question whether it may already be a group.

2.6. **Theorem.** *A finite cancellative monoid is already a group.*

Proof. The idea of the proof is to show that every element of M is right invertible. It follows that every element is invertible and so M is in fact a group: let $a \in M$, then there is $b \in M$ with $ab = e$ and there is $c \in M$ with $bc = e$, so b is invertible, and we must have $a = c$ (equality of right and left inverses), so $ba = e$, and b is the inverse of a.

Here is another important idea. We use the binary operation on M to construct maps $M \to M$. Let $a \in M$, then left multiplication by a gives a map

$$\ell_a : M \longrightarrow M , \quad m \longmapsto am .$$

Clearly, this map is injective if M is left cancellative (essentially by definition). Now M is finite, therefore every injective map $M \to M$ is also surjective (this is a possible definition of finiteness!). So ℓ_a is surjective, hence there is $b \in M$ such that $ab = \ell_a(b) = e$, meaning that a is right invertible. Since a was an arbitrary element of M, all elements of M are right invertible, hence M is a group. □

2.7. **Corollary.** $U(\mathbb{Z}_n, \cdot) = \{a \in \mathbb{Z}_n : \gcd(a, n) = 1\}$.

Proof. If a is invertible mod n, then there are integers b and k such that $ab = 1+kn$, which implies $\gcd(a, n) = 1$. To show the reverse inclusion, we note that the right hand side is a cancellative monoid (exercise). By the preceding theorem, it is a group and therefore contained in $U(\mathbb{Z}_n, \cdot)$. □

2.8. **Corollary.** *$\mathbb{Z}_p$ is a field.*

Often, the field $\mathbb{Z}_p$ is denoted $\mathbb{F}_p$.

Now for some criteria for when a subset of a group is a subgroup.

2.9. **Proposition.** *Let G be a group, $H \subset G$ a nonempty subset. Then H is a subgroup of G if one of the following conditions is satisfied.*

(1) *For all $h_1, h_2 \in H$, we have $h_1 h_2^{-1} \in H$.*
(2) *H is closed under the group operation and finite.*

Proof. Assume the first condition holds. Since H is nonempty, there is some $h \in H$. Therefore, we have $e = hh^{-1} \in H$ and then also $h^{-1} = eh^{-1} \in H$. Finally, with $h_1, h_2 \in H$, we have $h_2^{-1} \in H$ and then $h_1 h_2 = h_1 (h_2^{-1})^{-1} \in H$.

Now consider the second condition. Let $h \in H$. The set of all powers of h, $\{h^n : n \in \mathbb{Z}\}$, is finite, so there are $m > n$ such that $h^m = h^n$. It follows that $e = h^{m-n} \in H$. Now H is a finite monoid and cancellative (because it is a submonoid of a group), hence by the theorem, it is already a group. □

2.10. **Lemma.** *Let G be a group, $(H_i)_{i \in I}$ a collection of subgroups of G. Then $H = \bigcap_{i \in I} H_i$ is again a subgroup of G.*

Proof. Every H_i contains e, so $e \in H$ as well. Now let $h_1, h_2 \in H$. Then $h_1 h_2^{-1} \in H_i$ for all $i \in I$, so $h_1 h_2^{-1} \in H$. By the preceding proposition, H is a subgroup. □

This observation shows that it makes sense to talk of the smallest subgroup containing a given subset of G: we just take the intersection of all subgroups containing the subset. (Note that there is always one such subgroup, namely G itself.)

2.11. **Definition.** Let G be a group and $S \subset G$ a subset. The smallest subgroup of G containing S is called the subgroup *generated by* S and denoted $\langle S \rangle$ (or $\langle s_1, s_2, \dots \rangle$ if $S = \{s_1, s_2, \dots\}$).

If $\langle S \rangle = G$, we say that S *generates* G or that S is a *generating set.* If G has a finite generating set, we say that G is *finitely generated.*

A group that is generated by one element is called *cyclic.* If $G = \langle g \rangle$, then $G = \{g^n : n \in \mathbb{Z}\}$ consists of all powers of g.

2.12. **Definition.** If G is a group, then its cardinality $\#G$ is called the *order* of G. The *order* $o(g)$ of an element $g \in G$ is the order of the group $\langle g \rangle$ it generates. This is the smallest positive integer n such that $g^n = e$, if such an n exists; otherwise, g is of infinite order.

2.13. **Remark.** There are essentially only the following cyclic groups (up to *isomorphism;* see later).

(1) For each $n \geq 1$, the group $(\mathbb{Z}_n, +)$ — the cyclic group of order n.
(2) The infinite cyclic group $(\mathbb{Z}, +)$.

3. Cosets and Lagrange's Theorem

A subgroup H of a group G leads in a natural way to an equivalence relation on G.

3.1. **Defintion.** Let H be a subgroup of the group G. A *right coset* of H in G is a set of the form

$$Hg = \{hg : h \in H\}$$

for some $g \in G$. A *left coset* of H in G is a set of the form

$$gH = \{gh : h \in H\}$$

for some $g \in G$.

Note that right (left) multiplication by g gives a bijection between H and Hg (gH).

3.2. **Lemma.** *If H is a subgroup of the group G and $g, g' \in G$, then the following statements are equivalent.*

(1) $Hg' \subset Hg$
(2) $g' \in Hg$
(3) $g'g^{-1} \in H$
(4) $Hg' = Hg$

Proof. The equivalence of the first three is easy. To see that they imply the last, note that $g'g^{-1} \in H$ implies $g(g')^{-1} \in H$, which implies $Hg \subset Hg'$. □

3.3. **Proposition.** *If H is a subgroup of a group G and $g, g' \in G$, then either $Hg = Hg'$, or $Hg \cap Hg' = \emptyset$. In other words, the right cosets form a partition of G.*

Proof. Assume that $Hg \cap Hg' \neq \emptyset$. We have to show that $Hg = Hg'$. So let $x \in Hg \cap Hg'$. Then $x \in Hg$ and $x \in Hg'$, so by the preceding lemma, $Hg = Hx = Hg'$. □

Another way of phrasing this is to say that

$$g \sim g' \iff Hg = Hg' \iff g'g^{-1} \in H$$

defines an equivalence relation on G. The set of equivalence classes is denoted by $H\backslash G$. (If we consider left cosets gH, then we write G/H.)

3.4. **Theorem (Lagrange).** *If G is finite, then $\#H$ divides $\#G$.*

Proof. G is partitioned into disjoint subsets Hg, each of the same size as H. □

3.5. **Definition.** The cardinality of G/H is called the *index* of H in G and denoted $(G : H)$.

For example, the index of the subgroup $2\mathbb{Z}$ of $\mathbb{Z}$ is 2, even though $\mathbb{Z}$ is an infinite group.

Note that for this definition, it does not matter whether we consider left or right cosets. In fact, $Hg \mapsto g^{-1}H$ provides a bijection between the set of right cosets $H\backslash G$ and the set of left cosets G/H. (Exercise!)

3.6. **Corollary.** *If G is a finite group and $g \in G$, then $o(g)$ divides $\#G$. In particular, $g^{\#G} = e$.*

Proof. $\langle g\rangle \le G$, and $o(g) = \#\langle g\rangle$. □

3.7. **Corollary.** *If G is a group of prime order p, then G has no nontrivial subgroups. In particular, if $g \in G \setminus \{e\}$, then $G = \langle g\rangle$ (and so G is cyclic).*

Proof. If $H \le G$, then $\#H$ must be 1 or p, so $H = \{e\}$ or $H = G$. □

As an application, we can prove *Fermat's Little Theorem.*

3.8. **Theorem (Fermat).** *If p is a prime number and a an integer not divisible by p, then $a^{p-1} \equiv 1 \bmod p$ (i.e., p divides $a^{p-1} - 1$).*

Proof. It suffices to consider $1 \le a < p$. Then $a \in \mathbb{Z}_p^\times = \mathbb{Z}_p \setminus \{0\}$, a group of order $p - 1$. So $a^{p-1} = 1$ in $\mathbb{Z}_p^\times$, which means that $a^{p-1} \equiv 1 \bmod p$. □

We have seen that the order of an element divides the order of the group. This raises the question whether every divisor of the group order occurs as the order of an element. This will certainly be false in general: if $o(g) = \#G$, then $G = \langle g\rangle$ is cyclic, which is not true for all groups. But there is an important special case.

3.9. **Theorem (Cauchy).** *If G is finite and the prime number p divides $\#G$, then there is an element of G of order p.*

Proof. The proof uses a nice counting argument that is perhaps not very obvious. Let us consider the following set

$$T = \{(g_1, g_2, \dots, g_p) \in G^p : g_1 g_2 \cdots g_p = e\} .$$

Since for a tuple $(g_1, g_2, \dots, g_p) \in T$, the last entry g_p is uniquely determined by the first $p-1$ entries, which are otherwise arbitrary, T must have $\#G^{p-1}$ elements, a number that is divisible by p (since $\#G$ is, and $p - 1 \ge 1$).

On the other hand, if not all the g_j are the same, then we get p distinct elements of T by looking at

$$(g_1, g_2, \dots, g_p) , \quad (g_2, g_3, \dots, g_p, g_1) , \quad \dots , (g_p, g_1, \dots, g_{p-1}) .$$

Hence the number of elements $(g_1, g_2, \dots, g_p) \in T$, for which not all the g_j are identical, is also divisible by p.

This implies that the number of elements $(g, g, \dots, g) \in T$ must also be divisible by p. Now $(g, g, \dots, g) \in T$ is eqivalent with $g^p = e$, and there is at least one such element, namely e. Therefore there must be at least one other element $g \in G$, $g \ne e$, such that $g^p = e$. But this means that $o(g) = p$. □

Products of Subgroups.
We have seen that the intersection of two subgroups is again a subgroup. The union of two subgroups is rarely a subgroup. The next reasonable thing one can try is to consider the *product*

$$HK = \{hk : h \in H, k \in K\}$$

of two subgroups H and K. It is easy to find examples where HK is not a subgroup. However, there is the following result.

3.10. **Proposition.** *Let H, K be subgroups of a group G such that $HK = KH$. Then HK is again a subgroup of G.*

Proof. We use Prop. 2.9. HK is certainly nonempty. Let $h, h' \in H$ and $k, k' \in K$, so $hk, h'k' \in HK$. We need to show that $(hk)(h'k')^{-1} \in HK$. Now,

$$(hk)(h'k')^{-1} = hk(k')^{-1}(h')^{-1} \in HKH = HHK = HK\,.$$

□

In particular, when G is abelian, then HK is always a subgroup of G.

3.11. **Proposition.** *If H and K are finite subgroups of a group G, then*

$$\#(HK) = \frac{\#H\,\#K}{\#(H \cap K)}\,.$$

Proof. Exercise. □

4. Homomorphisms and Normal Subgroups

As usual in mathematics, we do not just want to study structures like groups in isolation, but we want to relate them to each other. So we need to introduce suitable structure-preserving maps.

4.1. **Definition.** Let G and G' be two groups. A *group homomorphism* from G to G' is a map $\phi : G \to G'$ such that $\phi(g_1g_2) = \phi(g_1)\phi(g_2)$ for all $g_1, g_2 \in G$. (Note that on the left, we have the operation in G, whereas on the right, it is in G'.)

If ϕ is bijective, then ϕ is called a *(group) isomorphism*, and G and G' are *isomorphic*. We then write $G \cong G'$. An isomorphism $\phi : G \to G$ is called an *automorphism* of G; the set of all automotphisms of G forms a group $\operatorname{Aut}(G)$ under composition of maps.

If $\phi : G \to G'$ is a group homomorphism, then $\ker\phi = \{g \in G : \phi(g) = e'\}$ (where e' is the identity element of G') is called the *kernel* of ϕ.

4.2. **Remark.** One would want to require that $\phi(e) = e'$ and $\phi(g^{-1}) = \phi(g)^{-1}$. Fortunately, these properties follow (Exercise). Also, if ϕ is bijective, then ϕ^{-1} is again a group homomorphism (Exercise).

(For a *monoid homomorphism*, on the other hand, one has to require that $\phi(e) = e'$, and for a *monoid isomorphism*, one needs to require that the inverse is again a homomorphism. Conclusion: groups are nicer than monoids!)

4.3. **Lemma.** *Let $\phi : G \to G'$ be a group homomorphism.*

(1) $\ker \phi$ *is a subgroup of* G.
(2) *If* $H \leq G$, *then* $\phi(H) \leq G'$.
(3) *If* $H' \leq G'$, *then* $\ker \phi \subset \phi^{-1}(H') \leq G$.
(4) *If* $\phi' : G' \to G''$ *is another group homomorphism, then* $\phi' \circ \phi : G \to G''$ *is also a group homomorphism.*
(5) ϕ *is injective if and only if* $\ker \phi = \{e\}$.

Proof. The first four statements are easy. For the last, note that if ϕ is injective, then e' can only have one preimage, which must be e. This proves one direction. Conversely, suppose $\ker \phi$ is trivial. Let $g_1, g_2 \in G$ such that $\phi(g_1) = \phi(g_2)$. Then

$$e' = \phi(g_1)\phi(g_2)^{-1} = \phi(g_1 g_2^{-1}),$$

so $g_1 g_2^{-1} \in \ker \phi = \{e\}$, whence $g_1 = g_2$. So ϕ is injective. □

In fact, more is true than just that $\ker \phi$ is a subgroup.

4.4. **Definition.** A subgroup $H \leq G$ is called *normal*, if one (and hence all) of the following equivalent conditions is satisfied.

(1) Left and right cosets of H coincide: $gH = Hg$ for all $g \in G$.
(2) For all $g \in G$, we have $gHg^{-1} = H$.
(3) For all $g \in G$, we have $gHg^{-1} \subset H$.

We write $H \lhd G$ if H is a normal subgroup of G.

Normal subgroups are nice and important because they allow us to define a natural group structure on the quotient G/H (which is the same as $H\backslash G$ for normal H). "Natural" means that we want the canonical map

$$\phi : G \longrightarrow G/H, \quad g \longmapsto gH$$

to be a group homomorphism. This implies that the only possible way to define a natural group structure on G/H is to set

$$(gH) \cdot (g'H) = gg'H.$$

In general, this will not even be well-defined.

4.5. **Lemma.** *Assume $H \leq G$. The definition $(gH) \cdot (g'H) = gg'H$ gives a well-defined binary operation on G/H if and only if H is a normal subgroup. In this case, this binary operation makes G/H into a group, and the canonical map $\phi : G \to G/H$ is a surjective group homomorphism with kernel H.*

Proof. In order for the binary operation to be well-defined, we need to have that $hgH = gH$ for all $h \in H$, $g \in G$ (since $(eH) \cdot (gH)$ should be gH, whereas $(hH) \cdot (gH)$ should be hgH, but $eH = H = hH$). This implies $g^{-1}hg \in H$ for all $g \in G$, $h \in H$; whence $H \lhd G$.

Conversely, if $H \lhd G$, then $gH\,g'H = g\,g'H\,H = gg'H$, and the definition of $\cdot$ coincides with the product of cosets and so is well-defined. Finally, it is a general fact that if $\phi : G \to X$ is a surjective map from a group G to a set X with binary operation such that $\phi(gg') = \phi(g)\phi(g')$, then X is also a group (Exercise). It is clear that the kernel of ϕ is H. □

The group G/H is called the *quotient group* (or *factor group*) of G by H. The basic idea is that G/H is some sort of coarser image of G, the "information lost" being what is contained in H.

4.6. **Remark.** If $H \leq G$ and $(G : H) = 2$, then $H \triangleleft G$. (Exercise!)

4.7. **Remark.** The intersection of any family of normal subgroups is again a normal subgroup.

The preceding lemma shows that every normal subgroup is the kernel of a group homomorphism. The converse is also true.

4.8. **Theorem.** *Let $\phi : G \to G'$ be a group homomorphism. Then* $\ker \phi$ *is a normal subgroup of G, and there is a unique isomorphism $\psi : G/\ker\phi \to \phi(G')$ making the diagram below commutative.*

$$\begin{tikzcd}
G \arrow[r, "\phi"] \arrow[d] & G' \\
G/\ker\phi \arrow[r, "\cong", "\psi"'] & \phi(G) \arrow[u, hook]
\end{tikzcd}$$

Proof. We first show that $\ker\phi$ is a normal subgroup. Let $h \in \ker\phi$ and $g \in G$. We have to show that $ghg^{-1} \in \ker\phi$ as well. Now

$$\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g^{-1}) = \phi(g)e'\phi(g)^{-1} = e'$$

and so $ghg^{-1} \in \ker\phi$.

Let $H = \ker\phi$. Now in order to make the diagram above commutative, ψ must satisfy $\psi(gH) = \phi(g)$. We need to check that this is a well-defined map. But this follows from $\phi(gh) = \phi(g)e' = \phi(g)$ for $g \in G$, $h \in H$. It is then clear that ψ is a homomorphism, and it remains to show that ψ is injective and has image $\phi(G)$. The latter is clear. For the former, we note that the kernel of ψ is just $\{H\}$ (the trivial one-element subgroup of G/H). □

4.9. **Corollary.** *If $\phi : G \to G'$ is a* surjective *group homomorphism, then we have $G' \cong G/\ker\phi$. In particular, $\#G' = (G : \ker\phi)$.*

4.10. **Corollary.** *If $\phi : G \to G'$ is a* surjective *group homomorphism, then there is a bijection between the subgroups of G containing* $\ker\phi$ *and the subgroups of G', given by images/preimages under ϕ.*

Proof. By Lemma 4.3, the preimage of a subgroup of G' is a subgroup of G containing $\ker\phi$, and the image of a subgroup of G is a subgroup of G'. Since ϕ is surjective, we have $\phi(\phi^{-1}(H')) = H'$ for any subgroup $H' \leq G'$. On the other hand, it is easy to see that in general, $\phi^{-1}(\phi(H)) = H \cdot \ker\phi$. So if $\ker\phi \subset H$, then $\phi^{-1}(\phi(H)) = H$. □

4.11. **Definition.** A group $G \neq \{e\}$ such that G has no nontrivial normal subgroups (i.e., the only normal subgroups are $\{e\}$ and G itself) is called *simple.*

Simple groups play a role in group theory that is somewhat analogous to the role played by prime numbers in number theory: in some sense, one can build up all groups from simple groups, and the simple groups needed for any given group are uniquely determined.

The *classification of the finite simple groups* has been completed (or so one believes) in the last century; the proof (which is distributed over many research papers) has several thousand pages. The result is that there are 18 infinite families of finite simple groups (two of which we will get to know soon), plus 26 so-called "sporadic" simple groups, which are rather interesting (and sometimes mysterious) objects.

4.12. **Example.** A group of prime order is simple. Indeed, such a group is not the trivial group, and we have seen that it only has the two trivial subgroups. In fact, these are exactly the *abelian* finite simple groups. To see why, let G be a finite abelian group of non-prime order and let p be a prime divisor of $\#G$. By Cauchy's Theorem 3.9, G has an element g of order p. Then $\langle g \rangle$ is a subgroup of G of order p, so it is a nontrivial subgroup, and it is normal since G is abelian.

4.13. **Example.** The subgroups of $\mathbb{Z}$ are all of the form

$$n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$$

for some $n \geq 0$. This is clear for the trivial subgroup $\{0\}$ (take $n = 0$). For a nontrivial subgroup H of $\mathbb{Z}$, let n be the smallest positive element of H (H has positive elements: it has a nonzero element k, and then it also contains $-k$, and one of the two is positive). Then $n\mathbb{Z} = \langle n \rangle \subset H$. Now let $k \in H$ be any element. Then we can write $k = nq + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < n$. So $r \in H$, but then r must be zero, since n was the *smallest* positive element of H. Hence $k = nq \in n\mathbb{Z}$, and so $H = n\mathbb{Z}$.

Now $\mathbb{Z}$ is abelian, and so all subgroups are already normal. Therefore, for every $n \geq 1$, there exists the quotient group $\mathbb{Z}/n\mathbb{Z}$. The above argument using division with remainder shows that as a set,

$$\mathbb{Z}/n\mathbb{Z} = \{0 + n\mathbb{Z}, 1 + n\mathbb{Z}, \ldots, (n-1) + n\mathbb{Z}\},$$

and addition in $\mathbb{Z}/n\mathbb{Z}$ is addition "mod n" of representatives. We conclude that $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$ (as additive groups).

Another way of looking at this is to say that the map

$$\phi : \mathbb{Z} \longrightarrow \mathbb{Z}_n\,, \qquad k \longmapsto k \bmod n$$

is a surjective group homomorphism (by definition of the addition in $\mathbb{Z}_n$) and has kernel $n\mathbb{Z}$. Therefore $\mathbb{Z}_n \cong \mathbb{Z}/n\mathbb{Z}$.

4.14. **Example.** You remember the sign of a permutation from Linear Algebra: Any element $\sigma \in S_n$ can be written as a product of transpositions; we write $\operatorname{sign}(\sigma) = 1$ if one (and then every) such product has an even number of transpositions (and call σ an *even* permutation), and $\operatorname{sign}(\sigma) = -1$ if the number is odd (and call σ an *odd* permutation). Then

$$\operatorname{sign} : S_n \longrightarrow \{\pm 1\}$$

is a group homomorphism. Its kernel is denoted A_n and called the *alternating group* (on n objects). It consists of all the even permutations.

As soon as $n \geq 2$, there exist odd permutations (take a transposition), hence the sign homomorphism is surjective. Therefore, A_n is a normal subgroup of index 2 in S_n when $n \geq 2$.

Are there other normal subgroups of S_n? Well, here is one example. Consider the case $n = 4$. Then

$$V = \{(), (12)(34), (13)(24), (14)(23)\} \subset S_4$$

is, as you can easily check, a normal subgroup. It is called the *Klein four group* ("Vierergruppe" in German, whence the V) and is (up to isomorphism) the only noncyclic group of order 4.

On the other hand, you will be hard put to find other examples: For $n \geq 5$, the only nontrivial normal subgroup of S_n is A_n (and this is also true for $n = 3$; for $n \leq 2$, A_n is the trivial group). In fact, a somewhat stronger statement is true:

A_n is a simple group when $n \geq 5$.

The proof is not hard (you can find it in every Algebra textbook), but would cost us too much time.

In fact, the smallest nonabelian simple group is A_5 of order 60.

4.15. **Remark.** The groups A_4, S_4 and A_5 are important: they are the symmetry groups of the platonic solids (i.e., the groups of rotations preserving the set of vertices (say) of a platonic solid). The tetrahedron has symmetry group A_4 (acting on the vertices or faces), the cube and octahedron have symmetry group S_4 (acting for example on the four long diagonals of the cube), and the dodecahedron and icosahedron have symmetry group A_5 (acting on the five tetrahedra that are inscribed into a dodecahedron).

4.16. **Example.** We learned in Linear Algebra that the determinant is multiplicative (and therefore does not vanish on invertible matrices). This means that we have a group homomomorphism

$$\det : \mathrm{GL}_n(F) \longrightarrow F^\times$$

(for any field F; $F^\times$ denotes the multiplicative group $(F \setminus \{0\}, \cdot)$). Its kernel is the normal subgroup

$$\mathrm{SL}_n(F) = \{A \in \mathrm{GL}_n(F) : \det(A) = 1\},$$

the *special linear group.*

For $n \geq 1$, the determinant homomorphism is surjective, hence we get that

$$\mathrm{GL}_n(F)/\,\mathrm{SL}_n(F) \cong \mathbb{F}^\times .$$

But now back to some theorems!

4.17. **Theorem.** *Let H and K be normal subgroups of G such that $K \subset H$. Then there is a natural surjective group homomorphism $G/K \to G/H$. Its kernel is H/K. In particular, we have the isomorphism*

$$(G/K)\,/\,(H/K) \cong G/H .$$

Proof. The homomorphism sends gK to gH. Since $K \subset H$, this is well-defined (if $k \in K$ and $g \in G$, then $gk \in gH$). It is obviously surjective, and the kernel consists of the classes gK such that $gH = H$, i.e., such that $g \in H$. This is exactly the subset H/K of G/K. The last statement follows from Thm. 4.8. □

4.18. **Proposition.** *If $\phi : G \to G'$ is a group homomorphism and $H \lhd G$ is a normal subgroup contained in $\ker \phi$, then there is a unique group homomorphism $\phi' : G/H \to \phi$ making the following diagram commutative.*

$$\begin{tikzcd}
G \arrow[rr, "\phi"] \arrow[dr] & & G' \\
& G/H \arrow[ur, "\phi'"'] &
\end{tikzcd}$$

Proof. Uniqueness is clear: we need to have $\phi'(gH) = \phi(g)$. To see existence, note that we have maps as in the following diagram.

$$\begin{tikzcd}
G \arrow[r, "\phi"] \arrow[d] & G' \\
G/H \arrow[r] \arrow[ur, dashed] & G/\ker\phi \arrow[u]
\end{tikzcd}$$

The dashed arrow does what we want. □

4.19. **Theorem.** *Let $H \lhd G$ be a normal subgroup and $K \leq G$ a subgroup. Then KH is a subgroup of G and has H as a normal subgroup, and the canonical homomorphism $K \to KH/H$, $k \mapsto kH$, is surjective with kernel $K \cap H$. In particular, we have the isomorphism*

$$KH/H \cong K/(H \cap K)\,.$$

Proof. First note that $kH = Hk$ for all $k \in K$ (since H is normal), so $KH = HK$, and by Prop. 3.10, $KH \leq G$. Also, it is clear that $H \lhd KH$.

We then have a canonical homomorphism as described; it is the composition $K \to KH \to KH/H$. It is surjective, since the general element of KH has the form kh with $k \in K$ and $h \in H$, and so $khH = kH$ is in the image. Finally, the kernel consists of those $k \in K$ that satisfy $kH = H$, which means $k \in H$. □

5. Direct Products

There are several ways to construct new groups out of given ones. One of the most important among these is the *direct product.*

5.1. **Definition.** Let G_1 and G_2 be two groups. Then $G_1 \times G_2$ with "componentwise" binary operation $(g_1, g_2) \cdot (g_1', g_2') = (g_1 g_1', g_2 g_2')$ is a group, called the *direct product* of G_1 and G_2.

There are canonical injective group homomorphisms

$$\begin{aligned} i_1 : G_1 &\longrightarrow G_1 \times G_2\,, \quad g_1 \longmapsto (g_1, e_2) \qquad \text{and} \\ i_2 : G_2 &\longrightarrow G_1 \times G_2\,, \quad g_2 \longmapsto (e_1, g_2) \end{aligned}$$

with commuting images $G_1 \times \{e_2\}$ and $\{e_1\} \times G_2$. There are canonical surjective group homomorphisms

$$\begin{aligned} p_1 : G_1 \times G_2 &\longrightarrow G_1\,, \quad (g_1, g_2) \longmapsto g_1 \qquad \text{and} \\ p_2 : G_1 \times G_2 &\longrightarrow G_2\,, \quad (g_1, g_2) \longmapsto g_2\,. \end{aligned}$$

5.2. **Lemma.** *The direct product has the following universal properties.*

(1) *If H is a group with homomorphisms $\phi_1 : H \to G_1$ and $\phi_2 : H \to G_2$, then there is a unique homomorphism $\Phi : H \to G_1 \times G_2$ such that $p_1 \circ \Phi = \phi_1$ and $p_2 \circ \Phi = \phi_2$.*

$$\begin{tikzcd} & G_1 \\ H \arrow[ur, "\phi_1"] \arrow[r, dashed, "\Phi"] \arrow[dr, "\phi_2"'] & G_1 \times G_2 \arrow[u, "p_1"'] \arrow[d, "p_2"] \\ & G_2 \end{tikzcd}$$

(2) *If H is a group with homomorphisms $\psi_1 : G_1 \to H$ and $\psi_2 : G_2 \to H$ such that their images commute, then there is a unique homomorphism $\Psi : G_1 \times G_2 \to H$ such that $\Psi \circ i_1 = \psi_1$ and $\Psi \circ i_2 = \psi_2$.*

$$\begin{tikzcd} G_1 \arrow[d, "i_1"'] \arrow[dr, "\psi_1"] & \\ G_1 \times G_2 \arrow[r, dashed, "\Psi"] & H \\ G_2 \arrow[u, "i_2"] \arrow[ur, "\psi_2"'] & \end{tikzcd}$$

Proof.

(1) As usual, uniqueness is clear — we have to define $\Phi(h) = (\phi_1(h), \phi_2(h))$. It is then easy to check that this is indeed a group homomorphism.

(2) Here we need to define $\Psi(g_1, g_2) = \psi_1(g_1)\psi_2(g_2)$. This is a group homomorphism, since $\psi_1(G_1)$ and $\psi_2(G_2)$ commute in H:

$$\begin{aligned} \Psi\big((g_1, g_2) \cdot (g_1', g_2')\big) &= \Psi(g_1 g_1', g_2 g_2') = \psi_1(g_1 g_1')\psi_2(g_2 g_2') \\ &= \psi_1(g_1)\psi_1(g_1')\psi_2(g_2)\psi_2(g_2') = \psi_1(g_1)\psi_2(g_2)\psi_1(g_1')\psi_2(g_2') \\ &= \Psi(g_1, g_2)\Psi(g_1', g_2') \end{aligned}$$

□

5.3. **Remark.** This definition and the properties extend naturally to an arbitrary finite number of factors $G_1, \ldots, G_n$. It is easy to see that this direct product is commutative and associative up to natural isomorphism:

$$G_1 \times G_2 \cong G_2 \times G_1 \quad \text{and} \quad (G_1 \times G_2) \times G_3 \cong G_1 \times G_2 \times G_3 \cong G_1 \times (G_2 \times G_3) \,.$$

5.4. **Remark.** The direct product can still be defined for an arbitrary (possibly infinite) family of groups. In general, only the first of the universal properties in Lemma 5.2 will hold. (We would need infinite products in H to define a suitable map for the second property.)

In order to get a construction that satisfies the second universal property, one has to use the *restricted direct product*. Let $(G_i)_{i \in I}$ be a family of groups G_i with identity elements e_i. Then we set

$$\prod_{i \in I}{}' G_i = \{(g_i)_{i \in I} : g_i \in G_i \text{ for all } i \in I \text{ and } g_i = e_i \text{ for almost all } i \in I\}$$

(where "almost all" means "all but finitely many"). This is again a group under componentwise operation, and it satisfies the second universal property above (but in general not the first).

In the context of abelian groups, this is called the *direct sum* $\bigoplus_{i \in I} G_i$.

5.5. **Lemma.** *Let G_1 and G_2 be groups, $g_1 \in G_1$, $g_2 \in G_2$ elements of finite order. Then the order of $(g_1, g_2) \in G_1 \times G_2$ is the least common multiple of $o(g_1)$ and $o(g_2)$.*

Proof. We have

$$\begin{aligned} \{n \in \mathbb{Z} : g_1^n = e_1\} &= o(g_1)\,\mathbb{Z}\,, \\ \{n \in \mathbb{Z} : g_2^n = e_2\} &= o(g_2)\,\mathbb{Z}\,, \quad \text{and therefore} \\ \{n \in \mathbb{Z} : (g_1, g_2)^n = (e_1, e_2)\} &= o(g_1)\,\mathbb{Z} \cap o(g_2)\,\mathbb{Z} = \operatorname{lcm}(o(g_1), o(g_2))\,\mathbb{Z}\,. \end{aligned}$$

□

5.6. **Corollary.** *Let m and n be coprime positive integers. Then $\mathbb{Z}_m \times \mathbb{Z}_n \cong \mathbb{Z}_{mn}$.*

Proof. Consider $(1, 1) \in \mathbb{Z}_m \times \mathbb{Z}_n$. By the preceding lemma, we have $o(1, 1) = \operatorname{lcm}(m, n) = mn = \#(\mathbb{Z}_m \times \mathbb{Z}_n)$, so $\mathbb{Z}_m \times \mathbb{Z}_n$ is cyclic, generated by $(1, 1)$. □

5.7. **Example.** On the other hand, $\mathbb{Z}_2 \times \mathbb{Z}_2$ is not isomorphic to $\mathbb{Z}_4$: in the first group, all elements are of order 1 or 2, whereas the second group has elements of order 4. (We have $\mathbb{Z}_2 \times \mathbb{Z}_2 \cong V$, the Klein four group.)

5.8. **Definition.** Let G be a group and H_1, H_2 subgroups of G. We say that G is the *internal direct product* of H_1 and H_2, if the map

$$H_1 \times H_2 \longrightarrow G\,, \quad (h_1, h_2) \longmapsto h_1 h_2$$

is a group isomorphism.

Concretely, we need the following properties.

(1) H_1 and H_2 commute (i.e., $h_1 h_2 = h_2 h_1$ for all $h_1 \in H_1$, $h_2 \in H_2$). This is equivalent to the statement that the map is a homomorphism.
(2) H_1 and H_2 generate G (i.e., $H_1 H_2 = G$, assuming (1)). This means that the homomorphism is surjective.
(3) $H_1 \cap H_2 = \{e\}$. This means that the homomorphism is injective.

If G is finite, property (2) can be replaced by $\#G = \#H_1 \cdot \#H_2$.

5.9. **Examples.** If m and n are coprime positive integers, then $\mathbb{Z}_{mn}$ is the internal direct product of its subgroups $n\mathbb{Z}_{mn}$ and $m\mathbb{Z}_{mn}$. Indeed, the group is abelian, so the first property is automatic. Also, an integer that is a multiple of both m and n must be a multiple of mn; this implies property (3). Finally, $\mathbb{Z}_{mn}$ is finite and $\#\mathbb{Z}_{mn} = mn = \#(n\mathbb{Z}_{mn})\#(m\mathbb{Z}_{mn})$.

The Klein four group

$$V = \{(), (12)(34), (13)(24), (14)(23)\} \subset S_4$$

is the internal direct product of any two of its subgroups of order 2. Indeed, properties (1) and (3) are clear, and $\#V = 4 = 2 \cdot 2$.

6. Group Actions on Sets

Groups are not only important in their own right, but also because they occur as symmetry groups, automorphism groups and the like, in particular as groups acting on something. In the simplest case, this something is just a set, but in many contexts, the set has some additional structure, which is preserved by the group action. Here, we want to study the basic concept of a group action on a set (and then apply it to a situation where the group acts on itself!).

6.1. **Definition.** Let G be a group and X a set. A *left action* of G on X is given by a map $G \times X \to X$, usually written $(g, x) \mapsto g \cdot x$ (or even just gx), that has the following properties.

(1) $\forall x \in X : e \cdot x = x$;
(2) $\forall g, g' \in G, x \in X : (gg') \cdot x = g \cdot (g' \cdot x)$.

Equivalently, a left action of G on X is given by a group homomorphism $\phi : G \to S(X)$ (recall that $S(X)$ is the group of permutations of X). To see this, observe that the map $G \times X \to X$ induces a homomorphism $\phi : G \to S(X)$ by setting $\phi(g) : x \mapsto g \cdot x$. Note that $\phi(g)$ is really a bijection $X \to X$, since it has the inverse $\phi(g^{-1})$ (by the properties (1) and (2) above). Property (2) then shows that ϕ is a group homomorphism. Conversely, from a homomorphism $\phi : G \to S(X)$, we obtain a map $G \times X \to X$ by setting $g \cdot x = \phi(g)(x)$.

Right actions are defined analogously; here the map is of the type $X \times G \to X$ and written $(x, g) \mapsto x \cdot g$.

6.2. **Definition.** Assume G acts on X from the left. Let $x \in X$. The *stabilizer* of x is

$$G_x = \{g \in G : g \cdot x = x\} \subset G\,.$$

G_x is a subgroup of G.

The *orbit of* x is

$$Gx = G \cdot x = \{g \cdot x : g \in G\} \subset X\,.$$

The set of all orbits is $G\backslash X = \{Gx : x \in X\}$.

If $g \cdot x = x$, then x is called a *fixed point* of g.

6.3. **Remark.** If we consider the homomorphism $\phi : G \to S(X)$ corresponding to the left action, then its kernel is the intersection of all stabilizers:

$$\ker \phi = \bigcap_{x \in X} G_x \,.$$

6.4. **Lemma.** *The orbits form a partition of X: two orbits are either disjoint or equal.*

Proof. The proof is virtually identical to that of Prop. 3.3. □

6.5. **Lemma.** *Let G act on X from the left. Then for every $x \in X$, we have a canonical bijection*

$$G/G_x \longrightarrow G \cdot x \,, \qquad gG_x \longmapsto g \cdot x \,.$$

In particular, we have $\#(G \cdot x) = (G : G_x)$.

Proof. The map is well-defined: if $g' \in G_x$, then $gg' \cdot x = g \cdot (g' \cdot x) = g \cdot x$. The map is surjective by definition of the orbit of x. Finally, assume that $g \cdot x = g' \cdot x$. Then we have $g^{-1}g' \cdot x = x$, so $g^{-1}g' \in G_x$, and therefore $gG_x = g'G_x$. So the map is also injective. □

6.6. **Example.** Let G be a group. Then G acts on the set G from the left by left multiplication: $g \cdot h = gh$. All stabilizers are trivial, and there is just one orbit. We therefore get an embdedding (injetctive homomorphism) of G into $S(G)$. In particular, every finite group of order n can be embedded into S_n.

6.7. **Example.** There is another (left) action of G on the set G. It is given by *conjugation.* We set $g \cdot h = ghg^{-1}$. In this context, the stabilizer of $g \in G$ is called the *centralizer* of g, $C_G(g)$. It is the largest subgroup of G in which g is central, i.e., commutes with all elements. The intersection of all centralizers is the *center* of the group:

$$\bigcap_{g \in G} C_G(g) = Z(G) = \{h \in G : gh = hg \text{ for all } g \in G\}$$

More generally, the *centralizer* of a subset $H \subset G$ is

$$C_G(H) = \{g \in G : gh = hg \text{ for all } h \in H\} = \bigcap_{h \in H} C_G(h) \,.$$

In this sense, we have $Z(G) = C_G(G)$.

This action is not just by permutations, but by automorphisms of G: we get a homomorphism $G \to \operatorname{Aut}(G)$ with kernel $Z(G)$. The elements of its image are called *inner automorphisms;* they are of the form $h \mapsto ghg^{-1}$. One can check that the image is a normal subgroup of $\operatorname{Aut}(G)$, the *inner automorphism group* $\operatorname{Inn}(G) \cong G/Z(G)$. The quotient group $\operatorname{Out}(G) = \operatorname{Aut}(G)/\operatorname{Inn}(G)$ is the *outer automorphism group;* every element of $\operatorname{Aut}(G) \setminus \operatorname{Inn}(G)$ is an *outer automorphism.*

It is a fact that S_n has no outer automorphisms, except when $n = 6$, when $\#\operatorname{Out}(S_6) = 2$. (Bonus Homework Problem)

On the other hand, all nontrivial automorphisms of an abelian group are outer automorphims.

The orbits under this action by conjugation are called *conjugacy classes;* two elements in the same orbit are said to be *conjugate.*

The size of the conjugacy class of $g \in G$ is $(G : C_G(g)) = \#G/\#C_G(g)$ (if G is finite). The class has only one element if and only if $g \in Z(G)$; otherwise its size is a divisor of $\#G$ larger than 1. We get the following identity.

6.8. **Class Formula.**

$$\#G = \#Z(G) + \sum_{g \in R} (G : C_G(g)) \,,$$

where $R \subset G \setminus Z(G)$ is a system of representatives of the conjugacy classes outside the center of G.

This seemingly innocent relation has a nice application.

6.9. **Definition.** Let p be a prime number. A finite group G is called a *p-group* if its order is a power of p: $\#G = p^f$ with $f \ge 1$.

6.10. **Proposition.** *A p-group G has nontrivial center $Z(G)$.*

Proof. In the equation above, we have that p divides $\#G$. Also, for any $g \in R$, the index $(G : C_G(g))$ is larger than 1 and a divisor of $p^f = \#G$, hence p divides $(G : C_G(g))$. Therefore, p must divide $\#Z(G)$, and so $Z(G)$ cannot be trivial. □

6.11. **Example.** G also acts by conjugation on its subgroups: if $H \le G$, then $gHg^{-1} \le G$. In this context, the stabilizer is called the *normalizer* of a subgroup:

$$N_G(H) = \{g \in G : gH = Hg\} \,;$$

it is the largest subgroup K of G such that $H \triangleleft K$. A subgroup H is normal if and only if $N_G(H) = G$, which is the case if and only if its orbit contains just one element. The orbits are again called *conjugacy classes* (of subgroups), subgroups in the same orbit are *conjugate.*

For the size of the orbit of H, we find

$$\#\{gHg^{-1} : g \in G\} = (G : N_G(H)) = \frac{(G : H)}{(N_G(H) : H)} \,,$$

a divisor of $(G : H)$.

In many cases, one is really interested in the set X "modulo" the action of G, which means that one wants to know about the quotient set $G\backslash X$. For example, one would like to know its size.

6.12. **Lemma (Burnside).** *If G and X are finite, then $\#(G\backslash X)$ is the "average number of fixed points":*

$$\#(G\backslash X) = \frac{1}{\#G} \sum_{g \in G} \#\{x \in X : gx = x\}$$

Proof. This is a classical proof by "double counting": we count the same set in two ways. The set is $S = \{(g, x) \in G \times X : gx = x\}$. We count its elements on the one side by first counting elements with fixed g and on the other side by first counting elements with fixed x.

$$\begin{aligned}\sum_{g\in G} \#\{x \in X : gx = x\} = \#S \\ &= \sum_{x\in X} \#\{g \in G : gx = x\} = \sum_{x\in X} \#G_x \\ &= \sum_{x\in X} \frac{\#G}{\#Gx} = \sum_{O\in G\backslash X} \sum_{x\in O} \frac{\#G}{\#O} \\ &= \#G \sum_{O\in G\backslash X} = \#G\,\#(G\backslash X)\end{aligned}$$

□

7. The Sylow Theorems

We will now use the action of G on itself by conjugation to prove a number of nontrivial statements on subgroups of finite groups, known as the Sylow Theorems.

7.1. **Theorem.** *Let G be a finite group, and let p^f be a prime power dividing $\#G$. Then G has a subgroup of order p^f.*

Proof. The proof is by induction on $\#G$. The case $f = 0$ is of course trivial. If G has a proper subgroup H such that $p^f \mid \#H$, then we know by induction that H has a subgroup $K \le H$ of order p^f, but then $K \le G$ as well.

So let us assume that G has no such subgroup H. This implies that for every proper subgroup $H < G$, p divides the index $(G : H)$. Looking at the Class Formula 6.8, we deduce that p divides $\#Z(G)$. By Cauchy's Theorem 3.9, $Z(G)$ has a subgroup K of order p, which then is a normal subgroup of G (since it is contained in the center). Now $\#(G/K) = \#G/p$ is less than $\#G$ and divisible by p^{f-1}, so by induction, G/K has a subgroup H' of order p^{f-1}. But then $H' = H/K$ with $H \le G$ a subgroup of G containing K, and $\#H = p\,\#H' = p^f$. □

This shows in particular that G has a subgroup of order p^t, where p^t is the maximal power of p dividing $\#G$.

7.2. **Definition.** Let G be a finite group, and let p be a prime number dividing $\#G$. A subgroup $H \le G$ is called a *p-Sylow subgroup* if H is a p-group and p does not divide the index $(G : H)$.

In other words, H is a maximal (in size) p-subgroup of G. By the previous result, p-Sylow subgroups always exist.

7.3. **Lemma.** *Let H be a p-Sylow subgroup of G and let $K \le G$ be a p-subgroup such that $K \subset N_G(H)$. Then $K \subset H$.*

Proof. We have $H \lhd N_G(H)$ and $K \le N_G(H)$, so HK is a subgroup of $N_G(H)$. Its order divides $\#H\,\#K$, which is a power of p. On the other hand, $HK \le G$ and so $\#HK \mid \#G$. So $\#HK$ divides the largest power of p dividing $\#G$, which is $\#H$. Since we also have $H \subset HK$, this implies $H = HK$, whence $K \subset H$. □

7.4. **Theorem.** *Let G be a finite group with p-Sylow subgroup H. Let K be any p-subgroup of G. Then K is contained in a conjugate of H. In particular, all p-Sylow subgroups are conjugate.*

Proof. Let $S = \{gHg^{-1} : g \in G\}$ be the set of conjugates of H. We have $\#S = (G : N_G(H))$, which divides $(G : H)$ and therefore is not divisible by p.

The subgroup K acts on S by conjugation. Now every K-orbit in S has size a power of p (because it equals the index of a subgroup of the p-group K). On the other hand, $\#S$ is not divisible by p, so there has to be at least one singleton orbit $\{gHg^{-1}\}$. Let $H' = gHg^{-1}$. Since conjugation by elements of K fixes H', we have $K \subset N_G(H')$, hence by the preceding lemma, $K \subset H'$. (Note that H' is a p-Sylow subgroup).

The last statement follows by taking K to be any p-Sylow subgroup. □

7.5. **Theorem.** *Let G be a finite group, and let p be a prime divisor of $\#G$. Write $\#G = p^t r$ with r not divisible by p. Then the number m_p of p-Sylow subgroups of G divides r, and $m_p \equiv 1 \bmod p$.*

Proof. Let H be a p-Sylow subgroup of G. By the preceding theorem, the p-Sylow subgroups form one orbit under conjugation by elements from G. The size m_p of the orbit equals the index $(G : N_G(H))$, which divides $(G : H) = r$. This proves the first assertion.

For the second assertion, we consider the conjugation action of H on the set of p-Sylow subgroups. We have seen (in the proof of the preceding theorem) that H fixes an element H' of this set if and only if $H \subset H'$, which means here that $H = H'$. This means that there is exactly one orbit of size 1 (namely $\{H\}$), and all the other orbit sizes are multiples of p. The claim follows, since m_p is the sum of all the orbit sizes. □

7.6. **Remark.** A p-Sylow subgroup H of G is normal if and only if $m_p = 1$.

The strength of these theorems lies in the fact that they can provide us with nontrivial normal subgroups (or, failing that, with some fairly strong information on the structure of G). For example, one can use them to prove the following result.

7.7. **Proposition.** *If G is a group of order $1 < \#G < 60$, then G is of prime order, or else G has a nontrivial normal subgroup.*

Proof. We only discuss one exemplary case here. For a complete treatment, see for example Rowen's book.

Let us consider the case $\#G = 56 = 2^3 \cdot 7$. We know that m_7 is $\equiv 1 \bmod 7$ and divides 8, therefore $m_7 = 1$ or $m_7 = 8$. In the first case, G has a normal subgroup of order 7 (the 7-Sylow subgroup). In the second case, we have 8 subgroups of order 7, any two of which only meet in the identity element (this is because they are of prime order and so any nontrivial element generates the subgroup). So me must have $8 \cdot (7 - 1) = 48$ elements of order 7 in G. This leaves exactly 8 elements of order a power of 2, which is just sufficient to make up one 2-Sylow subgroup. Hence $m_2 = 1$, and there is a normal subgroup of order 8. □

7.8. **Corollary.** *A_5 is the smallest nonabelian simple group.*

8. Rings — Basic Definitions

We will now discuss structures with two binary operations, commonly called addition and multiplication.

8.1. **Definition.** A *ring* is a set R with two binary operations $+$ and $\cdot$ and elements $0, 1 \in R$ such that $(R, +, 0)$ is an abelian group and $(R, \cdot, 1)$ is a monoid, and such that the distributive laws hold:

$$\forall a, b, c \in R : \quad a(b + c) = ab + ac \quad \text{and} \quad (a + b)c = ac + bc\,.$$

The ring R is called *commutative* if $(R, \cdot)$ is commutative.

As you will recall, a ring is called a *skew field* or *division ring* if every nonzero element has a multiplicative inverse (i.e., if $(R \setminus \{0\}, \cdot)$ is a group). If the division ring is also commutative, it is called a *field.*

Here are a few elementary properties.

8.2. **Lemma.** *Let R be a ring. Then*

(1) $\forall a \in R : a \cdot 0 = 0 \cdot a = 0.$
(2) $\forall a \in R : a \cdot (-1) = (-1) \cdot a = -a.$

Proof. We have $a \cdot 0 = a(0 + 0) = a \cdot 0 + a \cdot 0$, hence (subtracting $a \cdot 0$ from both sides) $0 = a \cdot 0$ and similarly $0 = 0 \cdot a$. For the second statement, note that $0 = a \cdot 0 = a(1 + (-1)) = a \cdot 1 + a \cdot (-1)$. Subtracting a from both sides gives $-a = a \cdot (-1)$. Similarly for $(-1) \cdot a$. □

Note that this implies rules like $(-a)b = -(ab)$, $(-a)(-b) = ab$ and so on.

8.3. **Remark.** In the definition above, it is not required that 0 and 1 are distinct. However, if $0 = 1$, then for all $a \in R$, we have $a = 1 \cdot a = 0 \cdot a = 0$, so $R = \{0\}$ is the *trivial ring.* In all other cases, 0 and 1 are two distinct elements of R.

8.4. **Examples.** The prototypical example of a commutative ring is the ring of integers $\mathbb{Z}$. But also the sets $\mathbb{Z}_n$ with addition and multiplication "mod n" are commutative rings (and we will soon recognize them as quotient rings $\mathbb{Z}/n\mathbb{Z}$).

Any field (like $\mathbb{Q}, \mathbb{R}, \mathbb{C}$) is a commutative ring.

Another important example of a commutative ring is the *polynomial ring* $R[x]$, where R is a commutative ring (for example, a field). We will discuss this in some detail later.

To see an example of a non-commutative ring, consider the $n \times n$ matrices $\mathrm{Mat}(n, R)$ over some ring R (for example, a field). This forms a ring with the usual addition and multiplication of matrices; it is not commutative when $n \geq 2$. It is also not a division ring (again when $n \geq 2$).

To see a "real" skew field, recall the *quaternions* $\mathbb{H} = \mathbb{R} + \mathbb{R}i + \mathbb{R}j + \mathbb{R}k$ with (associative) multiplication satisfying $i^2 = j^2 = k^2 = ijk = -1$. It can be shown that this is a division ring, but it is not commutative, since for example $ij = k \neq -k = ji$.

It is especially nice when we can multiplicatively cancel nonzero elements. This is equivalent to the absence of *zero divisors.*

8.5. **Lemma.** *Let R be a ring. Then the following statements are equivalent.*

(1) $\forall a, b, c \in R : a \neq 0$ *and* $ab = ac \implies b = c$.
(2) $\forall a, b \in R : ab = 0 \implies a = 0$ *or* $b = 0$.

Proof. To see that (1) implies (2), set $c = 0$ in (1). To see that (2) implies (1), note that $ab = ac$ implies $a(b - c) = 0$, so $b - c = 0$ by (2). □

A nonzero element $a \in R$ such that there is a nonzero $b \in R$ with $ab = 0$ is called a *zero divisor*. We can state the lemma in the form *"A ring R does not have zero divisors if and only if $(R \setminus \{0\}, \cdot)$ is a cancellative monoid."*

8.6. **Definition.** A nontrivial ring R having the properties given in Lemma 8.5 is called a *domain*. If R is also commutative, it is called an *integral domain*.

8.7. **Examples.** Obviously, $\mathbb{Z}$ is an integral domain (and that is where the name comes from).

Any division ring is a domain, any field is an integral domain.

A matrix ring $\text{Mat}(n, R)$ for $n \geq 2$ is never a domain.

A finite ring $\mathbb{Z}_n$ is an integral domain if and only if n is a prime number: $m \in \mathbb{Z}_n$ is a zero divisor if and only if $\gcd(m, n) > 1$ (and n does not divide m), hence zero divisors exist if and only if n has a nontrivial divisor (excluding $n = 1$ from the discussion when $\mathbb{Z}_n$ is the trivial rnig). In fact, more is true, as the following result shows.

8.8. **Proposition.** *If R is a finite integral domain, then R is a field.*

Proof. By assumption, $(R \setminus \{0\}, \cdot)$ is a cancellative monoid. Theorem 2.6 then tells us that $(R \setminus \{0\}, \cdot)$ is already a group, so R is a field. □

In particular, $\mathbb{Z}_p$ is a field for p a prime number. This field is usually denoted $\mathbb{F}_p$.

8.9. **Remark.** In the same way, one proves that a finite domain is a skew field. Now there is another (deep) result that says that a finite skew field is already a field (i.e., commutative). Hence any finite domain is already a field.

However, not every finite ring is commutative; counterexamples are given by the matrix rings $\text{Mat}(n, \mathbb{F}_p)$ over finite fields ($n \geq 2$).

As usual, we can define substructures.

8.10. **Definition.** Let R be a ring, $S \subset R$ a subset. We call S a *subring* of R, if S is an additive subgroup of R and a multiplicative submonoid of R (in particular, $1 \in S$).

It can be checked that S with the binary operations coming from R is then a ring (Exercise).

Any intersection of subrings is again a subring. Hence it is possible to talk of the subring *generated* by a subset, in the same way as we did for subgroups.

8.11. **Examples.** The only subring of $\mathbb{Z}$ is $\mathbb{Z}$ itself.

Any subring of an (integral) domain is again an (integral domain). In particular, any subring of a field is an integral domain. As a concrete example, consider the ring of *dyadic fractions*, consisting of all rational numbers whose denominator is a power of 2.

8.12. **Definition.** Let R be a ring. The *unit group* $R^\times$ is the group of invertible elements of the monoid $(R, \cdot)$.

8.13. **Examples.** We have $\mathbb{Z}^\times = \{\pm 1\}$ and $\mathbb{Z}_n^\times = \{a \in \mathbb{Z}_n : \gcd(a,n) = 1\}$.

If F is a skew field, then $F^\times = F \setminus \{0\}$.

9. Left, Right and Two-Sided Ideals and Homomorphisms

Contrary to the case of groups, there is another type of subobject that is relevant when studying rings. Later (when talking about modules) we will see that it corresponds to submodules of R as a module over itself.

9.1. **Definition.** Let R be a ring. A subset $I \subset R$ is a *left ideal* of R if I is an additive subgroup of R such that $\forall r \in R, a \in I : ra \in I$. Similarly, an additive subgroup $I \subset R$ is a *right ideal* if $\forall r \in R, a \in I : ar \in I$. Finally, I is a *two-sided ideal* (or just *ideal*) of R if I is both a left and right ideal.

Note that for a commutative ring R, there is no difference between left, right and two-sided ideals, and so one just talks about ideals.

Here are some basic properties.

9.2. **Lemma.** *Let R be a ring.*

(1) *Any intersection of left/right/two-sided ideals of R is again a left/right/two-sided ideal of R.*
(2) *The sum $I + J = \{a + b : a \in I, b \in J\}$ of two left/right/two-sided ideals I and J of R is again a left/right/two-sided ideal of R. It is the smallest left/right/two-sided ideal of R containing both I and J.*
(3) *If $I_1 \subset I_2 \subset I_3 \subset \ldots$ is an increasing sequence of left/right/two-sided ideals of R, then their union $I = \bigcup_n I_n$ is again a left/right/two-sided ideal of R.*

Proof.

(1) This is easy (Exercise).
(2) $I + J$ is certainly an additive subgroup and the smallest additive subgroup containing I and J. One only has to check the additional property of ideals. For example, if $a \in I$, $b \in J$, $r \in R$ and I, J are left ideals, then $r(a + b) = ra + rb \in I + J$.
(3) This is true because all properties that one has to check only involve finitely many elements and therefore can be verified in one of the I_n.

□

Thus it makes again sense to talk about the left/right/two-sided ideal generated by a subset of R.

9.3. **Examples.** If R is a ring and $a \in R$, then $Ra = \{ra : r \in R\}$ is a left ideal. In fact, it is the smallest left ideal containing a. We call it the *principal left ideal* generated by a. Similarly, aR is the smallest right ideal containing a. The smallest two-sided ideal containing a is more difficult to describe (Exercise).

The ideals of $\mathbb{Z}$ are all of the form $\mathbb{Z}n$ (since $\mathbb{Z}$ is commutative, left, right and two-sided ideals all coincide).

Any ring R has the *trivial ideals* $0 := \{0\}$ and R. R is a division ring if and only if these are the only left (or right) ideals, and $R \neq 0$. (In fact, if $a \in R$ is invertible, then $Ra = R$, and vice versa. So if R is a division ring and I is a nonzero left ideal, then there is some $0 \neq a \in I$. But a is invertible, so $R = Ra \subset I$. Conversely, assume there are no nontrivial ideals and pick $0 \neq a \in R$. Then $0 \neq Ra$, so $Ra = R$ and a is invertible.)

It can be shown that the matrix ring $\mathrm{Mat}(n, F)$ over a (skew) field F has no nontrivial two-sided ideals, even though it is not a division ring (for $n \geq 2$). However, it has many nontrivial left ideals, for example the matrices whose only nonzero entries are in the first column.

As always, we want to consider structure-preserving maps between rings.

9.4. **Definition.** A *ring homomorphism* between two rings R and R' is a map $\phi : R \to R'$ such that ϕ is a homomorphism of the additive groups $(R, +)$ and $(R', +)$ and a monoid homomorphism of $(R, \cdot)$ and $(R', \cdot)$. Concretely, we require that

$$\phi(a + b) = \phi(a) + \phi(b)\,, \quad \phi(1) = 1\,, \quad \phi(ab) = \phi(a)\phi(b)$$

for all $a, b \in R$.

ϕ is called an *endomorphism* of R if $R = R'$.

ϕ is an *isomorphism* if it is bijective; in this case, ϕ^{-1} is again a ring homomorphism (Exercise), and R and R' are called *isomorphic*, $R \cong R'$. An endomorphism of R that is an isomorphism is called an *automorphism* of R.

9.5. **Definition.** Let $\phi : R \to R'$ be a ring homomorphism. The *kernel* of ϕ is the kernel of ϕ as a homomorphism of additive groups:

$$\ker \phi = \{r \in R : \phi(r) = 0\}\,.$$

9.6. **Lemma.** *Let $\phi : R \to R'$ be a ring homomorphism. Then $\ker \phi$ is a two-sided ideal of R.*

Proof. Let $I = \ker \phi$. Since I is the kernel of a homomorphism of additive groups $(R, +) \to (R', +)$, it is certainly an additive subgroup of R. Now let $a \in I$ and $r \in R$. Then

$$\phi(ar) = \phi(a)\phi(r) = 0 \cdot \phi(r) = 0$$

and similarly $\phi(ra) = 0$, so $ar, ra \in I$. □

This raises, of course, the question if any given two-sided ideal can be the kernel of a homomorphism.

9.7. **Proposition.** *Let R be a ring and $I \subset R$ a two-sided ideal. On the quotient R/I of additive groups there is a unique multiplication turning R/I into a ring such that the natural map*

$$\phi : R \longrightarrow R/I\,, \quad r \longmapsto r + I$$

is a ring homomorphism. We then have $\ker \phi = I$.

The ring R/I constructed in this way is called the *quotient ring* of R modulo I.

Proof. If ϕ is to be a ring homomorphism, then we must define

$$(r + I)(r' + I) = \phi(r)\phi(r') = \phi(rr') = rr' + I\,.$$

We need to check that this makes sense. So pick $a, a' \in I$, then $r + a$, $r' + a'$ are alternative representatives of $r + I$ and $r' + I$, respectively. We must verify that $(r + a)(r' + a') \in rr' + I$. But we have

$$(r + a)(r' + a') = rr' + ra' + a(r' + a') \in rr' + RI + IR = rr' + I\,.$$

It is then clear that this multiplication turns R/I into a ring with unit element $1 + I$, and that ϕ is a ring homomorphism.

Finally, $\ker \phi = \{r \in R : r + I = I\} = I$. □

We have the usual kind of basic properties of ring homomorphisms.

9.8. **Lemma.** *Let $\phi : R \to R'$ be a ring homomorphism.*

1. *If $I \subset R$ is a left/right/two-sided ideal and ϕ is surjective, then $\phi(I) \subset R'$ is a left/right/two-sided ideal.*
2. *If $I' \subset R'$ is a left/right/two-sided ideal, then $\phi^{-1}(I') \subset R$ is a left/right/two-sided ideal.*
3. *If $S \subset R$ is a subring, then $\phi(S) \subset R'$ is a subring. In particular, the image of ϕ is a subring of R'.*
4. *If $S' \subset R'$ is a subring, then $\phi^{-1}(S') \subset R$ is a subring.*
5. *If $\phi' : R' \to R''$ is another ring homomorphism, then $\phi' \circ \phi : R \to R''$ is also a ring homomorphism.*
6. *ϕ is injective if and only if* $\ker \phi = 0$.

Proof. Easy. □

We now have isomorphism theorems analogous to the case of groups.

9.9. **Theorem.** *Let $\phi : R \to R'$ be a ring homomorphism. Let $I = \ker \phi$ be its kernel. Then the natural map*

$$\psi : R/I \longrightarrow R'\,, \quad r + I \longmapsto \phi(r)$$

is an injective ring isomorphism, which is an isomorphism if ϕ is surjective. In particular, we have $R/I \cong \phi(R)$.

Proof. The map is well-defined: if $r - r' \in I$, then $\phi(r) - \phi(r') = 0$. It is clear (by definition of the ring structure on R/I) that ψ is a ring homomorphism. It is injective, since its kernel is $\{I\}$, which is the zero ideal in R/I. Since $\psi(R/I) = \phi(R)$, ψ is surjective if ϕ is surjective, and then it is an isomorphism. □

In addition, we get a bijection between the set of left/right/two-sided ideals of R/I and the set of left/right/two-sided ideals of R containing I by taking images and preimages under ϕ.

9.10. **Theorem.** *Let $I \subset J \subset R$ be two two-sided ideals of R. Then there is a natural surjective ring homomorphism $R/I \to R/J$, with kernel J/I. In particular, we have the isomorphism*

$$(R/I) \Big/ (J/I) \cong R/J\,.$$

Proof. Same as for groups, compare Thm. 4.17. □

9.11. **Proposition.** *Let $\phi : R \to R'$ be a ring homomorphism, and let $I \subset R$ be a two-sided ideal contained in* $\ker\phi$. *Then there is a unique ring homomorphism $\phi' : R/I \to R'$ making the following diagram commutative.*

$$\begin{tikzcd}
R \arrow[rr, "\phi"] \arrow[dr] & & R' \\
& R/I \arrow[ur, "\phi'"'] &
\end{tikzcd}$$

Proof. Same as for groups, compare Prop. 4.18. □

9.12. **Example.** There is a unique ring homomorphism $\mathbb{Z} \to \mathbb{Z}_n$ (sending a to $a \bmod n$). Its kernel is $n\mathbb{Z}$, and it is clearly surjecive. So we get that $\mathbb{Z}_n \cong \mathbb{Z}/n\mathbb{Z}$ also as rings.

Now consider $\mathbb{Z}_n$ and $\mathbb{Z}_{mn}$. Then there is a unique ring homomorphism $\mathbb{Z}_{mn} \to \mathbb{Z}_n$ (which has to send 1 to 1). To see this, consider the ideals $mn\mathbb{Z} \subset n\mathbb{Z} \subset \mathbb{Z}$ and apply Prop. 9.11 above.

9.13. **Definition.** Let R be a *commutative* ring. An ideal $I \subset R$ is called *maximal* if I is a proper ideal, and there is no ideal strictly between I and R.

9.14. **Proposition.** *Let $\phi : R \to R'$ be a surjective homomorphism of commutative rings. Then R' is a field if and only if* $\ker\phi$ *is a maximal ideal of R.*

Proof. R' is a field if and only if the only ideals of R' are 0 and R'. We know that the ideals of R' are in inclusion-preserving 1-to-1 correspondence with the ideals of R containing $\ker\phi$. So R' is a field if and only if there are no ideals strictly between $\ker\phi$ and R, which means exactly that $\ker\phi$ is a maximal ideal. □

9.15. **Example.** We know that all ideals of $\mathbb{Z}$ are of the form $n\mathbb{Z}$. Also, we have that $n\mathbb{Z} \subset m\mathbb{Z}$ if and only if m divides n. Therefore $n\mathbb{Z}$ is maximal if and only if $n \neq 1$ and n does not have nontrivial divisors, which means that n is a prime number. So we see again that $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is a prime number.

9.16. **Example.** Consider the ring $R = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ continuous}\}$ (under point-wise addition and multiplication). Let $a \in \mathbb{R}$, then

$$e_a : R \longrightarrow \mathbb{R}\,, \quad f \longmapsto f(a)$$

is a surjective ring homomorphism onto a field. Hence

$$\ker e_a = \{f \in R : f(a) = 0\}$$

is a maximal ideal of R.

9.17. **Remark.** Using *Zorn's Lemma* (a statement equivalent to the axiom of choice), one can prove quite generally that if $I \subset R$ is a proper ideal in a commutative ring R, then there is a maximal ideal M of R containing I.

9.18. **Digression.** Zorn's Lemma is the following statement.

Let $(X, \leq)$ be a non-empty partially odered set such that every non-empty chain (i.e., totally ordered subset) in X has an upper bound in X. Then X has maximal elements.

It can be shown that this statement is equivalent (modulo the more "harmless" axioms of set theory) to the *Axiom of Choice:*

Let $(X_i)_{i \in I}$ be a family of non-empty sets. Then there exists a map $f : I \to \bigcup_i X_i$ such that $f(i) \in X_i$ for all $i \in I$. Equivalently, the product $\prod_i X_i$ is non-empty.

In the present case, we take X to be the set of all proper ideals of R containing I, ordered by inclusion. Then $I \in X$, so X is non-empty. If $\{I_j : j \in J\}$ is a non-empty chain in X, then $U = \bigcup_{j \in J} I_j$ is an ideal containing I, and U is proper, since $1 \notin U$ (otherwise, $1 \in I_j$ for some j, contradiction). So U is an upper bound for the chain, and Zorn's Lemma applies. We conclude that there is a maximal ideal containing I.

Similarly, one can prove that every vector space V has a basis, by taking X to be the set of all linearly independent subsets of V, again ordered by inclusion. X is nonempty, since $\emptyset \in X$, and the union of a chain of linearly independent sets is again linearly independent. If a maximal linearly independent set is not a basis, then it can be enlarged by including an element outside its span, which gives a contradiction.

10. Products of Rings and Chinese Remainder Theorem

In the same way as for groups, we can construct direct products of rings.

10.1. **Lemma and Definition.** Let R_i, for $i \in I$, be rings. Then we can turn the product $\prod_{i \in I} R_i$ of additive groups into a ring by defining multiplication component-wise:

$$(r_i)_{i \in I} \cdot (s_i)_{i \in I} = (r_i s_i)_{i \in I} .$$

The ring $R = \prod_{i \in I} R_i$ is called the *(direct) product* of the rings R_i. Its zero element is (0_{R_i}), its unit element is (1_{R_i}).

R comes with natural *projection* homomorphisms $\pi_i : R \to R_i$, given by sending $(r_i)_{i \in I}$ to its ith component r_i.

There is again a "universal property" regarding homomorphisms into a product of rings.

10.2. **Proposition.** *Let $(R_i)_{i \in I}$ be a family of rings, $R = \prod_i R_i$ their product. Let S be another ring, and let $\phi_i : S \to R_i$ be ring homomorphisms. Then there is a unique ring homomorphism $\Phi : S \to R$ such that $\pi_i \circ \Phi = \phi_i$ for all $i \in I$. The kernel of Φ is $\ker \Phi = \bigcap_i \ker \phi_i$.*

Proof. If $\pi_i \circ \Phi = \phi_i$ for all $i \in I$, then we have to set

$$\Phi(s) = (\phi_i(s))_{i \in I} \in R .$$

It is then easy to check that this is indeed a ring homomorphism. The statement on the kernel is then clear. □

10.3. **Remark.** If the rings R_i are all nontrivial and I has at least two elements, then R has zero divisors (for example, (r_i) with $r_{i_0} = 1$, $r_i = 0$ for all $i \in I \setminus \{i_0\}$) and so never is a domain.

Let $I = \{1, 2, \dots, n\}$ be finite. Then the elements

$$e_j = (0, \dots, 0, 1, 0, \dots, 0)$$

(where the 1 is in the jth position) are *idempotents* of $R = R_1 \times \cdots \times R_n$: they satisfy $e_j^2 = e_j$. They even form a system of *orthogonal idempotents* — we have $e_i e_j = 0$ for $i \neq j$ and $e_1 + \cdots + e_n = 1_R$.

Conversely, given a *commutative* ring R with a system of orthogonal idempotents $e_1, \dots, e_n$, the subgroups $R_i = Re_i$ are rings with the induced multiplication and unit e_i (they are not in general subrings, however, since they do not contain the unit of R). There are ring homomorphisms $\phi_i : R \to R_i$, $r \mapsto re_i$, hence by the preceding proposition, there is a unique ring homomorphism $\Phi : R \to R_1 \times \cdots \times R_n$, $r \mapsto (re_1, \dots, re_n)$. On the other hand, the homomorphism of additive groups $\Psi : R_1 \times \cdots \times R_n \to R$, $(r_1, \dots, r_n) \mapsto r_1 + \cdots + r_n$, is a ring homomorphism (this follows from the fact that the e_i are orthogonal idempotents), which is inverse to Φ. Hence $R \cong R_1 \times \cdots \times R_n$.

10.4. **Definition.** Two ideals I and J of a ring R are called *comaximal* or *coprime* if $I + J = R$.

10.5. **Theorem.** *Let $I_1, \dots, I_n$ be two-sided ideals of a ring R that are comaximal in pairs. Let $I = I_1 \cap \cdots \cap I_n$ be their intersection. Then the natural ring homomorphism*

$$\phi : R/I \to R/I_1 \times \cdots \times R/I_n$$

is an isomorphism.

Proof. We have the natural homomorphisms $\phi_j : R \to R/I_j$. Hence there is a homomorphism $\phi' : R \to R/I_1 \times \cdots \times R/I_n$; its kernel is $I_1 \cap \cdots \cap I_n = I$. By Thm. 9.9, this induces an injective ring homomorphism

$$\phi : R/I \to R/I_1 \times \cdots \times R/I_n \,.$$

We only need to show that ϕ is surjective. This is where we need that $I_i + I_j = R$ for $i \neq j$. Let $e_j \in R/I_1 \times \cdots \times R/I_n$ be the element that is zero everywhere except at the jth position, where it is 1. It suffices to show that all e_j are in the image (if $e_j = \phi(s_j + I)$, then $(r_1 + I_1, \dots, r_n + I_n) = \phi(r_1 s_1 + \cdots + r_n s_n + I)$). Without loss of generality, take $j = 1$. Since $I_1 + I_j = R$ for all $j = 2, \dots, n$, there are $a_j \in I_1$ and $b_j \in I_j$ such that $a_j + b_j = 1$. Let $s = b_2 \cdots b_n$. Then

$$s + I_1 = (1 - a_2) \dots (1 - a_n) + I_1 = 1 + I_1$$

and $s \in I_j$ for all $j = 2, \dots, n$, so $\phi(s + I) = \phi'(s) = e_1$. □

10.6. **Example.** In $\mathbb{Z}$, two ideals $n\mathbb{Z}$ and $m\mathbb{Z}$ are comaximal if and only if n and m are coprime (since $n\mathbb{Z} + m\mathbb{Z} = d\mathbb{Z}$ is the smallest ideal containing both $n\mathbb{Z}$ and $m\mathbb{Z}$, d is the largest divisor of both n and m). Applying the preceding theorem to $\mathbb{Z}$, we therefore get the standard **Chinese Remainder Theorem**:

Let $m_1, \dots, m_n$ be positive integers, coprime in pairs. Let $a_1, \dots, a_n \in \mathbb{Z}$. Then there is an integer $a \in \mathbb{Z}$ such that $a \equiv a_j \bmod m_j$ for all $j = 1, \dots, n$, and a is uniquely determined modulo $m = m_1 \cdots m_n$.

Note that $m_1\mathbb{Z} \cap \cdots \cap m_n\mathbb{Z} = m_1 \cdots m_n\mathbb{Z}$, since $m_1 \cdots m_n$ is the least common multiple of $m_1, \dots, m_n$ when these numbers are coprime in pairs.

11. The Field of Fractions

In the same way as the integral domain $\mathbb{Z}$ can be embedded into the field $\mathbb{Q}$, any integral domain can be embedded into a field. (Recall that an *integral domain* is a commutative ring without zero divisors.)

11.1. **Theorem.** *Let R be an integral domain, and let $S = R \setminus \{0\}$ be its multiplicative monoid of nonzero elements. Then there is a field $F = R_S = R[S^{-1}]$ and an injective ring homomorphism $\phi : R \to R_S$ such that for every ring homomorphism $\psi : R \to R'$ such that $\psi(s)$ is invertible for every $s \in S$, there is a unique ring homomorphism $\psi' : R_S \to R'$ such that $\psi = \psi' \circ \phi$.*

$$\begin{tikzcd}
 & R_S \arrow[dd, "\psi'"] \\
R \arrow[ur, "\phi"] \arrow[dr, "\psi"'] & \\
 & R'
\end{tikzcd}$$

F as above is called the *field of fractions* of R and denoted $\operatorname{Frac}(R)$. For example, $\operatorname{Frac}(\mathbb{Z}) = \mathbb{Q}$.

Proof. The construction is virtually identical to the construction of $\mathbb{Q}$ from $\mathbb{Z}$. We first define a relation on $R \times S$ via

$$(r, s) \sim (r', s') \iff rs' = r's$$

(think of fractions $\frac{r}{s} = \frac{r'}{s'}$). There is a number of things to check.

- $\sim$ is an equivalence relation.
  It is clear that $\sim$ is reflexive and symmetric. To see that $\sim$ is transitive, assume that $(r, s) \sim (r', s')$ and $(r', s') \sim (r'', s'')$. Then $rs' = r's$ and $r's'' = r''s'$. Multiplying the first equality by s'' and the second by s, we obtain $rs's'' = r'ss'' = r''s's$. Since $s' \neq 0$ and R is an integral domain, we can cancel s' from both sides to obtain $rs'' = r''s$.
- Let $R_S = R \times S/\sim$ be the set of equivalence classes, and write (as usual) $\frac{r}{s}$ for the equivalence class of the pair (r, s).
- Define addition and multiplication on R_S by
$$\frac{r}{s} + \frac{r'}{s'} = \frac{rs' + r's}{ss'}, \qquad \frac{r}{s} \cdot \frac{r'}{s'} = \frac{rr'}{ss'}$$
  and check that these are well-defined (do not depend on the representatives chosen) and make R_S into a ring (with zero $0 = \frac{0}{1}$ and unit $1 = \frac{1}{1}$). This is not hard, if a bit tedious.
- Define $\phi : R \to R_S$, $r \mapsto \frac{r}{1}$ and check that ϕ is an injective ring homomorphism (easy).
- Check that R_S is a field: For $\frac{r}{s} \neq 0$, we have $r \neq 0$, so $\frac{s}{r}$ makes sense, and
$$\frac{r}{s} \cdot \frac{s}{r} = \frac{rs}{sr} = 1\,.$$
- Check the universal property. Given a ring homomorphism $\psi : R \to R'$ such that $\psi(s) \in (R')^\times$ for all $s \in S$, we need to define
$$\psi'\Big(\frac{r}{s}\Big) = \psi'(\phi(r)\phi(s)^{-1}) = \psi'(\phi(r)) \cdot \psi'(\phi(s))^{-1} = \psi(r)\psi(s)^{-1}\,.$$

It is then easy to check that this is indeed a well-defined ring homomorphism.

□

Now, most arguments in the proof of this theorem also go through under somewhat weaker assumptions, if we just want to embed R into a ring in which all elements from a suitable multiplicative submonoid become invertible.

We state the result, without repeating the proof (which is identical to the previous one).

11.2. **Proposition.** *Let R be a commutative ring, and let $S \subset R$ be a multiplicative submonoid such that for all $r \in R$, $s \in S$, $rs = 0 \implies r = 0$. Then there exists a ring R_S and an injective ring homomorphism $\phi : R \to R_S$ such that $\phi(s) \in R_S^\times$ for all $s \in S$, and such that for every ring homomorphism $\psi : R \to R'$ with $\psi(s) \in (R')^\times$ for all $s \in S$, there is a unique ring homomomorphism $\psi' : R_S \to R'$ such that $\psi = \psi' \circ \phi$.*

The ring R_S (or $R[S^{-1}]$) is called the *localization* of R at S. (This strange name comes from algebraic geometry, where rings are used to describe geometric objects, and this operation corresponds to looking at some sort of "neighborhood" of a subobject related to S.)

11.3. **Example.** As an example, look at $\mathbb{Z}$. For every subset P of the set of prime numbers $\{2, 3, 5, 7, \dots\}$, there is a multiplicative submonoid S_P of $\mathbb{Z}$ consisting of the integers whose prime divisors are all from the set P. Then we get a subring $\mathbb{Z}_{S_P} \subset \mathbb{Q}$ of the field of rational numbers. Conversely, if $R \subset \mathbb{Q}$ is a subring, then let P be the set of prime numbers dividing the denominator of some element of R (when written in lowest terms); we then have $R = \mathbb{Z}_{S_P}$. Therefore the subrings of $\mathbb{Q}$ are in 1-to-1 correspondence with the subsets of the set of prime numbers. In particular, there are uncountably many such subrings.

11.4. **Remark.** It is possible to remove the cancellation property $rs = 0 \Rightarrow r = 0$ from the conditions in the proposition above. One then has to modify the definition of the equivalence relation $\sim$ as follows.

$$(r, s) \sim (r', s') \iff \exists s'' \in S : rs's'' = r'ss''$$

The price one has to pay for this is that the homomorphism ϕ is no longer injective in general. For example, when $0 \in S$, then $R_S = 0$ is the trivial ring.

12. Polynomials

In this section, we want to introduce the ring of polynomials over a given ring R. This generalizes the concept of polynomial functions, known from analysis. In algebra, however, we do not want to consider polynomials as functions, whose purpose it is to be evaluated, but as objects in their own right, that we want to add and multiply.

12.1. **Definition.** Let R be a ring. The *polynomial ring* $R[x]$ over R in one variable x is defined as the set of (finite) formal sums

$$R[x] = \Bigl\{\sum_{n=0}^{\infty} a_n x^n : a_n \in R,\, a_n = 0 \text{ for } n \gg 0\Bigr\}$$

with addition and multiplication given by

$$\sum_{n=0}^{\infty} a_n x^n + \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} (a_n + b_n) x^n \,,$$

$$\Bigl(\sum_{n=0}^{\infty} a_n x^n\Bigr) \cdot \Bigl(\sum_{n=0}^{\infty} b_n x^n\Bigr) = \sum_{n=0}^{\infty} \Bigl(\sum_{k=0}^{n} a_k b_{n-k}\Bigr) x^n \,.$$

(These are forced by the ring axioms and the wish to have $x^m \cdot x^n = x^{m+n}$.)

One has, of course, to check that these operations really turn the set $R[x]$ into a ring. This is, as usual, not hard, but tedious. There is another “higher-level” approach, which exhibits (or even defines) $R[x]$ as a subring of some already known ring. In this case, one can consider the abelian group $R^{\mathbb{N}}$ of sequences of elements of R (under component-wise addition; this is just the direct product of countably many copies of the additive group of R) and its *endomorphism ring* $E = \mathrm{End}(R^{\mathbb{N}})$.

12.2. **Example.** In general, if $(A, +)$ is an abelian group, then the set $\mathrm{End}(A)$ of endomorphisms of A (i.e., group homomorphisms $A \to A$) can be turned into a (in general noncommutative) ring by defining

$$f + g : a \mapsto f(a) + g(a)\,, \qquad f \cdot g = f \circ g : a \mapsto f(g(a))\,.$$

It is not hard to verify the ring axioms; for one of the distributive laws, one needs to use that the elements are homomorphisms. (Exercise)

Coming back to $R[x]$, we have a natural embedding $R \to E$ that sends $r \in R$ to the endomorphism $(r_0, r_1, \dots) \mapsto (rr_0, rr_1, \dots)$. Also, there is an element $x \in E$ that sends $(r_0, r_1, \dots)$ to $(0, r_0, r_1, \dots)$. We let $R[x]$ be the subring of E generated by R and x. It is then clear that every element of $R[x]$ will be a finite linear combination of powers of x with coefficients from R (note that $rx = xr$ in E for $r \in R$), that all such linear combinations are elements of $R[x]$ and that formally distinct linear combinations give distinct elements (consider the action on $(1, 0, 0, \dots)$). In this approach, we get the ring structure for free.

Here are a few basic definitions related to polynomials.

12.3. **Definition.** Let $f = \sum_n a_n x^n \in R[x]$ be a polynomial. If $f \neq 0$, we define the *degree* of f, $\deg f$, to be the largest n such that $a_n \neq 0$; then $a_{\deg f}$ is called the *leading coefficient* of f. A polynomial with leading coefficient 1 is called *monic*. We set $\deg 0 = -\infty$; the zero polynomial does not have a leading coefficient (and is therefore not monic). f is *constant* if $\deg f \le 0$, so $f = a_0$, with all other terms being zero.

12.4. **Lemma.** *There is a natural injective ring homomorphism $R \to R[x]$, mapping $r \in R$ to the constant polynomial $r \cdot x^0$.*

Proof. Easy. □

12.5. **Lemma.** *The degree has the following properties. Let $f, g \in R[x]$.*

(1) $\deg(f+g) \le \max\{\deg f, \deg g\}$, *with equality if* $\deg f \ne \deg g$.
(2) $\deg(fg) \le \deg f + \deg g$, *with equality if f and g are nonzero and the product of the leading coefficients does not vanish.*
(3) *If R is an (integral) domain, then so is $R[x]$.*

Proof. Let $f = \sum_n a_n x^n$, $g = \sum_n b_n x^n$.

(1) Let $N = \max\{\deg f, \deg g\}$; then $a_n = b_n = 0$ for $n > N$, implying $a_n + b_n = 0$ for $n > N$. This means that $\deg(f+g) \le N$.

If $\deg f < \deg g = N$ (say), then $a_N = 0$, $b_N \ne 0$, so $a_N + b_N = b_N \ne 0$, and $\deg(f+g) = N$.
(2) Let $N = \deg f$, $M = \deg g$. Then $a_n = 0$ for $n > N$ and $b_n = 0$ for $n > M$. Therefore, if $n > N * M$, then $a_n b_0 + \cdots + a_0 b_n = 0$, since in every summand, at least one of the factors is zero. Hence $\deg fg \le M + N$.

If $a_N b_M \ne 0$, then this is the leading coefficient of fg, and $\deg fg = N + M$.
(3) This follows from (2), since the product of the leading coefficients never vanishes when R is a domain.

□

As for all other algebraic constructions, there is a *universal property* of polynomial rings. It essentially says that we can evaluate polynomials (by substituting something for the variable) in a fairly general sense.

12.6. **Theorem.** *Let R be a ring, $R[x]$ the polynomial ring in one variable x over R. For every ring homomorphism $\phi : R \to R'$ and every element $c \in R'$ such that c commutes with the image of ϕ, there is a unique ring homomorphism $\Phi : R[x] \to R'$ such that $\Phi|_R = \phi$ (identifying R with its image in $R[x]$ under the canonical embedding given in Lemma 12.4) and such that $\Phi(x) = c$. In other words, Φ makes the following diagram commutative, where ψ is the map that sends x to c.*

$$\begin{tikzcd}
 & R[x] \arrow[dd, "\Phi"] & \\
R \arrow[ur] \arrow[dr, "\phi"'] & & \{x\} \arrow[ul] \arrow[dl, "\psi"] \\
 & R' &
\end{tikzcd}$$

Proof. By the condition on Φ, we must have

$$\Phi\Bigl(\sum_{n=0}^{\infty} a_n x^n\Bigr) = \sum_{n=0}^{\infty} \Phi(a_n)\Phi(x)^n = \sum_{n=0}^{\infty} \phi(a_n) c^n \,.$$

It remains to check that this Φ is a ring homomorphism. This is easy; the only point here is that we need c to commute with $\phi(r)$ for $r \in R$ in order to have $\Phi(rx) = \Phi(x \cdot r)$. □

12.7. **Example.** Let F be a field, V an F-vector space, and $f : V \to V$ a linear map. The set $\operatorname{End}(V)$ of all endomorphisms of V forms a ring (under pointwise addition and composition of maps), and there is a natural ring homomorphism $\phi : F \to \operatorname{End}(V)$, sending $\lambda \in F$ to $v \mapsto \lambda v$.

Applying the theorem, we obtain a ring homomorphism $\Phi : F[x] \to \operatorname{End}(V)$ that restricts to ϕ on F and sends x to f. In this context, the Cayley-Hamilton Theorem states that $\Phi(\operatorname{char} f) = 0$.

We will come back to this example later in the context of the Jordan Normal Form Theorem.

Seeing the universal property in Thm. 12.6 above, a natural question is whether one can construct more general polynomial rings in any set of variables. Let R be a ring, X a set. Then this polynomial ring $R[X]$ (which we write $R[x_1, x_2, \dots, x_n]$ if $X = \{x_1, x_2, \dots, x_n\}$) should have a natural ring homomorphism $R \to R[X]$ and a natural map $X \to R[X]$ such that the following universal property holds.

12.8. **Theorem.** *For every ring homomorphism $\phi : R \to R'$ and every map $\psi : X \to R'$ such that the elements of $\psi(X)$ commute with each other and the image of ϕ, there is a unique ring homomorphism $\Phi : R[X] \to R'$ such that $\Phi|_R = \phi$ and $\Phi|_X = \psi$. In other words, Φ makes the following diagram commutative.*

$$\begin{tikzcd}
 & R[X] \arrow[dd, "\Phi"] & \\
R \arrow[ur] \arrow[dr, "\phi"'] & & X \arrow[ul] \arrow[dl, "\psi"] \\
 & R' &
\end{tikzcd}$$

We can construct $R[X]$ inductively when X is finite.

12.9. **Definition.** Let $X = \{x_1, x_2, \dots, x_n\}$ be a finite set. We define the ring $R[X] = R[x_1, x_2, \dots, x_n]$ to be R when $X = \emptyset$ and to be $R[x_1, \dots, x_{n1}][x_n]$ otherwise.

It is then an easy induction based on Thm. 12.6 to prove the universal property in this case. The universal property implies that any two polynomial rings $R[X]$ (with the same R and X) are isomorphic in a unique way. This tells us, for example, that in the iterative construction above, the ordering of the elements of X does not matter.

Now we come to a very important property of polynomials: there is a procedure of division with remainder, like for integers.

12.10. **Proposition.** *Let $f, g \in R[x]$, with g monic. Then there are $q, r \in R[x]$ such that $\deg r < \deg g$ (this includes $r = 0$) and such that $f = qg + r$.*

Proof. The proof is by induction on $\deg f$ and mimics the "long division" procedure for polynomials.

First assume that $\deg f < \deg g$. The we can take $q = 0$ and $r = f$.

Now assume that $\deg f \geq \deg g$ (and that the statement is valid for f of smaller degree). Write $f = ax^{n+m} + f_1$, where $\deg g = n$, $\deg f = n + m$, $a \neq 0$ is the leading coefficient of f and $\deg f_1 < \deg f$. Write $g = x^n + g_1$ with $\deg g_1 < n$. Set

$$f_2 = f - ax^m g = ax^{n+m} + f_1 - ax^{n+m} - ax^m g_1 = f_1 - ax^m g_1 \,.$$

By Lemma 12.5, $\deg f_2 < n + m$, so by induction, there are $q_2, r \in R$ with $f_2 = q_2 g + r$ and $\deg r < \deg g$. Then $f = qg + r$, where $q = q_2 + ax^m$. □

Note that we need to assume that g is monic, since otherwise, we would have to divide by the leading coefficient of g, which is not possible in general.

The existence of a procedure like in the preceding proposition is so important that there is a special name for rings that have it

12.11. **Definition.** An integral domain R is a *Euclidean domain*, if there is a map $N : R \setminus \{0\} \to \mathbb{N}$ such that

(1) for all $a, b \in R \setminus \{0\}$, $N(ab) \ge N(b)$;
(2) for all $a \in R$, $b \in R \setminus \{0\}$, there exist $q, r \in R$ with $r = 0$ or $N(r) < N(b)$ such that $a = qb + r$.

12.12. **Examples.** The ring of integers in a Euclidean domain. If F is a field, then the polynomial ring $F[x]$ is a Euclidean domain. Indeed, all nonzero constants are invertible in this case, so we can scale g above to be monic and then apply the proposition.

We have seen earlier that all ideals of $\mathbb{Z}$ are principal. This is true more generally for Euclidean domains, with essentially the same proof.

12.13. **Definition.** An integral domain R is a *principal ideal domain* or *PID* if every ideal of R is principal (i.e., of the form Ra for some $a \in R$).

12.14. **Theorem.** *If R is a Euclidean domain, then R is a PID.*

Proof. Let $I \subset R$ be an ideal. The zero ideal is always principal, so we can assume $I \neq 0$. Let $a \in I \setminus \{0\}$ be an element with smallest possible "norm" $N(a)$ (this exists, since norms are natural numbers). Then $Ra \subset I$. To see the other inclusion, take $b \in I$. By the Euclidean property, there are q and r such that $b = qa + r$ and $r = 0$ or $N(r) < N(a)$. But the latter possibility cannot occur, since $r = b - qa \in I$, and a has smallest norm among all nonzero elements of I. So r must be zero, hence $b = qa \in Ra$. □

12.15. **Remark.** There are PIDs that are not Euclidean; an example is given by $R = \mathbb{Z}[\frac{1+\sqrt{-19}}{2}] = \mathbb{Z}[x]/(x^2 - x + 5)\mathbb{Z}[x]$.

There are integral domains that are not PIDs. For example, in $\mathbb{Z}[x]$, the ideal generated by 2 and x is not principal.

13. Divisibility and Factorization

The motivation for what we will do in the following is the wish to generalize the "Fundamental Theorem of Arithmetic", which states that every (positive, say) integer can be written as a product of prime numbers in an essentially unique way. We would like to prove this for arbitrary PIDs.

In order to do this, we first have to generalize the relevant notions from $\mathbb{Z}$ to more general rings.

13.1. **Definition.** Let R be an integral domain.

(1) Let $a, b \in R$. We say that *a divides b*, or that *b is divisible by a*, written $a \mid b$, if there is $c \in R$ such that $b = ac$.
This is equivalent to saying that $b \in Ra$ or that $Rb \subset Ra$.

(2) Two elements $a, b \in R$ are *associate*, written $a \sim b$, if $a \mid b$ and $b \mid a$.
Equivalently, $Ra = Rb$, or $a = bu$ with a unit $u \in R^\times$.

(3) An element $a \in R \setminus (R^\times \cup \{0\})$ is called *irreducible*, if it cannot be written as a product in a nontrivial way: $a = bc$ implies $b \sim a$ or $b \in R^\times$ (note that $b \sim a \iff c \in R^\times$ when $a = bc$).
Equivalently, Ra is maximal among proper principal ideals.

Note that the definition of irreducible elements is analogous to the usual definition of prime numbers in $\mathbb{Z}$. We will, however, later introduce the notion of *prime* elements, which captures a different property. However, as we will see, in a PID both notions coincide.

Now we can already show that we can write every nonzero nonunit as a product of irreducibles.

13.2. **Theorem.** *Let R be a PID, $a \in R \setminus (\{0\} \cup R^\times)$. Then a can be written as a product of irreducible elements of R.*

Proof. If R were a Euclidean domain, then we could mimick the usual proof for the integers by doing an induction on $N(a)$. (Note that in a nontrivial factorization $a = bc$, we have $N(b), N(c) < N(a)$.) But a general PID is not Euclidean, and so we have to use some other sort of induction principle. Let us start by assuming the statement is false. Then there is some $a_0 \in R \setminus (\{0\} \cup R^\times)$ that is not a product of irreducibles. In particular, a_0 itself is not irreducible, and so we can write $a_0 = bc$ with b and c non-units. If b and c both are products of irreducibles, then so would be a_0, so at least one of the two is not a product of irreducibles. Call it a_1. Then a_1 is not a product of irreducibles, and $Ra_0 \subsetneq Ra_1$. Continuing in this way, we obtain a strictly increasing sequence of principal ideals

$$Ra_0 \subsetneq Ra_1 \subsetneq Ra_2 \subsetneq Ra_3 \subsetneq \dots$$

Let $I = \bigcup_n Ra_n$. Since R is a PID, $I = Rr$ for some $r \in R$. But then there must be some n such that $r \in Ra_n$. It follows that

$$Rr \subset Ra_n \subsetneq Ra_{n+1} \subset I = Rr\,,$$

a contradiction. Therefore, our initial assumption must be false, and the statement of the theorem must be true. □

Looking at this proof, we realize that we only need to know that there cannot be an ever-increasing sequence of ideals, and so the theorem should hold for even more general rings.

13.3. **Definition.** A commutative ring R is called *noetherian* (derived from the name of the female German mathematician Emmy Noether) if every ascending chain of ideals of R

$$I_0 \subset I_1 \subset I_2 \subset \dots$$

becomes stationary, i.e., there is some n such that

$$I_n = I_{n+1} = I_{n+1} = \dots\,.$$

13.4. **Corollary.** *Let R be a noetherian integral domain. Then every element $a \in R \setminus (\{0\} \cup R^\times)$ is a product of irreducible elements.*

Proof. Same proof as before, using the noetherian property to derive the contradiction. □

Here are two other useful characterization of noetherian rings.

13.5. **Proposition.** *Let R be a commutative ring. Then the following statements are equivalent.*

(1) *R is noetherian.*
(2) *Every nonempty set of ideals of R has a maximal element.*
(3) *Every ideal of R is finitely generated.*

Proof. "(1) ⇒ (2)": Assume R is noetherian, and let X be a nonempty set of ideals of R. Assume X has no maximal element. Then there is $I_0 \in X$ (since $X \neq \emptyset$). I_0 is not maximal, so there is $I_1 \in X$ such that $I_0 \subsetneq I_1$. Continuing in this way, we find a chain of ideals

$$I_0 \subsetneq I_1 \subsetneq I_2 \subsetneq I_3 \subsetneq \dots ,$$

contradicting the noetherian property of R.

"(2) ⇒ (1)": Assume (2) and let

$$I_0 \subset I_1 \subset I_2 \subset I_3 \subset \dots$$

be an ascending chain of ideals. Let $X = \{I_0, I_1, I_2, \dots\}$. By assumption, X has a maximal element I_n. But this implies $I_n = I_{n+1} = I_{n+1} = \dots$, so R is noetherian.

"(1) ⇒ (3)": Suppose there is an ideal I that is not finitely generated. Pick $r_0 \in I$ and let $I_0 = Rr_0 \subset I$. Since I is not finitely generated, $I_0 \subsetneq I$, so we can pick $r_1 \in I \setminus I_0$. Let $I_1 = I_0 + Rr_1$; then $I_0 \subsetneq I_1 \subset I$. I_1 is finitely generated, but I is not, so there is $r_2 \in I \setminus I_1$. Pick $r_2 \in I \setminus I_1$, and let $I_2 = I_1 + Rr_2$; then $I_0 \subsetneq I_1 \subsetneq I_2 \subset I$. Continuing in this way, we find a chain $I_0 \subsetneq I_1 \subsetneq I_2 \subsetneq \dots$, contradicting the noetherian property of R.

"(3) ⇒ (1)": Assume that all ideals are finitely generated. Let $I_0 \subset I_1 \subset I_2 \subset \dots$ be an ascending chain of ideals. Let $I = \bigcup_n I_n$. By assumption, I is finitely generated, so $I = Rr_1 + Rr_2 + \dots + Rr_m$ for elements $r_1, \dots, r_m \in R$. Then there are indices n_j such that $r_j \in I_{n_j}$. Taking $n = \max\{n_1, \dots, n_m\}$, we have that $r_1, \dots, r_m \in I_n$. But then $I_n = I$, and therefore $I_n = I_{n+1} = I_{n+2} = \dots$. □

13.6. **Examples.** Clearly, any field is a noetherian ring, as there are only two ideals.

Property (3) above implies (again) that any PID is noetherian: in a PID, every ideal is even generated by one element.

It is not so easy to come up with an example of a commutative ring that is *not* noetherian (we will see why this is so later in this course). The usual example is the polynomial ring R in countably infinitely many variables $x_1, x_2, x_3, \dots$ over a field F (say); then

$$0 \subsetneq Rx_1 \subsetneq Rx_1 + Rx_2 \subsetneq Rx_1 + Rx_2 + Rx_3 \subsetneq \dots$$

is an ascending chain of ideals that does not become stationary.

This example also shows that a subring of a noetherian ring need not be noetherian — R is a subring of its field of fractions, which is a noetherian ring (being a field).

We have seen that *existence* of a factorization into irreducibles holds quite generally for noetherian integral domains. What about *uniqueness*?

We first have to figure out how unique we can such a factorization expect to be. The most obvious way of changing a given factorization into another one is to reorder the factors, and we certainly do not want to consider two factorizations differing only in the ordering of the factors as essentially distinct (bear in mind that we are working in commutative rings). But there is also another way in which we can modify a factorization, and this is by multiplying each factor by a unit (such that the product of all the units is 1). We also do not want to count two factorizations as essentially distinct if they only differ in this way.

13.7. **Definition.** An integral domain R is a *unique factorization domain* or *UFD*, if every element $a \in R \setminus (\{0\} \cup R^\times)$ can be written as a product of irreducible elements, and if $a = r_1 \cdots r_m = s_1 \cdots s_n$ are two such factorizations, then $m = n$, and there is a permutation $\sigma \in S_n$ such that $r_j \sim s_{\sigma(j)}$ for all $j = 1, \dots, n$.

The “Fundamental Theorem of Arithmetic” states that $\mathbb{Z}$ is a UFD.

How do we prove that factorization into primes in unique in $\mathbb{Z}$? We use another property of prime numbers: if p is prime, and p divides a product ab, then p must divide one of the factors.

13.8. **Definition.** Let R be a commutative ring.

(1) Suppose R is an integral domain. An element $p \in R \setminus (\{0\} \cup R^\times)$ is called *prime*, if $p \mid ab$ implies $p \mid a$ or $p \mid b$, for all $a, b \in R$.

(2) A proper ideal $P \subset R$ is called a *prime ideal* if $ab \in P$ implies $a \in P$ or $b \in P$, for all $a, b \in R$.

It is clear that $p \in R \setminus \{0\}$ is prime if and only if the principal ideal Rp is a prime ideal.

Note that in the definition of “prime ideal”, $P = 0$ is allowed. The zero ideal is prime if and only if R is an integral domain.

Let us first investigate uniqueness of factorization and then apply the results to the case when R is a PID.

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

(1) *If $p \in R$ is prime, then p is irreducible.*

(2) *If $p, q \in R$ are two prime elements such that $p \mid q$, then $p \sim q$.*

Proof.
(1) If $p = ab$, then $p \mid a$ or $p \mid b$, so $p \sim a$ or $p \sim b$, i.e., $a \sim p$ or $a \in R^\times$.

(2) By assumption, $q = ap$ with some $a \in R$. By part (1), $p \sim q$, or else p wold have to be a unit, which is not the case. □

13.10. **Proposition.** *Let R be an integral domain and suppose*

$$r_1 r_2 \cdots r_m \sim s_1 s_2 \cdots s_n$$

with prime elements $r_1, \ldots, r_m$ and $s_1, \ldots, s_n$. Then $m = n$, and up to reordering the factors, $r_j \sim s_j$ for $j = 1, \ldots, n$.

Proof. The proof is by induction on n (say). If $n = 0$, then $s_1 \cdots s_n = 1$, and so $r_1 \cdots r_m$ is a unit. If $m > 0$, then this would imply that r_1 is a unit, contradicting the assumption that r_1 is prime.

Now assume that $n > 0$ and that the statement is true for smaller values of n. s_n divides $r_1 \cdots r_m$, so $s_n \mid r_j$ for some $j \in \{1, \ldots, m\}$, and (up to reordering) we can assume that $j = m$. By Lemma 13.9, $s_n \sim r_m$, and therefore, cancelling these factors, $r_1 \cdots r_{m-1} \sim s_1 \cdots s_{n-1}$. By the induction hypothesis, $m - 1 = n - 1$, so $m = n$, and the factors are associate, as claimed. □

13.11. **Corollary.** *If R is a noetherian integral domain such that every irreducible element of R is prime, then R is a UFD.*

Proof. Since R is noetherian, factorizations into irreducibles exist by Cor. 13.4. By assumption, all irreducibles are prime, so by Prop. 13.10, any two factorization of the same element into irreducibles (and therefore primes) only differ by units and ordering of the factors. □

13.12. **Lemma.** *If R is a UFD and $a \in R$ is irreducible, then a is prime.*

Proof. Assume that a divides bc, so $bc = ad$ for some $d \in R$. If $d \in R^\times$, then $b \sim a$ or $c \sim a$, since a is irreducible. If $b \in R^\times$ (or $c \in R^\times$), then $a \mid c$ ($a \mid b$). So we cann assume that none of b, c, d is a unit. Write b, c and d as products of irreducibles: $b = b_1 \cdots b_m$, $c = c_1 \cdots c_n$, $d = d_1 \cdots d_k$. Then

$$b_1 \cdots b_m c_1 \cdots c_n = a d_1 \cdots d_k$$

are two factorizations into irreducibles of the same element. Hence a must be associate to one of the factors in the left hand side, implying that a divides b or c. □

So in a UFD, the notions of irreducible and prime elements coincide. Therefore one usually talks about *prime factorization* in a UFD.

13.13. **Lemma.** *Let R be a commutative ring.*

(1) *An ideal $I \subset R$ is prime if and only if R/I is an integral domain.*
(2) *If $M \subset R$ is a maximal ideal, then M is a prime ideal.*

Proof. (1) Write $\bar{a}$ for $a + I$, the image of a in R/I. I is a prime ideal if and only if $ab \in I$ implies $a \in I$ or $b \in I$, if and only if $ab + I = I$ implies $a + I = I$ or $b + I = I$, if and only if $\overline{ab} = \bar{0}$ implies $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$ in R/I, if and only if R/I has no zero divisors.

(2) If M is a maximal ideal, then R/M is a field and therefore an integral domain. By part (1), M is a prime ideal. □

13.14. **Theorem.** *Let R be an integral domain. If R is a PID, then R is a UFD.*

Proof. Since R is a PID, R is noetherian. It remains to show that every irreducible element of R is prime. So assume that $a \in R$ is irreducible. Then Ra is maximal among proper principal ideals. But all ideals are principal in a PID, so Ra is a maximal ideal. By Lemma 13.13, Ra is a prime ideal, and so a is a prime element. □

13.15. **Corollary.** *Let F be a field. Then $F[x]$ is a UFD.*

14. Greatest Common Divisors

We have seen that a Euclidean domain is a PID, and a PID is a UFD. In a UFD, *greatest common divisors* exist, as we will see. In a PID, greatest common divisors have a nice interpretation, and in Euclidean domains, we can compute them easily.

But first let us give a slightly different formulation of the unique factorization property.

14.1. **Proposition.** *Let R be a UFD. Pick a system $\mathcal{P}$ of representatives of the prime elements of R modulo $\sim$. Then every $a \in R \setminus \{0\}$ can be written uniquely as*

$$a = u \prod_{p \in \mathcal{P}} p^{e_p}$$

with $u \in R^\times$ and $e_p \in \{0, 1, 2, \dots\}$ with $e_p \neq 0$ for only finitely many $p \in \mathcal{P}$.

If

$$a = u \prod_{p \in \mathcal{P}} p^{e_p} \quad \text{and} \quad b = v \prod_{p \in \mathcal{P}} p^{f_p} ,$$

then $a \mid b$ if and only if $e_p \leq f_p$ for all $p \in \mathcal{P}$.

Proof. Existence: If a is a unit, then we can take $u = a$ and $e_p = 0$ for all p. Otherwise, we can write a as a product of primes: $a = r_1 \cdots r_n$. Let

$$e_p = \#\{j \in \{1, \dots, n\} : p \sim r_j\} ;$$

then $e_p = 0$ for all but finitely many p, $\sum_p e_p = n$, and $\prod_p p^{e_p} \sim r_1 \cdots r_n = a$, so there is a unit u such that $a = u \prod_p p^{e_p}$.

Uniqueness: Assume that

$$a = u \prod_{p \in \mathcal{P}} p^{e_p} = v \prod_{p \in \mathcal{P}} p^{f_p} .$$

If a is a unit, then all the e_p and f_p must be zero, so $a = u = v$. Otherwise, we can temporarily combine the units u and v with one of the factors in the product. Then by uniqueness of factorization, for every $p \in \mathcal{P}$, the number of factors associate to p must be the same in both products. This means that $e_p = f_p$ for all p, which in turn implies $u = v$.

The second statement is clear. □

In some cases, there is a natural choice for the set $\mathcal{P}$. If $R = \mathbb{Z}$, we take for $\mathcal{P}$ the set of prime numbers (i.e., positive prime (= irreducible) elements of $\mathbb{Z}$). If $R = F[x]$ for a field F, we take for $\mathcal{P}$ the set of monic irreducible polynomials. (Note that $F[x]^\times = F^\times$, so associate polynomials only differ by scaling with a nonzero constant factor.)

14.2. **Definition.** Let R be an integral domain, $a, b \in R \setminus \{0\}$. An element $d \in R$ is a *greatest common divisor* of a and b if $d \mid a$ and $d \mid b$, and for every $r \in R$ such that $r \mid a$ and $r \mid b$, we have $r \mid d$. We write $\gcd(a, b) = d$, even though in general, d is not uniquely determined — it is determined up to associates (if it exists).

In a similar way, we can define the greatest common divisor of any set S of elements of R: $d = \gcd(S)$ if $d \mid s$ for all $s \in S$, and if for all $r \in R$ such that $r \mid s$ for all $s \in S$, we have $r \mid d$. Note that the greatest common divisor of the empty set exists and is zero (0 is the largest element of R with respect to divisibility).

It is then clear that if gcds of pairs of elements exist, then the gcd of any finite set of elements exists, and

$$\gcd(a_1, \dots, a_n) = \gcd(a_1, \gcd(a_2, \gcd(a_3, \dots))) .$$

14.3. **Proposition.** *Let R be a UFD. Then any two elements a and b of R have a greatest common divisor.*

Proof. If $a = 0$, then b is a gcd of a and b; if $b = 0$, then a is a gcd of a and b. So we can assume that $a, b \neq 0$. By Prop. 14.1, we can write

$$a = u \prod_{p \in \mathcal{P}} p^{e_p} , \qquad b = v \prod_{p \in \mathcal{P}} p^{f_p} .$$

Then $d = \prod_{p \in \mathcal{P}} p^{\min\{e_p, f_p\}}$ is a gcd of a and b, by the second statement in Prop. 14.1. □

In the same way, one can prove that $\gcd(S)$ exists for any set $S \subset R$.

So to find a gcd in a general UFD, we first need to find the factorizations of a and b. This can be rather hard — even for integers, there are no efficient algorithms known.

14.4. **Proposition.** *Let R be a PID, $a, b \in R$. Then there is $d \in R$ such that $Ra + Rb = Rd$, and $\gcd(a, b) = d$. In particular, any multiple of a greatest common divisor of a and b can be written as $ua + vb$ for suitable $u, v \in R$.*

Proof. The ideal $Ra+Rb$ is principal, hence there is $d \in R$ such that $Ra+Rb = Rd$. Since $a \in Ra \subset Rd$, $b \in Rb \subset Rd$, d divides both a and b. Conversely, suppose $r \in R$ divides both a and b. Then $Ra, Rb \subset Rr$, hence $Rd = Ra + Rb \subset Rr$ (recall that $Ra + Rb$ is the smallest ideal containing Ra and Rb), so r divides d. □

14.5. **Remark.** An analogous statement is true for any set $S \subset R$: the ideal generated by S equals Rd for some $d \in R$, and then $d = \gcd(S)$, and d is a finite linear combination of elements of S. In particular, d is already the GCD of a finite subset of S. In fact, this statement is already true in a general UFD (Exercise).

So in a PID, we can reduce to the problem of finding a generator of an ideal generated by two elements. If the ring is Euclidean, then there is a efficient way of doing this.

14.6. **Theorem.** *Let R be a Euclidean domain, $a, b \in R$. We can compute a gcd of a and b as follows. Set $r_0 = a$, $r_1 = b$, $n = 1$. While $r_n \neq 0$, write $r_{n-1} = q_n r_n + r_{n+1}$ with $r_{n+1} = 0$ or $N(r_{n+1}) < N(r_n)$, and replace n by $n + 1$. This loop terminates, and upon exit, we have* $\gcd(a, b) = r_{n-1}$.

Proof. Since $N(r_1) > N(r_2) > N(r_3) > \dots$, the loop must terminate with $r_n = 0$. Let $I = Ra + Rb$. Then $Rr_{n-1} + Rr_n = I$ for all n (such that r_n is defined): this is certainly true for $n = 0$, and

$$Rr_n + Rr_{n+1} = Rr_n + R(r_{n-1} - q_n r_n) = Rr_{n-1} + Rr_n \,.$$

Upon exit from the loop, $r_n = 0$, therefore $I = Rr_{n-1}$, and so r_{n-1} is a gcd of a and b. □

This algorithm is the *Euclidean Algorithm.* It can be extended to also provide elements $u, v \in R$ such that $ua + vb = \gcd(a, b)$.

In the case $R = \mathbb{Z}$ (or, for example, $R = \mathbb{Q}[x]$), this provides a polynomial time algorithm for computing greatest common divisors.

14.7. **Remark.** This illustrates a general fact. While essentially all the nice properties can be proved for PIDs, the Euclidean property allows for efficient algorithms, which may be much harder to find for general PIDs.

The notion of greatest common divisor is crucial in the proof of the important fact that $R[x]$ is a UFD if R is a UFD.

14.8. **Definition.** Let R be a UFD, $f \in R[x]$. The *content* of f is the gcd of its (non-zero) coefficients. f is called *primitive* if it has content 1.

If $f \neq 0$, then f can be written as $f = cf_0$, where c is constant (the content) and f_0 is primitive.

14.9. **Gauss' Lemma.** *Let R be a UFD. If $f, g \in R[x]$ are primitive, then so is their product fg.*

Proof. Assume fg is not primitive. Then there is a prime element $p \in R$ that divides all the coefficients of fg. On the other hand, p does not divide all the coefficients of f, and p does not divide all the coefficients of g. Let $f = \sum_n a_n x^n$, $g = \sum_n b_n x^n$ and define

$$k = \min\{n : p \nmid a_n\}\,, \quad m = \min\{n : p \nmid b_n\}\,.$$

The coefficient of x^{k+m} in fg is

$$c = a_{k+m}b_0 + \dots + a_k b_m + \dots + a_0 b_{k+m}\,.$$

In all terms except $a_k b_m$, p divides one of the factors, but p does not divide $a_k b_m$. Therefore p does not divide c, a contradiction. □

14.10. **Remark.** Let R be an integral domain, F its field of fractions. Then there is a canonical embedding $R[x] \to F[x]$ (using the embedding of R into F and sending x to x).

If R is a UFD and $0 \neq f \in F[x]$, then f can be written as cf_0 with $c \in F^\times$ and $f_0 \in R[x]$ primitive: some multiple rf with $r \in R \setminus \{0\}$ is in $R[x]$ (take for r the product of all denominators of the coefficients, for example); then we can write $rf = af_0$ with f_0 primitive. Putting $c = a/r$ gives the result.

In particular, we can choose a set $\mathcal{P}$ of representatives of the irreducible polynomials in $F[x]$ up to associates that contains only primitive polynomials with coefficients in R.

14.11. **Corollary.** *Let R be a UFD, F its field of fractions. Let $f \in R[x]$ be a primitive polynomial. Then f is irreducible in $R[x]$ if and only if (the image of) f is irreducible in $F[x]$.*

Proof. Assume that f is irreducible in $R[x]$. Suppose $f = gh$ in $F[x]$. Write $g = ag_0$, $h = bh_0$ with $a, b \in F^\times$, $g_0, h_0 \in R[x]$ primitive. Then $f = (ab)f_0$ with $f_0 = g_0h_0 \in R[x]$ primitive (by Gauss' Lemma). Since f is primitive, this implies $ab \in R^\times$. But then $f = (abg_0)h_0$ is a factorization in $R[x]$, therefore (f being irreducible) $\deg g = \deg g_0 = 0$ or $\deg h = \deg h_0 = 0$. This shows that f is also irreducible in $F[x]$.

For the converse, assume that f is irreducible in $F[x]$. If $f = gh$ in $R[x]$, then the same factorization holds in $F[x]$, and so we must have that g or h is constant. Since f is primitive, this constant must be a unit, showing that f is irreducible in $R[x]$. □

14.12. **Example.** We can use this result in order to prove that a given polynomial of degree 3 is irreducible in $\mathbb{Q}[x]$. Consider for example $f = 5\,x^3 + 7\,x + 1$. f has integral coefficients and is primitive, so f is irreducible in $\mathbb{Q}[x]$ if and only if f is irreducible in $\mathbb{Z}[x]$. If f is not irreducible in $\mathbb{Z}[x]$, then it has a factor $ax + b$ of degree 1, with $a, b \in \mathbb{Z}$. Then a must divide 5 and b must divide 1, so there are only four possibilities for the linear factor (note that we can assume a to be positive): $x+1$, $x-1$, $5x+1$, $5x-1$. Since none of $f(-1)$, $f(1)$, $f(-1/5)$, $f(1/5)$ is zero, none of these polynomials divides f, and so f is irreducible.

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

Proof. As before, let F be the field of fractions of R. We know that $F[x]$ is a UFD. Let $\mathcal{P}$ be the set specified in the remark above. By the corollary above, all the elements of $\mathcal{P}$ are irreducible in $R[x]$. Every nonzero element f of $R[x] \subset F[x]$ can be written uniquely

$$f = c \prod_{p \in \mathcal{P}} p^{e_p}$$

with $c \in F^\times$. Since the product is primitive by Gauss' Lemma, c is in fact in $R \setminus \{0\}$. Since R is a UFD, c is a product of irreducibles of R (which stay irreducible in $R[x]$). This proves existence of the factorization. Uniqueness follows from the fact that the e_p are uniquely determined and uniqueness of factorization in R. □

14.14. **Examples.** $\mathbb{Z}[x]$, and more generally, $\mathbb{Z}[x_1, x_2, \ldots, x_n]$ are UFDs.

$F[x_1, x_2, \ldots, x_n]$ is a UFD when F is a field.

15. More About Noetherian Rings

We have seen that some properties of a ring R carry over to the polynomial ring $R[x]$ —

$$\begin{aligned} R \text{ commmutative} &\implies R[x] \text{ commutative} \\ R \text{ is a domain} &\implies R[x] \text{ is a domain} \\ R \text{ is a UFD} &\implies R[x] \text{ is a UFD} \end{aligned}$$

There is at least one other very important implication of this kind, which we want to prove (among other things) in this section:

$$R \text{ noetherian} \implies R[x] \text{ noetherian}$$

15.1. **Theorem (Hilbert Basis Theorem).** *Let R be a noetherian commutative ring. Then $R[x]$ is also noetherian.*

Proof. Let $I \subset R[x]$ be an ideal. The main idea in this proof is to consider the "ideals of leading coefficients" associated to I: for $n \geq 0$, set

$$L_n(I) = \{a \in R : a\,x^n + f \in I \text{ for some } f \in R[x] \text{ with } \deg f < n\}\,.$$

It is clear that $L_n(I)$ is an ideal. We also have that (use that $x \cdot I \subset I$)

$$L_0(I) \subset L_1(I) \subset L_2(I) \subset \dots$$

Since R is noetherian, this chain stabilizes, and we have

$$L_n(I) = L_{n+1}(I) = L_{n+2}(I) = \dots =: L(I)$$

for some $n = n(I)$. Now the key point is the following claim.

If $I \subset I'$ are two ideals of $R[x]$ such that $L_n(I) = L_n(I')$ for all n, then $I = I'$.

To prove this claim, assume that $I \subsetneq I'$ and consider $f \in I' \setminus I$ with $m = \deg f$ minimal. Since $L_m(I) = L_m(I')$, there is $g \in I$ with $\deg g = m$ and $\deg(f-g) < m$. Now $f, g \in I'$, so $f - g \in I'$, hence by minimality of m, $f - g \in I$. But this implies $f = (f - g) + g \in I$, a contradiction.

Now consider a chain

$$I_0 \subset I_1 \subset I_2 \subset \dots$$

of ideals of $R[x]$. We obtain a chain

$$L(I_0) \subset L_(I_1) \subset L(I_2) \subset \dots$$

of ideals of R, which beomes stationary at some m:

$$L(I_m) = L(I_{m+1}) = L(I_{m+2}) = \dots =: L$$

Let $n = n(I_m)$ be such that $L_n(I_m) = L_{n+1}(I_m) = \dots = L(I_m)$. Then for all $N \geq n$ and all $M \geq m$, we have

$$L = L(I_m) = L_n(I_m) \subset L_n(I_M) \subset L_N(I_M) \subset L(I_M) = L\,,$$

so $L_N(I_M) = L$. Now consider the chains

$$L_k(I_0) \subset L_k(I_1) \subset L_k(I_2) \subset \dots$$

for $k = 0, 1, \dots, n-1$. They all become stationary eventually, so there is some $m' \geq m$ such that

$$L_k(I_{m'}) = L_k(I_{m'+1}) = L_k(I_{m'+2}) = \dots$$

for all $0 \leq k < n$ and therefore for *all* $k \geq 0$. By the claim proved above, this implies

$$I_{m'} = I_{m'+1} = I_{m'+2} = \dots ,$$

so our original chain of ideals of $R[x]$ also stabilizes. □

There is another construction that preserves the noetherian property.

15.2. **Lemma.** *Let R be a commutative ring, $I \subset R$ an ideal. If R is noetherian, then the quotient ring R/I is also noetherian.*

Proof. Indeed, let $\phi : R \to R/I$ be the quotient map, and let

$$J_0 \subset J_1 \subset J_2 \subset \dots$$

be a chain of ideals of R/I. Then

$$\phi^{-1}(J_0) \subset \phi^{-1}(J_1) \subset \phi^{-1}(J_2) \subset \dots$$

is a chain of ideals of R, which by assumption becomes stationary:

$$\phi^{-1}(J_n) = \phi^{-1}(J_{n+1}) = \dots$$

for some n. But then

$$J_n = \phi(\phi^{-1}(J_n)) = \phi(\phi^{-1}(J_{n+1})) = J_{n+1}$$

etc., so the chain of ideals of R/I also becomes stationary. □

In order to summarize these results in a convenient way, we introduce yet another notion.

15.3. **Definition.** Let R be a commutative ring. An *R-algebra* is a ring A, together with a ring homomorphism $\phi : R \to A$ such that $\phi(R) \subset Z(A)$ (where $Z(A) = \{a \in A : \forall b \in A : ab = ba\}$ is the center of A).

Let (A, ϕ), (A', ϕ') be two R-algebras. An *R-algebra homomorphism* $A \to A'$ is a ring homomorphism $\psi : A \to A'$ such that $\psi \circ \phi = \phi'$.

$$\begin{tikzcd}
A \arrow[rr, "\psi"] & & A' \\
& R \arrow[ul, "\phi"] \arrow[ur, "\phi'"'] &
\end{tikzcd}$$

15.4. **Examples.** R, together with the identity homomorphism, is an R-algebra. More generally, every quotient R/I of R (with the quotient map) is an R-algebra.

The polynomial ring $R[x]$, or more generally, $R[x_1, x_2, \dots, x_n]$ (or even $R[X]$ for any set X) is an R-algebra (with the canonical embedding $R \to R[X]$). The universal property of the polynomial ring $R[x_1, \dots, x_n]$ can then be stated as follows.

Let A be an R-algebra, and let $a_1, \dots, a_n \in A$ be commuting elments. Then there is a unique R-algebra homomorphism $\psi : R[x_1, \dots, x_n] \to A$ satisfying $\psi(x_j) = a_j$ for all $j = 1, \dots, n$.

Compare this to the universal property of the F-vector space V on the basis $x_1, \dots, x_n$:

Let W be an F-vector space, $w_1, \dots, w_n$. Then there is a unique F-linear map $\psi : V \to W$ satisfying $\psi(x_j) = w_j$ for all $j = 1, \dots, n$.

In this sense, $R[x_1, \dots, x_n]$ is the *free commutative R-algebra* on the set $\{x_1, \dots, x_n\}$.

The matrix ring $\mathrm{Mat}_n(R)$ of $n \times n$ matrices with entries in R is an R-algebra (with the embedding sending r to r times the identity matrix). This algebra is not commutative when $n \geq 2$.

15.5. **Definition.** Let (A, ϕ) be an R-algebra. An *R-subalgebra* of A is a subring $B \subset A$ containing the image of ϕ, together with the restricted homomorphism $\phi_B : R \to B$.

The intersection of any family of R-subalgebras of A is again an R-subalgebra. Therefore we can speak of the R-subalgebra *generated* by a subset $S \subset A$. The R-algebra A is *finitely generated* if it equals the R-subalgebra generated by a finite subset $S \subset A$.

15.6. **Lemma.** *Let (A, ϕ) be a commutative R-algebra. Then A is a finitely generated R-algebra if and only if A is isomorphic to a quotient $R[x_1, \dots, x_n]/I$ of a polynomial ring in finitely many variables over R.*

Proof. Assume A is finitely generated, with generators $a_1, \dots, a_n$. By the universal property of the polynomial ring, there is an R-algebra homomorphism $\psi : R[x_1, \dots, x_n] \to A$ sending x_j to a_j. Its image is an R-subalgebra containing the generating set, hence ψ is surjective. Let $I = \ker \psi$, then $A \cong R[x_1, \dots, x_n]/I$.

Conversely, let $A = R[x_1, \dots, x_n]/I$. Then A is generated by the images of $x_1, \dots, x_n$ as an R-algebra. □

Using the notion of R-algebra, the two main results of this section can be conveniently summarized as follows.

15.7. **Corollary.** *Let R be a noetherian commutative ring, A a finitely generated commutative R-algebra. Then A is also noetherian.*

15.8. **Remark.** Let F be a field. Finitely generated F-algebras (also called F-algebras *of finite type*) are important because they occur in Algebraic Geometry as the "affine coordinate rings" of "affine algebraic sets" — subsets of affine n-space defined by a collection of polynomial equations.

Note that the Hilbert Basis Theorem implies that every set of polynomial equations in a finite number of variables $x_1, \dots, x_n$ over a field F is equivalent to a *finite* set of such equations. Indeed, the set of all equations that can be derived from the collection of equations $f = 0$ for $f \in S \subset F[x_1, \dots, x_n]$ is given by $f = 0$ for f in the ideal I generated by S. By the Hilbert Basis Theorem, this ideal is finitely generated, leading to a finite set of equations that is equivalent to the original set. (The name Hilbert *Basis* Theorem refers to the fact that it implies that every ideal in $F[x_1, \dots, x_n]$ has a finite "basis", i.e., generating set.)

16. Modules — Basics

Modules are a generalization of vector spaces. Basically, they are like vectors spaces, but instead of the field of scalars, we now have a commutative ring. The fact that rings are in some respects less "nice" than fields (or, put positively, that rings have a richer structure) leads to new phenomena. For example, it is no longer true that a finitely generated module necessarily has a basis.

16.1. **Definition.** Let R be a commutative ring. An *R-module* is an abelian group M, together with a “scalar multiplication” $R \times M \to M$, usually written $(r, m) \mapsto rm$, satisfying the following conditions.

(1) Distributivity: $r(m + m') = rm + rm'$ for all $r \in R$, $m, m' \in M$,
$(r + r')m = rm + r'm$ for all $r, r' \in R$, $m \in M$.
(2) Identity: $1m = m$ for all $m \in M$
(3) Associativity: $(rr')m = r(r'm)$ for all $r, r' \in R$, $m \in M$.

16.2. **Remark.** This definition makes sense for arbitrary (not necessarily commutative) rings R. In this case, what we have defined is called a *left* R-module, and *right* R modules are defined in a similar way, using a multiplication map $M \times R \to M$.

16.3. **Remark.** Alternatively, if M is an abelian group, giving an R-module structure on M is equivalent to specifying a ring homomorphism $R \to \mathrm{End}(M)$, where $\mathrm{End}(M)$ is the endomorphism ring of M (addition point-wise, multipication is composition of endomorphisms).

This may (and should!) remind you of group actions — an R-module is essentially a ring acting on an abelian group.

16.4. **Definition.** Let M be an R-module. An *R-submodule* of M is an abelian subgroup M' of M, closed under multiplication by elements from R: $rm' \in M'$ for all $r \in R$, $m' \in M'$.

Every module has the *trivial submodules* $0 = \{0\}$ and M.

The intersection of any family of submodules of M is again a submodule. Therefore it makes sense to speak of the submodule *generated by a subset $S \subset M$*. The R-module M is *finitely generated* if there is a finite subset $S \subset M$ such that M is the submodule generated by S. M is called *cyclic* if M is generated by one element m; then M has the form $Rm = \{rm : r \in R\}$.

If $M_1, M_2 \subset M$ are two submodules, then their sum $M_1 + M_2$ as abelian groups is the submodule generated by $M_1 \cup M_2$, i.e., the smallest submodule containing both M_1 and M_2. The analogous statement holds for any family of submodules (where the sum $\sum_{i \in I} M_i$ is defined to be the set of all *finite* sums of elements from $\bigcup_{i \in I} M_i$).

16.5. **Examples.** An abelian group A is the same as a $\mathbb{Z}$-module. Indeed, there is always a natural multiplication $\mathbb{Z} \times A \to A$, given by taking multiples of an element. Equivalently, there is always a unique ring homomorphism $\mathbb{Z} \to \mathrm{End}(A)$.

The ring R is itself an R-module, using the ring multiplication. The submodules are then exactly the ideals of R. Similarly, R^n (the set of n-tuples of elements of R) is an R-module with addition and scalar multiplication defined component-wise. R^n is called the *(finitely generated) free R-module of rank n.*

More generally, any R-algebra (A, ϕ) is an R-module, with scalar multiplication given by $ra = \phi(r)a$.

If F is a field, then an F-module is the same as an F-vector space.

Let F be a field, V an F-vector space and $\phi : V \to V$ an F-linear map. Then V is an $F[x]$-module via the homomorphism $F[x] \to \mathrm{End}(V)$ that sends x to ϕ (and F to scalar mulitplications). If $f \in F[x]$ is the characteristic or minimal polynomial of ϕ, then V is an $F[x]/F[x]f$-module in the same way (since the homomorphism $F[x] \to \operatorname{End}(V)$ above has kernel containing f, it factors through $F[x]/F[x]f$). We will use this point of view for the proof of the Jordan Normal Form Theorem for matrices.

16.6. **Definition.** Let M and M' be two R-modules. An *R-module homomorphism* or *R-linear map* from M to M' is a homomorphism $\phi : M \to M'$ of abelian groups such that $\phi(rm) = r\phi(m)$ for all $r \in R$, $m \in M$.

The image of ϕ is a submodule of M'.

If ϕ is bijective, it is called an *R-module isomorphism*; in this case M and M' are called *isomorphic*, $M \cong M'$. As usual, in this case the inverse ϕ^{-1} is again an R-module homomorphism.

The composition of two R-module homomorphisms is again an R-module homomorphism.

The *kernel* of ϕ is the kernel as homomorphism of additive groups:

$$\ker\phi = \{m \in M : \phi(m) = 0\}\,.$$

Note that $\ker\phi$ is a submodule of M, and ϕ is injective if and only if $\ker\phi = 0$ is the zero submodule.

We denote by $\operatorname{Hom}_R(M, M')$ the set of all R-module homomorphisms $M \to M'$. This set has the structure of an R-module under point-wise addition and scalar multiplication:

$$(\phi + \psi)(m) = \phi(m) + \psi(m)\,, \qquad (r\phi)(m) = r\phi(m)\,.$$

In particular, there is always the *zero homomorphism* sending every $m \in M$ to $0 \in M'$.

16.7. **Example.** An R-algebra homomorphism is the same as a ring homomorphism that is at the same time an R-module homomorphism.

A homomorphism of abelian groups is the same as a $\mathbb{Z}$-module homomorphism.

Let M be an R-module. Then there is a unique R-module homomorphism $0 \to M$ and a unique R-module homomorphism $M \to 0$; in both cases, it is the zero homomorphism.

16.8. **Proposition.** *Let M be an R-module, $M' \subset M$ a submodule. Then there is a unique R-module structure on the abelian group quotient M/M' that makes the quotient map $\phi : M \to M/M'$ an R-module homomorphism.*

M/M' is called the *quotient module* of M by M'.

Proof. Uniqueness is clear, as usual — we need to have

$$r(m + M') = r\phi(m) = \phi(rm) = rm + M'\,.$$

It remains to check that this is well-defined (if it is, it inherits the axioms from M). So let $m + M' = m' + M'$; then $m - m' \in M'$, and $rm - rm' = r(m - m') \in M'$ as well, hence $rm + M' = rm' + M'$. □

We have the usual isomorphism theorems.

16.9. **Proposition.**

(1) *If $\phi : M \to M'$ is an R-module homomorphism, then $\phi(M) \cong M/\ker\phi$.*
(2) *If $\phi : M \to M'$ is an R-module homomorphism and $N \subset M$ is a submodule contained in $\ker\phi$, then there is a unique R-module homomorphism $\psi : M/N \to M'$ such that $\psi(m + N) = \phi(m)$.*
(3) *If M is an R-module and $M_1 \subset M_2 \subset M$ are two submodules, then the natural R-module homomorphism $M/M_1 \to M/M_2$ is surjective with kernel M_2/M_1; in particular, $(M/M_1)/(M_2/M_1) \cong M/M_2$.*
(4) *If M is an R-module and $M_1, M_2 \subset M$ are two submodules, then the natural R-module homomorphism $M_1 \to (M_1 + M_2)/M_2$ is surjective with kernel $M_1 \cap M_2$; in particular, $M_1/(M_1 \cap M_2) \cong (M_1 + M_2)/M_2$.*

Proof. Use the corresponding statements for abelian groups and check that everything is compatible with the scalar multiplication. □

16.10. **Definition.** A diagram of R-modules and R-module homomorphisms

$$M_0 \xrightarrow{\phi_1} M_1 \xrightarrow{\phi_2} M_2 \xrightarrow{\phi_3} \cdots \xrightarrow{\phi_{n-1}} M_{n-1} \xrightarrow{\phi_n} M_n$$

is called *exact at* M_j $(1 \le j \le n-1)$ if the image of ϕ_j equals the kernel of ϕ_{j+1}. The diagram is called an *exact sequence* if it is exact at all M_j for $j = 1, \ldots, n-1$. An exact sequence of the form

$$0 \longrightarrow M' \xrightarrow{\alpha} M \xrightarrow{\beta} M'' \longrightarrow 0$$

(where 0 denotes the zero module) is called a *short exact sequence.* In this case, α is injective, β is surjective, so $\alpha(M') \cong M'$ and $M'' \cong M/\alpha(M')$.

As for vector spaces, we can define external and internal direct sums and direct products of R-modules

16.11. **Definition.** Let $(M_j)_{j\in J}$ be a family of R-modules. The *direct product* of the M_j is

$$\prod_{j\in J} M_j = \{(m_j)_{j\in J} : m_j \in M_j \text{ for all } j \in J\}\,,$$

with component-wise addition and scalar multiplication. We have natural *projection* homomorphisms

$$\pi_k : \prod_{j\in J} M_j \longrightarrow M_k\,, \quad (m_j)_{j\in J} \longmapsto m_k$$

for $k \in J$. If $J = \{1, \ldots, n\}$ is finite, we also write $M_1 \times \cdots \times M_n$ for the product.

The *(external) direct sum* of the M_j is the submodule

$$\bigoplus_{j\in J} M_j = \{(m_j)_{j\in J} \in \prod_{j\in J} M_j : m_j = 0 \text{ for all but finitely many } j \in J\}\,.$$

We have natural *inclusion* homomorphisms

$$\iota_k : M_k \longrightarrow \bigoplus_{j\in J} M_j\,, \quad m \longmapsto (m_j)_{j\in J} \quad \text{where } m_j = 0 \text{ for } j \ne k \text{ and } m_k = m$$

for $k \in J$. If $J = \{1, \ldots, n\}$ is finite, we also write $M_1 \oplus \cdots \oplus M_n$ for the direct sum.

Note that the direct sum and product coincide when the index set J is finite, but are distinct when J is infinite.

We have the usual universal properties.

16.12. **Proposition.** *Let $(M_j)_{j\in J}$ be a family of R-modules, M another R-module.*

(1) *If $\phi_j : M \to M_j$ are R-module homomorphisms, then there is a unique R-module homomorphism $\Phi : M \to \prod_{j\in J} M_j$ such that $\pi_j \circ \Phi = \phi_j$ for all $j \in J$.*

(2) *If $\phi_j : M_j \to M$ are R-module homomorphisms, then there is a unique R-module homomorphism $\Phi : \bigoplus_{j\in J} M_j \to M$ such that $\Phi \circ \iota_j = \phi_j$ for all $j \in J$.*

$$\begin{tikzcd}
M \arrow[r, "\Phi"] \arrow[dr, "\phi_j"'] & \prod\limits_{j\in J} M_j \arrow[d, "\pi_j"] \\
& M_j
\end{tikzcd}
\qquad
\begin{tikzcd}
M & \bigoplus\limits_{j\in J} M_j \arrow[l, "\Phi"'] \\
& M_j \arrow[u, "\iota_j"'] \arrow[ul, "\phi_j"]
\end{tikzcd}$$

Proof. As usual: uniqueness is clear; it remains to check that the definition of Φ that is forced gives a well-defined R-module homomorphism, which is easy. □

16.13. **Definition.** Let M be an R-module, $(M_j)_{j\in J}$ a family of submodules of M. M is the *internal direct sum* of the M_j if the natural R-module homomorphism $\bigoplus_{j\in J} M_j \to M$ is an isomorphism. In this case, we write $M = \bigoplus_{j\in J} M_j$.

A submodule $M' \subset M$ is called a *direct summand* of M if there is a *complement* $M'' \subset M$, i.e., a submodule M'' such that $M = M' \oplus M''$ is the internal direct sum of M' and M''.

16.14. **Remark.** M is the internal direct sum of submodules M_j, $j \in J$, if and only if

(1) $M = \sum_{j\in J} M_j$ is generated by the M_j, and
(2) for every $k \in J$, $M_k \cap \sum_{j\in J\setminus\{k\}} M_j = 0$.

(Exercise.)

16.15. **Example.** Let $M_1, M_2 \subset M$ be submodules of an R-module M. Then we have a short exact sequence

$$0 \longrightarrow M_1 \cap M_2 \xrightarrow{\alpha} M_1 \oplus M_2 \xrightarrow{\beta} M_1 + M_2 \longrightarrow 0$$

with the maps $\alpha : m \mapsto (m, -m)$ and $\beta : (m_1, m_2) \mapsto m_1 + m_2$.

16.16. **Examples.** If $R = F$ is a field (so that R-modules are F-vector spaces), then every submodule is a direct summand: we can pick a basis of the subspace and extend it to a basis of the full vector space. This is not true for general R-modules. For example, the $\mathbb{Z}$-submodules of $\mathbb{Z}$ are all of the form $n\mathbb{Z}$, but only 0 and $\mathbb{Z}$ itself are direct summands. (Exercise.)

The $\mathbb{Z}$-module $\mathbb{Z}/n\mathbb{Z}$ has (an isomorphic copy of) $\mathbb{Z}/m\mathbb{Z}$ as a direct summand if and only if $n = mm'$ with m and m' coprime. In particular, if $n = p_1^{e_1} \cdots p_k^{e_k}$ is the prime factorization of n, then

$$\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/p_1^{e_1}\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/p_k^{e_k}\mathbb{Z}\,.$$

(Compare the Chinese Remainder Theorem.) A similar statement is true for any PID in place of $\mathbb{Z}$.

17. Noetherian Modules

Similarly as for rings, we can consider a finiteness condition for modules, expressed in terms of submodules.

17.1. **Definition.** An R-module M is called *noetherian* if every increasing sequence of submodules of M

$$M_0 \subset M_1 \subset M_2 \subset M_3 \subset \dots$$

stabilizes, i.e., $M_n = M_{n+1} = \dots$ for some n.

17.2. **Proposition.** *Let M be an R-module. Then the following statements are equivalent.*

(1) *M is noetherian.*
(2) *Every non-empty set of submodules of M has a maximal element.*
(3) *Every submodule of M is finitely generated.*

Proof. Identical to the proof of Prop. 13.5. □

With this definition, a commutative ring R is noetherian as a ring if and only if it is noetherian as an R-module.

17.3. **Proposition.** *Let*

$$0 \longrightarrow M' \xrightarrow{\alpha} M \xrightarrow{\beta} M'' \longrightarrow 0$$

be a short exact sequence of R-modules. Then M is noetherian if and only if both M' and M'' are noetherian.

Proof. First assume that M' and M'' are both noetherian. Let $M_0 \subset M_1 \subset \dots$ be an increasing sequence of submodules of M. Then $\beta(M_0) \subset \beta(M_1) \subset \dots$ is an increasing sequence of submodules of M'' and $\alpha^{-1}(M_0) \subset \alpha^{-1}(M_1) \subset \dots$ is an increasing sequence of submodules of M'. By assumption, both become stationary, so there is some n such that $\beta(M_n) = \beta(M_{n+1}) = \dots$ and $\alpha^{-1}(M_n) = \alpha^{-1}(M_{n+1}) = \dots$. Now I claim that if $N \subset N'$ are two submodules of M such that $\beta(N) = \beta(N')$ and $\alpha^{-1}(N) = \alpha^{-1}(N')$, then $N = N'$. This implies that $M_n = M_{n+1} = \dots$, so the sequence of submodules of M also stabilizes.

To prove the claim, let $x \in N'$. Since $\beta(N) = \beta(N')$, there is some $y \in N$ such that $\beta(x) = \beta(y)$, so $x - y \in \ker\beta \cap N' = \alpha(\alpha^{-1}(N'))$. Since $\alpha^{-1}(N) = \alpha^{-1}(N')$, it follows that $x - y \in \alpha(\alpha^{-1}(N)) \subset N$, and finally that $x = y + (x - y) \in N$.

Now assume that M is noetherian. Let $M''_0 \subset M''_1 \subset \dots$ be an increasing sequence of submodules of M''. Then $\beta^{-1}(M''_0) \subset \beta^{-1}(M''_1) \subset \dots$ is an increasing sequence of submodules of M, which by assumption stabilizes. Since $\beta(\beta^{-1}(M''_n)) = M_n$, the original sequence of submodules of M'' also stabilizes.

Let $M'_0 \subset M'_1 \subset \dots$ be an increasing sequence of submodules of M'. Then $\alpha(M'_0) \subset \alpha(M'_1) \subset \dots$ is an increasing sequence of submodules of M, which by assumption stabilizes. Since $\alpha^{-1}(\alpha(M'_n)) = M_n$, the original sequence of submoudles of M' stabilizes as well. □

In particular, every submodule and every quotient module of a noetherian module is again noetherian.

17.4. **Theorem.** *Let R be a noetherian commutative ring, M a finitely generated R-module. Then M is a noetherian R-module.*

Proof. Since M is finitely generated, say by $m_1, \dots, m_n$, M is a quotient of the free module R^n, for some n. It therefore suffices to pove the claim for $M = R^n$. This now follows by induction n, the case $n = 1$ being the assumption that R is a noetherian ring, using the obvious exact sequences

$$0 \longrightarrow R^{n-1} \longrightarrow R^n \longrightarrow R \longrightarrow 0 .$$

□

17.5. **Corollary.** *Let R be a PID and M a finitely generated R-module. Then every R-submodule of M is again finitely generated.*

Proof. As a PID, R is noetherian, so M is noetherian as well. □

18. Finitely Generated Modules over Principal Ideal Domains

We have seen at the end of the last section that finitely generated modules over a PID R are noetherian. In particular, every submodule of the free module R^n is finitely generated. In this section, we want to study the structure of submodules and quotients of R^n in more detail. The main result will be that submodules of R^n are again free modules of rank at most n and that quotients of R^n are products of at most n cyclic modules.

In this section, R will *always be a PID.*

For the proof, we need a few results about matrices over R.

18.1. **Definition.** Two $m \times n$ matrices A and A' with entries in R are called *equivalent* if there are invertible matrices $U \in \mathrm{GL}_m(R)$ and $V \in \mathrm{GL}_n(R)$ such that $A' = UAV$. We write $A \sim A'$. This is clearly an equivalence relation.

We write $\gcd(A)$ for a gcd of all the entries of A. More generally, we write $\gcd_r(A)$ for a gcd of all $r \times r$ *minors* of A, i.e., determinants of $r \times r$ matrices obtained by extracting any choice of r rows and columns from A.

18.2. **Remarks.** We can think of the matrix A as specifying an R-linear map $R^n \to R^m$ (identifying the elements of R^n and R^m with column vectors of the appropriate length). Then multiplying A on the left or right with invertible matrices corresponds to changing bases in the two free modules. So two matrices are equivalent if and only if they describe the same map with respect to suitable bases. Compare with the situation in Linear Algebra.

Our goal in the following will be to find a normal form for matrices with respect to equivalence. Over a field, the normal form is diagonal (i.e., with zero entries off the main diagonal), with diagonal entries $1, \dots, 1, 0, \dots, 0$. What we do here is a generalization of this to PIDs instead of fields.

Note that (as in Linear Algebra), elementary row and column operations on A (swapping two rows/columns, multiplying a row/column by a unit, adding a multiple of a row/column to another) correspond to multiplying A on the left/right by certain invertible matrices.

18.3. **Lemma.** *If $A \sim A'$, then $\gcd_r(A) \sim \gcd_r(A')$ for all r.*

Proof. Exercise. □

18.4. **Lemma.**

(1) *Let $A = (a\ b)$. Then A is equivalent to $A' = (g\ 0)$ where $g = \gcd(A)$.*
(2) *Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then A is equivalent to $A' = \left(\begin{smallmatrix} g & 0 \\ 0 & h \end{smallmatrix}\right)$ where $g = \gcd(A)$.*
(3) *Let A be any $m \times n$ matrix over R, with $m, n \geq 1$. Then A is equivalent to a matrix whose upper left entry is a gcd of the entries of A.*

Proof.
(1) There are $u, v \in R$ such that $ua + vb = g$. Write $a = ga'$, $b = gb'$. We set $U = (1)$ and $V = \left(\begin{smallmatrix} a' & b' \\ -v & u \end{smallmatrix}\right)$, then $U(g\ 0)V = (a\ b)$.

(2) Among all matrices equivalent to A, consider one with upper left entry minimal with respect to divisibility, say $\left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right)$. We must have that a' divides b' and c'; otherwise we could apply part (1) (or its transpose) to get a smaller upper left entry $\gcd(a', b')$ (or $\gcd(a', c')$). But then we can add a suitable multiple of the first row/column to the second row/column (this amounts to multiplying on the left/right by invertible matrices) and obtain an equivalent matrix $\left(\begin{smallmatrix} a' & 0 \\ 0 & h \end{smallmatrix}\right)$. If a' does not divide h, then adding the second row to the first and then performing column operations, we get a matrix with left upper entry $\gcd(a', h)$, which is smaller than a'. This is not possible by our choice of a', hence a' divides h and so $a' = \gcd(a', h) = \gcd(a, b, c, d)$.

(3) As in the proof of part (2), consider an equivalent matrix A' with minimal upper left entry a (w.r.t. divisibility). If a is not a gcd of the entries of A', then there is a 1×2 or a 2×1 or a 2×2 submatrix of A' with left upper entry a that has the same property. But then (by parts (1) or (2)) we can find an equivalent matrix with a smaller left upper entry, a contradiction. □

18.5. **Definition.** An $m \times n$ matrix $A = (a_{ij})$ is called *diagonal* if $a_{ij} = 0$ for all $i \neq j$. We write $A = \operatorname{diag}_{mn}(a_{11}, a_{22}, \ldots, a_{kk})$, where $k = \min\{m, n\}$. (This is ad-hoc notation.)

18.6. **Proposition.** *Let A be an $m \times n$ matrix with entries in R. Then A is equivalent to a diagonal $m \times n$ matrix $A' = \operatorname{diag}_{mn}(d_1, \ldots, d_k)$ (where $k = \min\{m, n\}$) such that $d_1 \mid d_2 \mid \ldots \mid d_k$. The sequence of diagonal entries $(d_i)_i$ is uniquely determined up to associates.*

The diagonal entries $d_1, d_2, \ldots, d_k$ with $k = \min\{m, n\}$ are called the *elementary divisors* of A.

Proof. To prove existence, we proceed by induction on k. There is nothing to prove if $k = 0$. Otherwise, by Lemma 18.4, A is equivalent to a matrix whose upper left entry is a gcd d_1 of the entries of A. We can then perform row and column operations to make all other entries of the first row and column zero, so

$$A \sim \left(\begin{array}{c|c} d_1 & 0 \\ \hline 0 & A_1 \end{array}\right)$$

with some $(m-1) \times (n-1)$ matrix A_1. By induction, $A_1 \sim \operatorname{diag}_{m-1,n-1}(d_2, \ldots, d_k)$ with $d_2 \mid \ldots \mid d_k$, so $A \sim \operatorname{diag}_{mn}(d_1, d_2, \ldots, d_k)$. Note also that $d_1 \mid d_2$, since d_1 divides all entries of A_1.

It remains to show uniqueness. Assume that

$$A = \operatorname{diag}_{mn}(d_1, \dots, d_k) \sim \operatorname{diag}_{mn}(d'_1, \dots, d'_k) = A'$$

and $d_1 \mid \dots \mid d_k$, $d'_1 \mid \dots \mid d'_k$. Then $d_1 d_2 \cdots d_r = \gcd_r(A) \sim \gcd_r(A') = d'_1 d'_2 \cdots d'_r$ for all $1 \le r \le k$. This implies $d_j \sim d'_j$ for $1 \le j \le k$ (note that $d_r = 0$ implies $d_j = 0$ for $j \ge r$). □

18.7. **Example.** Let us see how this works in a specific example. Consider

$$M = \begin{pmatrix} -3 & 1 & 0 & 1 \\ 1 & -5 & 0 & 2 \\ 0 & 0 & -2 & 1 \\ 1 & 2 & 1 & -2 \end{pmatrix}$$

over $R = \mathbb{Z}$. We can interchange the first two rows in order to get a 1 into the upper left corner, and then clear out the first row and column. This gives

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -14 & 0 & 7 \\ 0 & 0 & -2 & 1 \\ 0 & 7 & 1 & -4 \end{pmatrix} .$$

We now move the 1 in the remaining 3×3 matrix into its upper left corner (by interchanging rows 2 and 4 and columns 2 and 3) and then clear out row and column 2:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 14 & -7 \\ 0 & 0 & -14 & 7 \end{pmatrix}$$

Now we move 7 to the next diagonal position and clear out, leading to

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

so the elementary divisors of M are $1, 1, 7, 0$.

18.8. **Remark.** The result is valid for any PID; however in order to *compute* the elementary divisors of a matrix, you need to be able to express the gcd of two elements $a, b \in R$ as a linear combination of a and b. This is possible in *Euclidean* rings with an extension of the Euclidean Algorithm, but may be hard in more general PIDs.

18.9. **Definition.** An element $(r_1, \dots, r_n) \in R^n$ is called *primitive* if we have $\gcd(r_1, \dots, r_n) = 1$. Note that this is equivalent to saying that $Rr_1 + \dots + Rr_n = R$.

18.10. **Corollary.** *The group $\mathrm{GL}_n(R) \cong \mathrm{Aut}_R(R^n) = \mathrm{End}_R(R^n)^\times$ of invertible $n \times n$ matrices over R acts transitively on the set of primitive elements of R^n. More generally, two elements are in the same orbit if and only if the gcds of their entries agree.*

Proof. Identifying elements of R^n with column vectors, i.e., $n \times 1$ matrices, this is just a special case of Prop. 18.6. □

18.11. **Corollary.** *Given a primitive element* $(r_1, \dots, r_n) \in R^n$*, there exists a matrix* $A \in \mathrm{GL}_n(R)$ *whose first column is* $(r_1, \dots, r_n)$*.*

Proof. By the preceding corollary, there is $A \in \mathrm{GL}_n(R)$ such that

$$A \cdot \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} .$$

But this means that that $(r_1, \dots, r_n)$ is the first column of A. □

18.12. **Corollary.** *Let* $M \subset R^n$ *be a submodule. Then* M *is free of rank* $\leq n$*.*

Proof. We know that M is finitely generated. Let $(r_{1j}, \dots, r_{nj})$ be generators of M, $j = 1, \dots, m$. Consider the matrix A whose columns are $(r_{1j}, \dots, r_{nj})$. By Prop. 18.6 above, there are invertible matrices U and V such that $A' = UAV$ is diagonal. The submodule M' of R^n generated by the columns of A' is obviously free of rank $\leq n$. Since V is invertible, this is the same as the submodule generated by the columns of UA. Since U is an automorphism of R^n, M is isomorphic to M' and therefore also free of rank $\leq n$. □

18.13. **Corollary.** *Let* M *be an* R*-module generated by* n *elements. Then* M *is isomorphic to a direct product (or sum) of at most* n *cyclic modules:*

$$M \cong R/d_1R \times R/d_2R \times \cdots \times R/d_mR$$

with $m \leq n$ *and such that* d_1 *is not a unit and* $d_1 \mid d_2 \mid \ldots \mid d_m$*. The sequence of the* d_j*'s is uniquely determined up to associates.*

Proof. Let $m_1, \dots, m_n$ be generators of M and let $\phi : R^n \to M$ be the R-linear map that sends $(r_1, \dots, r_n)$ to $r_1m_1 + \ldots + r_nm_n$. The kernel of ϕ is a submodule of R^n, and by considerations similar to the ones in the proof of the preceding corollary, we can assume that $\ker \phi$ is generated by $d_1e_1, \dots, d_me_m$ (where the e_j are the standard basis elements of R^n) with $m \leq n$ and $d_1 \mid d_2 \mid \ldots \mid d_m$. Let k be the smallest index such that d_k is not a unit. Then $M \cong R^n / \ker \phi$ is isomorphic to $R/d_kR \times \ldots \times R/d_mR \times R^{n-m}$. Uniqueness follows from the uniqueness statement in Prop. 18.6. □

18.14. **Definition.** Let M be an R-module. An element $m \in M$ is *torsion* if there is $r \in R \setminus \{0\}$ such that $rm = 0$. M is called *torsion free* if the only torsion element of M is 0.

18.15. **Corollary.** *A finitely generated torsion free* R*-module is free.*

Proof. In the preceding corollary, the product of cyclic modules is torsion free if and only if all $d_j = 0$; then $M \cong R^m$. □

Note that the assumption on finite generation is necessary. For example, $\mathbb{Q}$ as a $\mathbb{Z}$-module is torsion free, but not free. (Which proves that $\mathbb{Q}$ is not a finitely generated abelian group.)

Now we want to apply these results to prove two important theorems. One is the classification theorem for finitely generated abelian groups; the other is the Jordan Normal Form Theorem for matrices.

18.16. **Theorem.** *Let G be a finitely generated abelian group. Then there are uniquely determined integers $d_1, \dots, d_m$ with $d_1 > 1$ and $d_1 \mid d_2 \mid \dots \mid d_m$ and an integer $r \geq 0$ such that*

$$G \cong \mathbb{Z}/d_1\mathbb{Z} \times \mathbb{Z}/d_2\mathbb{Z} \times \dots \times \mathbb{Z}/d_m\mathbb{Z} \times \mathbb{Z}^r .$$

G is finite if and only if $r = 0$, and G is a free abelian group if and only if $m = 0$.

Proof. Apply Corollary 18.13 to the case $R = \mathbb{Z}$ and note that every nonzero integer is associate to a unique positive integer. □

If we use that $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \dots \times \mathbb{Z}/p_k^{e_k}\mathbb{Z}$ when $n = p_1^{e_1} \cdots p_k^{e_k}$ is the prime factorization, then we can state the classification theorem also in the following form.

Every finitely generated abelian group is a product of finitely many cylic groups of prime order and a finitely generated free abelian group.

A similar statement is true over a general PID.

18.17. **Theorem.** *Let M be a finitely generated R-module. Then M is isomorphic to a direct product (or sum) of a finitely generated free R-module and finitely many cyclic modules of the form R/p^eR, where p is a prime element of R and $e \geq 1$.*

Now we want to use our results in order to prove the Jordan Normal Form Theorem. Recall its statement.

18.18. **Theorem.** *Let F be a field, $M \in \mathrm{Mat}_n(F)$ a matrix such that the characteristic polynomial f factors into linear factors. Then there is an invertible matrix $T \in \mathrm{GL}_n(F)$ such that $T^{-1}MT$ is a block diagonal matrix whose blocks are "Jordan blocks" of the form*

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

Proof. The statement amounts to saying that we can write the vector space F^n as a direct sum of subspaces V_j such that each V_j is invariant under M, and the action of M on V_j with respect to a suitable basis $v_{j1}, \dots, v_{jk}$ is given by a Jordan block, i.e.,

$$Mv_{jk} = \lambda v_{jk} + v_{j,k-1} , \quad \dots , \quad Mv_{j2} = \lambda v_{j2} + v_{j1} , \quad Mv_{j1} = \lambda v_{j1} .$$

In order to show this, we exhibit F^n as a finitely generated $F[x]$-module. We have an F-algebra homomorphism $F[x] \to \mathrm{Mat}_n(F)$ that sends x to M, and we know by the Cayley-Hamilton Theorem that the characteristic polynomial f is in the kernel. We can therefore consider F^n as an $F[x]$-module, in which every element is "killed" by f: if $v \in F^n$, then $fv = f(M) \cdot v = 0$. Since F^n is finitely generated as a vector space, it is also finitely generated as an $F[x]$-module (a vector space basis will generate F^n also as an $F[x]$-module). By Theorem 18.17, we can write

$$F^n \cong F[x]/\langle p_1^{e_1} \rangle \oplus \dots \oplus F[x]/\langle p_k^{e_k} \rangle .$$

(There is no free part, since f kills everything.) This induces a splitting $F^n = V_1 \oplus \dots \oplus V_k$ of F^n as an internal direct sum of $F[x]$-submodules $V_j \cong F[x]/\langle p_j^{e_j} \rangle$.

Note also that every p_j must divide f, and since f splits into linear factors, every p_j must be of the form $x - \lambda$, with λ an eigenvalue of M. So we only have to study the action of x on $F[x]$-modules of the form $M_{\lambda,e} = F[x]/\langle (x-\lambda)^e \rangle$. Now division with remainder shows that the images of $1, x, \dots, x^{e-1}$ form an F-basis of $M_{\lambda,e}$. But then, $v_1 = (x-\lambda)^{e-1}, v_2 = (x-\lambda)^{e-2}, \dots, v_e = 1$ (modulo $(x-\lambda)^e$) also form an F-basis, and the action of x on the v_j is

$$xv_1 = \lambda v_1\,, \quad xv_2 = \lambda v_2 + v_1\,, \quad \dots\,, \quad xv_e = \lambda v_e + v_{e-1}\,.$$

□

19. Roots of Polynomials

In this section, F will always be a field. Note that any ring homomorphism $F \to R$, where R is not the zero ring, is injective (the kernel is an ideal of F and not all of F, so it must be the zero ideal). Therefore, we can identify F with a subring of the F-algebra R.

19.1. **Definition.** Let $f \in F[x]$ be a polynomial, and let $R \neq 0$ be an F-algebra. An element $r \in R$ is called a *root* or *zero* of f if $f(r) = 0$.

Recall that $f(r)$ is the image of f under the unique F-algebra homomorphism $F[x] \to R$ that sends x to r.

19.2. **Lemma.** *Let $f \in F[x]$ be a polynomial, $a \in F$. Then a is a root of f if and only if f is divisible by $x - a$.*

Proof. We use polynomial division, see Prop. 12.10: there is a constant polynomial r and a polynomial q such that $f(x) = q(x)(x-a) + r$. Evaluating this equation at a, we get $f(a) = r$. Hence a is a root of f iff $f(a) = 0$ iff $r = 0$ iff $x - a$ divides f. □

19.3. **Proposition.** *Let $0 \neq f \in F[x]$ be a polynomial of degree n. Then f has at most n distinct roots in F.*

Proof. By induction on n. The claim is clear for constant polynomials ($n = 0$). Otherwise, assume that $n > 0$ and that f has at least one root $a_1 \in F$. Then by the previous lemma, $f = (x - a_1)g$ with some polynomial g of degree $n - 1$. By induction, g has at most $n - 1$ distinct roots in F. Now $0 = f(a) = (a - a_1)g(a)$ implies that $a = a_1$ or a is a root of g, so f can have at most n distinct roots in F. □

Note that the statement is equivalent to the following.

19.4. **Corollary.** *Let $f \in F[x]$ be a polynomial of degree at most n. If f has at least $n + 1$ distinct roots in F, then f is the zero polynomial.*

19.5. **Examples.** The argument works more generally for integral domains in place of F (or use the field of fractions). However, it is essential that there are no zero divisors and that the ring is commutative, as the following examples show.

(1) Let R be the ring $\mathbb{Z}/8\mathbb{Z}$ and consider $f = x^2 - 1 \in R[x]$. Then f has the four distinct roots $1, 3, 5, 7$ in R.

(2) Now consider the quaternions $\mathbb{H}$ and let $f = x^2 + 1 \in \mathbb{H}[x]$. Then f has the six distinct roots $\pm i, \pm j, \pm k$ in $\mathbb{H}$. In fact, the roots of f in $\mathbb{H}$ are exactly the elements $ai + bj + ck$ with $a^2 + b^2 + c^2 = 1$, so there are (uncountably) infinitely many roots!

(Exercise: where does the proof go wrong when F is only a skew field?)

19.6. **Corollary.** *If the polynomial $f \in F[x]$ has the distinct roots $a_1, \dots, a_m$ in F, then f is divisible by $(x - a_1)\cdots(x - a_m)$. If f is monic of degree n and has the n distinct roots $a_1, \dots, a_n$ in F, then $f = \prod_{j=1}^{n}(x - a_j)$.*

Proof. The first statement follows by induction from Lemma 19.2: we can write $f = (x - a_1)f_1$; then $0 = (a_j - a_1)f_1(a_j)$, so $a_2, \dots, a_m$ are roots of f_1, and by induction, f_1 is divisible by $(x - a_2)\cdots(x - a_m)$.

As to the second statement, we know from the first part that f is divisible by the right hand side. But both sides are monic polynomials of the same degree, hence their quotient is 1. □

Now it is perhaps surprising that the seemingly innocuous result of Prop. 19.3 has a very interesting consequence.

19.7. **Theorem.** *Let $G \subset F^\times$ be a finite subgroup. Then G is cyclic.*

Proof. By Thm. 18.16, the classification theorem for finitely generated abelian groups, $G \cong \mathbb{Z}/d_1\mathbb{Z} \times \ldots \times \mathbb{Z}/d_k\mathbb{Z}$ with positive integers $1 < d_1 \mid d_2 \mid \ldots \mid d_k$. This implies that for all $\alpha \in G$, $\alpha^{d_k} = 1$ (recall that the group operation in G is written multiplicatively). So all elements of G are roots of the polynomial $x^{d_k} - 1$. By Prop, 19.3, $\#G \le d_k$. But $\#G = d_1 d_2 \cdots d_k$, so this implies that $k = 1$, and $G \cong \mathbb{Z}/d_1\mathbb{Z}$ is cyclic. □

19.8. **Corollary.** *Let F be a finite field, $n = \#F$. Then $F^\times$ is a cyclic group, and all elements of F are roots of $f = x^n - x$. In particular,*

$$x^n - x = \prod_{a \in F}(x - a) \qquad \text{in } F[x].$$

Proof. By the previous theorem, $F^\times$ is cyclic (as a finite subgroup of $F^\times$). Since the order of $F^\times$ is $n-1$, all elements $a \in F \setminus \{0\}$ are roots of $x^{n-1} - 1$. Therefore, all elements of F are roots of $x(x^{n-1} - 1) = x^n - x$. The last statement follows from Cor. 19.6. □

We will study finite fields more closely in a later section.

If $a \in F$ is an element of a finite subgroup of $F^\times$, then $a^n = 1$ for some $n \ge 1$. Such elements have a special name.

19.9. **Definition.** An element $a \in F$ such that $a^n = 1$ for some $n \geq 1$ is called an *nth root of unity.* It is a *primitive nth root of unity* if $a^m \neq 1$ for all $1 \leq m < n$ (i.e., if $o(a) = n$ in $F^\times$).

The nth roots of unity form a (finite) subgroup of $F^\times$ denoted by $\mu_n(F)$ or just μ_n in case $\#\mu_n(F) = n$. This group $\mu_n(F)$ is cyclic of order dividing n; if the order is n, then $\mu_n(F)$ contains primitive nth roots of unity, and these are exactly the generators of $\mu_n(F)$.

19.10. **Example.** What are the nth roots of unity in $\mathbb{C}$? Using polar coordinates, we have $(re^{i\phi})^n = r^n e^{ni\phi}$ (where $r > 0$ and $\phi \in \mathbb{R}/2\pi\mathbb{Z}$). This equals 1 if and only if $r = 1$ and $n\phi \in 2\pi\mathbb{Z}$. Therefore, there are exactly the n solutions

$$1, \quad e^{2\pi i/n}, \quad e^{4\pi i/n}, \quad \ldots, \quad e^{2(n-1)\pi i/n}.$$

In the complex plane, they are the vertices of a regular n-gon centered at the origin and with one vertex at 1. The nth root of unity $e^{2k\pi i/n}$ is primitive if and only if $\gcd(k, n) = 1$.

19.11. **Example.** If F is a finite field of size n, then $F^\times = \mu_{n-1}(F)$. For example, we have all fourth roots of unity in $\mathbb{F}_5 = \mathbb{Z}/5\mathbb{Z}$. The two primitive ones are 2 and 3; indeed $2^4 \equiv 3^4 \equiv 1 \bmod 5$, but $2^2 = 4$ and $3^2 = 9$ are not $\equiv 1$.

20. Algebraic Field Extensions

In this section, we will consider the relations between fields. F will continue to denote a field.

20.1. **Definition.** A *field extension* of F is another field F', together with a field (= ring) homomorphism $F \to F'$. We frequently write F'/F to indicate that F' is a field extension of F (and hope that it does not lead to confusion with the other uses of "/").

Note that F' is then an F-algebra, in particular, F' is an F-vector space (i.e., an F-module), and we can talk about its dimension over F. The extension F'/F is called *finite* if F' is a finite-dimensional F-vector space. In this case the dimension is called the *degree* of the extension and written $[F' : F]$. Extensions of degree 2 are called *quadratic*, extensions of degree 3, *cubic*, and so on.

We frequently identify F with its image in F'; then the homomorphism is just inclusion $F \subset F'$.

20.2. **Examples.** F/F (with the identity map) is a finite field extension; its degree is 1. A field extension of degree 1 is called *trivial.*

$\mathbb{C}/\mathbb{R}$ is a quadratic field extension. $\mathbb{R}/\mathbb{Q}$ is an infinite field extension.

20.3. **Lemma.** *If $F''/F'/F$ is a chain of two field extensions, then F''/F is finite if and only if F''/F' and F'/F are both finite. In this case,*

$$[F'' : F] = [F'' : F'][F' : F].$$

Furthermore, if $(a_i)_{i\in I}$ is an F-basis of F' and $(b_j)_{j\in J}$ is an F'-basis of F'', then $(a_i b_j)_{(i,j)\in I\times J}$ is an F-basis of F''.

Proof. If F''/F is finite, then clearly F''/F' and F'/F are both finite. Now assume that F''/F' and F'/F are both finite, and let $(a_i)_{i\in I}$ be an F-basis of F' and $(b_j)_{j\in J}$ an F'-basis of F'' (so I and J are finite). Then a_ib_j (for $i \in I$, $j \in J$) generate F'' as an F-vector space, hence F'' has finite dimension over F. (Given $\alpha \in F''$, we can write α as a linear combination of the b_j with coefficients in F'; then we can write the coefficients as linear combinations of the a_i with coefficients in F. Expanding out shows that α is a linear combination of the a_ib_j with coefficients in F.)

We have to show that $\dim_F F'' = \dim_{F'} F'' \cdot \dim_F F' = \#J\#I$. The argument above shows $\leq$, so we only have to show that the a_ib_j are linearly independent over F. So assume there are $\lambda_{ij} \in F$ such that $\sum_{j\in J}\sum_{i\in I}\lambda_{ij}a_ib_j = 0$. Since the b_j are linearly independent over F', this implies that $\sum_{i\in I}\lambda_{ij}a_i = 0$ for all $j \in J$. Now since the a_i are linearly independent over F, this in turn implies that all $\lambda_{ij} = 0$. □

20.4. **Definition.** Let K be a field. A *subfield* of K is a subring F of K that is a field. Then K/F is a field extension, and K therefore an F-algebra. It is clear that the intersection of arbitrary collections of subfields is again a subfield. Therefore the following definitions make sense.

Let $F \subset K$ be a subfield, $\alpha_1, \dots, \alpha_n \in K$. We denote by $F[\alpha_1, \dots, \alpha_n]$ the F-subalgebra of K generated by $\alpha_1, \dots, \alpha_n$, and by $F(\alpha_1, \dots, \alpha_n)$ the subfield of K generated by F and $\alpha_1, \dots, \alpha_n$.

Note that $F[\alpha_1, \dots, \alpha_n]$ is the image of the polynomial ring $F[x_1, \dots, x_n]$ under the F-algebra homomorphism to K that sends x_j to α_j.

Next, we look at elements in field extensions.

20.5. **Definition.** Let F'/F be a field extension and $\alpha \in F'$. We call α *algebraic* over F if α is a root of a nonzero polynomial $f \in F[x]$. If α is not algebraic over F, it is called *transcendental* over F. The extension F'/F is called *algebraic* if all elements $\alpha \in F'$ are algebraic over F; otherwise it is *transcendental.*

F is called *algebraically closed in F'* if the only elements of F' that are algebraic over F are those of F. F is called *algebraically closed* if F does not have nontrivial algebraic extensions.

20.6. **Lemma.** *Let F'/F be a field extension and $\alpha \in F'$ an algebraic element. Then there is a monic polynomial $f \in F[x]$ such that $f(\alpha) = 0$ and such that f divides every polynomial $g \in F[x]$ that has α as a root. The polynomial f is irreducible.*

If $g \in F[x]$ is monic and irreducible and $g(\alpha) = 0$, then $g = f$. In particular, f is uniquely determined.

This polynomial f is called the *minimal polynomial* of α over F. If $\deg f = n$, then α is called algebraic *of degree n* over F.

Proof. Consider the F-algebra homomorphism $\phi : F[x] \to F'$ that sends x to α. Its kernel is not the zero ideal, since by assumption there is some $g \in F[x]$ with $g(\alpha) = 0$. Since $F[x]$ is a PID, the kernel is a principal ideal; it is then generated by a unique monic polynomial f. Now α is a root of $g \in F[x]$ if and only if $g \in \ker\phi$, if and only if f divides g. Since the image of ϕ is an integral domain, $\ker\phi$ is a (nonzero) prime ideal, and therefore f is irreducible.

For the last statement, we certainly have that f divides g. But f is not constant and g is irreducible, so $f \sim g$. Since both are monic, $f = g$. □

20.7. **Examples.** Consider $i \in \mathbb{C}$, where $\mathbb{C}$ is a field extension of $\mathbb{R}$. The minimal polynomial is then $x^2 + 1$. More generally, if $\alpha = a + bi \in \mathbb{C}$ with $b \neq 0$, then its minimal polynomial over $\mathbb{R}$ is $x^2 - 2a\,x + a^2 + b^2$.

Since i is a root of $x^2 + 1$, it is also algebraic over $\mathbb{Q}$. More generally, if $\zeta \in \mathbb{C}$ is an nth root of unity, then ζ is a root of $x^n - 1$, and so ζ is algebraic over $\mathbb{Q}$.

By a counting argument from set theory, "most" real (or complex) numbers are transcendental (over $\mathbb{Q}$), but it is hard to prove that for specific interesting numbers. Lindemann's proof in 1882 that π is transcendental was a big achievement. The number e is also transcendental (proved by Hermite a decade earlier), but it is not known if both $e + \pi$ and $e\pi$ are transcendental (they cannot both be algebraic). On the other hand, e^π is known to be transcendental.

20.8. **Lemma.** *Let F'/F be a field extension, $\alpha \in F'$. The following statements are equivalent.*

1. *α is algebraic over F.*
2. *$\dim_F F[\alpha]$ is finite.*
3. *$F[\alpha] = F(\alpha)$.*
4. *α is contained in a finite subextension of F'/F.*

Proof. Let $\phi : F[x] \to F'$ be the F-algebra homomomorphism sending x to α. Then the image of ϕ is $F[\alpha]$, and the kernel of ϕ is a principal ideal $\langle f \rangle$. Since the image of ϕ is an integral domain, the kernel is a prime ideal, hence $f = 0$, or f is irreducible. We then have the following equivalences.

$$\begin{aligned}
\dim_F F[\alpha] < \infty &\iff f \neq 0 \\
&\iff \alpha \text{ is algebraic over } F \\
F[\alpha] = F(\alpha) &\iff F[\alpha] \text{ is a field} \\
&\iff \ker\phi \text{ is a maximal ideal} \\
&\iff f \text{ is irreducible} \\
&\iff f \neq 0
\end{aligned}$$

(For the first, note that $\dim_F F[x]/\langle f \rangle = \deg f$ if $f \neq 0$, but $\dim_F F[x] = \infty$.)

This proves the equivalence of the first three statements. Together, they clearly imply (4): $\alpha \in F(\alpha) = F[\alpha]$ is a finite subextension. Conversely, (4) obviously implies (2). □

20.9. **Corollary.** *If F'/F is a finite field extension, then F'/F is algebraic.*

20.10. **Corollary.** *Let F'/F be a field extension.*

1. *Let $\alpha, \beta \in F'$. If α is algebraic over F and β is algebraic over $F(\alpha)$, then β is algebraic over F.*
2. *The set of all elements of F' that are algebraic over F form a subfield of F' containing F: if α and β are algebraic, then $\alpha + \beta$ and $\alpha\beta$ are algebraic, and if $\alpha \neq 0$, then α^{-1} is algebraic.*

Proof.
(1) By assumption and Lemma 20.8, $F(\alpha,\beta) = F(\alpha)(\beta)$ is a finite extension of $F(\alpha)$ and $F(\alpha)$ is a finite extension of F. By Lemma 20.3, $F(\alpha,\beta)/F$ is finite, and so by Lemma 20.8 again, $\beta \in F(\alpha,\beta)$ is algebraic.

(2) Let $\alpha, \beta \in F'$ be algebraic over F. By the reasoning above, $F(\alpha,\beta)$ is finite over F, and so all its elements are algebraic, in particular $\alpha+\beta$, $\alpha\beta$ and α^{-1} (if the latter is defined). □

20.11. **Example.** The element $2\cos\frac{2\pi}{7} \in \mathbb{R}$ is algebraic over $\mathbb{Q}$ — it is a zero of $x^3 + x^2 - 2\,x - 1$, which is also its minimal polynomial, since it is monic and irreducible (it has no rational root). The fact that $\alpha = 2\cos\frac{2\pi}{7}$ is algebraic can be seen from $\alpha = \zeta + \zeta^{-1}$ where $\zeta = e^{2\pi i/7} \in \mathbb{C}$ is a seventh root of unity and therefore algebraic. ζ satisfies $\zeta^7 - 1 = 0$, $\zeta \neq 1$, so

$$\zeta^6 + \zeta^5 + \zeta^4 + \zeta^3 + \zeta^2 + \zeta + 1 = 0\,,$$

which can be rewritten in the form

$$(\zeta + \zeta^{-1})^3 + (\zeta + \zeta^{-1})^2 - 2(\zeta + \zeta^{-1}) - 1 = 0\,.$$

20.12. **Corollary.** *If F'/F and F''/F' are algebraic extensions, then so is F''/F.*

Proof. Let $\alpha \in F''$. Then α is algebraic over F', so it is a root of some polynomial $f \in F'[x]$. The coefficients of f are all algebraic over F. By an argument similar to that in the proof of the corollary above, the F-subalgebra K generated by the coefficients is of finite F-dimension. But now $K(\alpha)/K$ is finite and K/F is finite, so $K(\alpha)/F$ is also finite, hence α is algebraic over F. □

We now come to a very important construction that provides us with algebraic field extensions.

20.13. **Proposition.** *Let $f \in F[x]$ be an irreducible polynomial of degree n. Then $F' = F[x]/\langle f\rangle$ is a field extension of F of degree n, in which f has a root.*

If K/F is a field extension such that $a \in K$ is a root of f, then there is a unique field homomorphism $F' \to K$ sending the image of x in F' to a.

This construction is called *adjoining a root of f to F.*

Proof. Since f is irreducible, $\langle f\rangle$ is a maximal ideal of $F[x]$ and therefore the quotient ring $F' = F[x]/\langle f\rangle$ is a field. It is an F-algebra in a natural way and has F-dimension $n = \deg f$. Let $\alpha \in F'$ be the image of x. Then $f(\alpha) = 0$, since the evaluation homomorphism $F[x] \to F'$ that sends x to α is the natural quotient map, and f is in its kernel.

Now suppose K/F is a field extension and $a \in K$ is a root of f. The evaluation homomorphism $F[x] \to K$ that sends x to a then has kernel generated by f, therefore it induces an F-algebra homomorphism $F' \to K$ that sends the image α of x to a. □

20.14. **Examples.** This proposition allows us to construct fields like $\mathbb{Q}(\sqrt[3]{2})$ without the need to define them as subfields of an already known field like $\mathbb{R}$ or $\mathbb{C}$. We simply set $K = \mathbb{Q}[x]/\langle x^3 - 2\rangle$ (we need to show that $x^3 - 2$ is irreducible) and give the image of x the name $\sqrt[3]{2}$.

In a similar way, consider $f = x^2 + x + 1 \in \mathbb{F}_2[x]$. It does not have a zero in $\mathbb{F}_2$, hence f is irreducible. Therefore we can construct the field $K = \mathbb{F}_2[x]/\langle x^2+x+1\rangle$, which is a quadratic extension of $\mathbb{F}_2$ and therefore has four elements (and is usually called $\mathbb{F}_4$). If α is the image of x, then $\mathbb{F}_4 = \{0, 1, \alpha, \alpha + 1\}$, and we know how to compute with these elements if we keep in mind that $\alpha^2 = \alpha + 1$. We will look at finite fields more closely soon.

We see that irreducible polynomials are very important. Therefore, we need a way of knowing when a given polynomial is irreducible. For example, if $2 \le \deg f \le 3$, then f is irreducible in $F[x]$ if and only if f has no root in F. (If f is reducible, then it must have a factor of degree 1 and therefore a root.) Of course, linear polynomials (of degree 1) are always irreducible.

To get more powerful criteria, we assume that the field F is the field of fractions of a UFD R. Recall Gauss' Lemma one of whose consequences was that a primitive polyomial $f \in R[x]$ is irreducible if and only if it is irreducible in $F[x]$. This implies the following useful fact.

20.15. **Lemma.** *Let $f = a_nx^n + \ldots + a_0 \in R[x]$ with $a_n \neq 0$. If $\alpha \in F$ is a root of f, then $\alpha = r/s$ with $r, s \in R$ such that $r \mid a_0$ and $s \mid a_n$.*

Proof. In $F[x]$, f is divisible by $x - \alpha$, so in $R[x]$, f must be divisible by $c(x - \alpha)$, where $c \in R$ is such that $cx - c\alpha \in R[x]$ is primitive. Write $c(x - \alpha) = sx - r$; then $\alpha = r/s$, and s must divide the leading coefficient a_n of f, r must divide the constant coefficient a_0. □

20.16. **Example.** The polynomial $2x^3 + x^2 + 3 \in \mathbb{Q}[x]$ is irreducible. Otherwise, it would have to have a root $r/s \in \mathbb{Q}$, with $r \mid 3$ and $s \mid 2$. But none of the possibilities $\pm 1, \pm 3, \pm\frac{1}{2}, \pm\frac{3}{2}$ is a root.

20.17. **Lemma.** *Let $f = a_nx^n + \ldots + a_0 \in R[x]$ be primitive. If $p \in R$ is a prime element that does not divide a_n, and the image of f in $R/Rp[x]$ is irreducible, then f is irreducible.*

Proof. Note that there is a canonical R-algebra homomorphism $R[x] \to R/Rp[x]$ that extends $R \to R/Rp$ and sends x to x. (To be completely correct, one should use a different variable for the polynomial ring over R/Rp.) Write $a \mapsto \bar{a}$ for this homomorphism and assume that $\bar{f}$ is irreducible. Note that $\deg \bar{f} = \deg f$, since $\bar{a}_n \neq 0$. If f were reducible, then (since f is primitive) f would have to factor into two polynomials of smaller degree: $f = gh$. But then $\bar{f} = \bar{g}\bar{h}$ is a factorization of $\bar{f}$ into two polynomials of smaller degree, contradicting the irreducibility of $\bar{f}$. □

20.18. **Example.** The polynomial $x^3 + x + 10^{10^{100}} + 1 \in \mathbb{Z}[x] \subset \mathbb{Q}[x]$ is irreducible. In fact, its image in $\mathbb{F}_2[x]$ is irreducible (since it has no root there).

The most famous of all irreducibililty criteria is *Eisenstein's Criterion.* It is as follows.

20.19. **Proposition.** *Let $f = a_n x^n + \ldots + a_0 \in R[x]$ be primitive. If $p \in R$ is a prime element such that $p \nmid a_n$, but p divides all the other coefficients of f, and $p^2 \nmid a_0$, then f is irreducible.*

Proof. Suppose $f = gh$ in $R[x]$ with g and h of smaller degree. Consider again the homomorphism $R[x] \to R/Rp[x]$. By assumption, the image of f is $\bar{a}_n x^n$ with $\bar{a}_n \neq 0$. Since R/Rp is an integral domain, the only factors of $\bar{f}$ are (constant multiples of) polynomials of the form x^k for $0 \le k \le n$. In particular, $\bar{g} = bx^k$ and $\bar{h} = cx^{n-k}$ where $k = \deg g \ge 1$, $n - k = \deg h \ge 1$ and $b, c \neq 0$. This means that the constant terms in g and h are both divisible by p. But then their product a_0 must be divisible by p^2, a constradiction to our assumptions. □

A polynomial as in the statement above is called a *p-Eisenstein polynomial.*

20.20. **Example.** If $n \ge 1$ and p is a prime number, then $x^n - p$ and $x^n + p$ are p-Eisenstein polynomials (in $\mathbb{Z}[x]$) and therefore irreducible.

20.21. **Example.** If p is a prime number, then $f = x^{p-1} + \ldots + x + 1$ is irreducible in $\mathbb{Q}[x]$. Here we use a trick. Note that $f = (x^p - 1)/(x - 1)$ and therefore,

$$f(x+1) = \frac{(x+1)^p - 1}{x} = \sum_{j=0}^{p-1} \binom{p}{j+1} x^j = x^{p-1} + p\, x^{p-2} + \ldots + \binom{p}{2} x + p$$

is a p-Eisenstein polynomial. (We use the fact that the binomial coefficients occurring here are all divisible by p.) So $f(x+1)$ is irreducible, but then f is irreducible as well.

21. Splitting Fields and Algebraic Closure

As before, F denotes a field.

21.1. **Definition.** Let $f \in F[x]$ be a nonzero polynomial. A field extension F'/F is called a *splitting field* for f over F, if f splits into linear factors in $F'[x]$. If in addition, F' is generated (as an F-algebra) by the roots of f, then F'/F is a *minimal splitting field* for f over F.

21.2. **Proposition.** *Let $f \in F[x]$ be a nonzero polynomial. Then there is a splitting field F'/F for f. The subextension of F'/F generated by the roots of f is then a minimal splitting field for f.*

Proof. Write f as a product of linear polynomials times a product f_1 of irreducible polynomials of degree > 1. Denote by $n(F)$ the degree of f_1. We proceed by induction on $n(F)$.

If $n(F) = 0$, then f splits into linear factors over F, and F/F is a splitting field. Otherwise, let g be one of the irreducible factors of f_1, and let F_1/F be the extension obtained ba adjoining a root of g to F. Then g has a linear factor in $F_1[x]$, and therefore $n(F_1) < n(F)$ (f has more linear factors in $F_1[x]$ than in $F[x]$). By induction, there is a splitting field F'/F_1 for f over F_1. But then F'/F is a splitting field for f over F.

The last statement is clear. □

21.3. **Theorem.** *Let $0 \neq f \in F[x]$. If K/F is a minimal splitting field for f and F'/F is any splitting field for f, then there is an F-algebra homomorphism $\phi : K \to F'$. Any two minimal splitting fields for f are isomorphic as F-algebras.*

Proof. K is obtained from F by successively adjoining roots of f. By Prop. 20.13, we get an F-algebra homomorphism to F' for each of the intermediate fields, hence also for K.

If F' is also a minimal splitting field, then ϕ is surjective (since the roots of f in K, which generate K, are sent to the roots of f in F', which generate F'), hence an isomorphism (it is injective in any case). □

21.4. **Example.** The field $\mathbb{Q}(\zeta_n) \subset \mathbb{C}$, where $\zeta_n = e^{2\pi i/n}$ is a primitive nth root of unity, is a minimal splitting field of $x^n - 1$ over $\mathbb{Q}$. It is called the *nth cyclotomic field.*

For example, $\mathbb{Q}(\zeta_3) = \mathbb{Q}(\sqrt{-3})$, since $x^3 - 1 = (x-1)(x^2+x+1)$, and the roots of the quadratic factor are $(-1 \pm \sqrt{-3})/2$. Also, $\zeta_4 = i$, so $\mathbb{Q}(\zeta_4) = \mathbb{Q}(i)$.

21.5. **Proposition.** *Let F be a field. Then the following statements are equivalent.*

(1) *F is algebraically closed.*
(2) *Every non-constant polynomial $f \in F[x]$ has a root in F.*
(3) *All irreducible polynomials in $F[x]$ have degree 1.*

Proof. Assume F is not algebraically closed. Then there is a non-trivial algebraic extension F'/F, which contains an element $\alpha \in F' \setminus F$. Then the minimal polynomial of α is irreducible of degree > 1 and therefore does not have a root in F. This proves that (2) implies (1). Now, if there is an irreducible polynomial of degree > 1, then we can construct a non-trivial algebraic extension of F by Prop. 20.13. So (1) implies (3). Finally, if all irreducible polynomials are of degree 1, then every non-constant polynomial f is divisible by an irreducible polynomial of the form $x - a$ and so $f(a) = 0$. So (3) implies (2). □

21.6. **Example.** The "Fundamental Theorem of Algebra" can be stated as follows.

The field $\mathbb{C}$ of complex numbers is algebraically closed.

21.7. **Definition.** Let F'/F be a field extension. The subfield of F' consisting of all elements algebraic over F is called the *algebraic closure of F in F'*. The field F' is called an *algebraic closure* of F if it is algebraically closed and algebraic over F.

This raises the question whether every field necessarily has an algebraic closure. What about an algebraic closure of $\mathbb{Q}$?

21.8. **Lemma.** *If F'/F is a field extension such that F' is algebraically closed, then the algebraic closure of F in F' is an algebraic closure of F.*

Proof. Let K be the algebraic closure of F in F'. Then K/F is algebraic by definition. We have to show that K is algebraically closed. Assume that there is a non-trivial algebraic extension K'/K. We can assume that $K' = K(\alpha)$ for some $\alpha \notin K$, but algebraic over K. By Prop. 20.13, K' embeds into every field extension of K in which the minimal polynomial of α has a root. Since F' is algebraically closed, F' is such an extension, and so we can assume without loss of generality that $\alpha \in F'$. But then α is algebraic over F by Cor. 20.12 and so already in K, contradiction. □

So we can find an algebraic closure of $\mathbb{Q}$ by taking its algebraic closure in $\mathbb{C}$; this is the *field $\bar{\mathbb{Q}}$ of algebraic numbers.* Note that $\bar{\mathbb{Q}}$ is countable: if we map $\alpha \in \bar{\mathbb{Q}}$ to its minimal polynomial, then we get a map $\bar{\mathbb{Q}} \to \mathbb{Q}[x]$ with finite fibers, and $\mathbb{Q}[x]$ is countable. Since $\mathbb{C}$ is uncountable, "almost all" complex numbers are transcendental (over $\mathbb{Q}$).

But what can we do when we do not have a sufficiently large algebraically closed field at our disposal?

The idea is basically to enlarge the field by adjoining roots of irreducible polynomials until this cannot be done any further. However, countably many such steps may not be sufficient, and we need again a more powerful induction principle like Zorn's Lemma to help us out.

21.9. **Theorem.** *Every field F has an algebraic closure. If K/F is an algebraic closure and F'/F is algebraic, then there is an F-algebra homomorphism $F' \to K$. Any two algebraic closures of F are isomorphic as F-algebras.*

Proof. To show existence, we consider the set of all algebraic extensions of F, ordered by inclusion. (To avoid problems with paradoxes in set theory, one has to be a bit careful here. For example, one can restrict to extension fields whose underlying set is a subset of a fixed sufficiently large set.) This set is nonempty, since it contains the trivial extension F/F. If we have a (non-empty) chain in this set, then the union of its elements is again an algebraic extension of F. So we can apply Zorn's Lemma, which tells us that there is a maximal algebraic extension K/F. We have to show that K is algebraically closed. So let K'/K be an algebraic extension. Then K'/F is also algebraic by Cor. 20.12. But K was a maximal algebraic extension of F, hence we must have $K' = K$.

Now fix an algebraic closure K/F and let F'/F be algebraic. We consider the set of all pairs (L, ϕ), where L/F is a subextension of F' (i.e., $F \subset L \subset F'$ and L is a field) and $\phi : L \to K$ is an F-algebra homomorphism, ordered by inclusion on L such that the map on the larger L restricts to the map on the smaller L. This set is non-empty (it contains (F, i), where $i : F \to K$ is the inclusion), and if we have a non-empty chain $\{(L_i, \phi_i) : i \in I\}$, then $L = \bigcup_{i \in I} L_i$ is a subextension of F'/F, and the ϕ_i piece together to form an F-algebra homomorphism $\phi : L \to K$. So Zorn's Lemma applies again, and there is a maximal pair (L, ϕ). If $L \subsetneq F'$, then there is an algebraic element $\alpha \in F' \setminus L$, and by Prop. 20.13, we get a strictly larger pair $(L[\alpha], \phi')$. Hence $L = F'$, and we are done.

For the last statement, let K and K' be two algebraic closures. By the previous statement, there is an F-algebra homomorphism $\phi : K' \to K$; it is injective since K' is a field. The image $\phi(K')$ is isomorphic to K' and is therefore an algebraic closure of F. But $K/\phi(K')$ is an algebraic extension, hence $K = \phi(K')$. So ϕ is surjective as well and therefore an isomorphism. □

22. Finite Fields

In this section, we study the structure of finite fields. We already know the finite fields $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, but we will see that there are more.

22.1. **Definition.** Let F be a field. There is a unique ring homomorphism $\mathbb{Z} \to F$; its kernel is a principal ideal and therefore of the form $n\mathbb{Z}$ with $n \ge 0$. We call n the *characteristic* of F, written $\operatorname{char}(F)$.

22.2. **Proposition.** *Let F be a field.*

1. *The characteristic of F is either zero or a prime number.*
2. *If $\operatorname{char}(F) = 0$, then F is a $\mathbb{Q}$-algebra in a unique way. If $\operatorname{char}(F) = p$ is a prime number, then F is an $\mathbb{F}_p$-algebra in a unique way.*

Proof.
(1) Since F is a field, the image of the homomorphism $\mathbb{Z} \to F$ is an integral domain, Therefore the kernel is a prime ideal. The only prime ideals of $\mathbb{Z}$ are the zero ideal and the ideals generated by a prime number.

(2) Assume first that $\operatorname{char}(F) = p$ is prime. Then we have an injective ring homomorphism $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z} \to F$, exhibiting F as an $\mathbb{F}_p$-algebra. Now assume that $\operatorname{char}(F) = 0$. Then $\mathbb{Z} \to F$ is injective and therefore extends uniquely to a homomorphism $\mathbb{Q} \to F$, exhibiting F as a $\mathbb{Q}$-algebra. These algebra structures are unique since the homomorphism $\mathbb{Z} \to F$ is unique. □

22.3. **Corollary.** *If F is a finite field, then there is a prime number p and a positive integer f such that $\#F = p^f$. F is a vector space of dimension f over $\mathbb{F}_p$.*

Proof. If $\operatorname{char}(F) = 0$, then F would have to contain $\mathbb{Q}$, which is not possible. So $\operatorname{char}(F) = p$ is a prime number. We know then that F is an $\mathbb{F}_p$-algebra, hence an $\mathbb{F}_p$-vector space of some dimension $f \ge 1$. The claim follows. □

Do finite fields of size p^f exist for all prime powers p^f, and how unique are they? Before we answer these questions, we need a tool.

22.4. **Lemma.** *Let F be a finite field of characteristic p. Then the map $F \ni a \mapsto a^p \in F$ is a field automorphism of F.*

This automorphism is called the *Frobenius automorphism.*

Proof. Let $\phi(a) = a^p$ be the map. We obviously have $\phi(1) = 1$ and $\phi(ab) = (ab)^p = a^p b^p = \phi(a)\phi(b)$. Also,

$$\phi(a+b) = (a+b)^p = \sum_{k=0}^{p} \binom{p}{k} a^k b^{p-k} = a^p + b^p = \phi(a) + \phi(b)\,,$$

since the binomial coefficients except the extremal ones are divisible by p. Also, $\ker\phi = \{a \in F : a^p = 0\} = \{0\}$, so ϕ is injective. Since F is finite, ϕ must be bijective. □

22.5. **Lemma.** *If F is a finite field with $\#F = p^f$, then F is a minimal splitting field for the polynomial $x^{p^f} - x$ over $\mathbb{F}_p$.*

Proof. By Cor. 19.8, all elements of F are roots of $x^{p^f} - x$. Therefore, $F/\mathbb{F}_p$ is a splitting field for $x^{p^f} - x$. Since F obviously is generated by its elements (as an $\mathbb{F}_p$-algebra), it is a minimal splitting field. □

22.6. **Theorem.** *For every prime number p and positive integer f, there exists a field $\mathbb{F}_{p^f}$ of size p^f. Any two such fields are isomorphic.*

Proof. By Prop. 21.2, there exists a minimal splitting field F of $q = x^{p^f} - x$ over $\mathbb{F}_p$. We claim that F consists exactly of the roots of q. For this, we have to show that the roots of q form a field. Let ϕ be the Frobenius automorphism of F and let $\Phi(a) = a^{p^f}$ be its f-fold iterate. Then for $a \in F$, we have $q(a) = 0 \iff \Phi(a) = a$. Since Φ is a ring homomorphism, this implies $q(a) = q(b) = 0 \Rightarrow q(a+b) = q(ab) = 0$. Therefore the set of roots of q in F is a subring, and since it is finite, it is a subfield. Since F is generated by the roots, F must be this subfield. So $\#F = \deg q = p^f$.

By Thm. 21.3, any two minimal splitting fields of a polynomial are isomorphic. By the preceding lemma, this implies that any two finite fields of the same size are isomorphic. □

23. Constructions with Straightedge and Compass

Let us relate classical geometric constructions in the plane with field theory. We identify the complex plane $\mathbb{C}$ with the geometric plane $\mathbb{R}^2$.

23.1. **Definition.** Let $S \subset \mathbb{C}$ be a subset containing 0 and 1. We let S' be S, together with the points that can be obtained from points in S by the following constructions. A line is called an "S-line" if it contains two distinct points of S. A circle is called an "S-circle" if its center is a point in S and its radius is the distance of two points in S.

(1) The point of intersection of two S-lines.
(2) The points of intersection of an S-line and an S-circle.
(3) The points of intersection of two S-circles.

We define $S^{(0)} = S$ and $S^{(n+1)} = (S^n)'$. Let $\bar{S} = \bigcup_{n \ge 0} S^{(n)}$. Then we call $z \in \mathbb{C}$ *constructible from* S if $z \in \bar{S}$.

23.2. **Lemma.** *From $\{0, 1, a, b\}$, we can construct $a \pm b$, $a \cdot b$, a^{-1} (if $a \neq 0$) and $\sqrt{a}$.*

Proof. These are classical constructions and left as an exercise. (For the square root of a positive real, use a suitable right-angled triangle.) □

23.3. **Corollary.** *$F = \bar{S} \subset \mathbb{C}$ is a subfield that contains S, and every polynomial of degree 2 in $F[x]$ has a root in F.*

Proof. That F is a subfield follows from the preceding lemma. If $x^2 + 2ax + b$ is a polynomial of degree 2 over F, then its roots are $-a \pm \sqrt{a^2 - b}$ and so are in $F = \bar{S}$ as well by the lemma again. □

23.4. **Lemma.** *Let $F_0 = \mathbb{Q}(S)$ be the subfield of $\mathbb{C}$ generated by S. Assume S is stable under complex conjugation (i.e., $a + bi \in S$ implies $a - bi \in S$). If $z \in S'$, then z is in an extension of F_0 obtained by a sequence of successive quadratic extensions, and S' is stable under complex conjugation.*

Proof. If S is stable under complex conjugation and $z = a + bi \in \mathbb{C}$, then $z \in \mathbb{Q}(S, i)$ if and only if $a, b \in \mathbb{Q}(S, i)$. Also, $\mathbb{Q}(S, i) = F_0(i)$ is a quadratic extension of F_0. It therefore suffices to prove the claim for the real and imaginary parts of the newly constructed points. Now the equations of S-lines and S-circles (in terms of real and imaginary part as coordinates) have coefficients that are rational expressions in the real and imaginary parts of points in S. If we intersect two S-lines, then the point of intersection has coordinates that are again rational expressions in real and imaginary parts of points in S. If we intersect an S-line and an S-circle, we obtain a quadratic equation for one of the coordinates (and then a linear equation for the other one). We therefore only have to make a quadratic extension. If we intersect two S-circles, the the difference of their equations (in standard form $(x - a)^2 + (y - b)^2 = r^2$) is a linear equation, and we are reduced to the previous case.

It is clear that S' will again be stable under complex conjugation. □

23.5. **Theorem.** *Let $\{0, 1\} \subset S \subset \mathbb{C}$ be stable under complex conjugation. Then $F = \bar{S} \subset \mathbb{C}$ is the smallest subfield of $\mathbb{C}$ containing S such that every polynomial of degree 2 in $F[x]$ has a root in F.*

Proof. We already know that F is "closed under square roots". On the other hand, the lemma above implies that $\bar{S}$ is contained in every subfield of $\mathbb{C}$ containing S and closed under square roots. So F must be the smallest such subfield. □

23.6. **Corollary.** *Let S be as above. Then $z \in \mathbb{C}$ is constructible from S if and only if there is a chain $\mathbb{Q}(S) = F_0 \subset F_1 \subset \cdots \subset F_n$ of quadratic field extensions such that $z \in F_n$.*

23.7. **Corollary.** *Let S be as above, and let $F_0 = \mathbb{Q}(S)$. If $z \in \mathbb{C}$ is constructible from S, then z is algebraic over F_0 of degree a power of 2.*

Proof. By the above, there is a chain of quadratic extensions $F_0 \subset F_1 \subset \ldots \subset F_n$ such that $z \in F_n$. Since $[F_n : F_0] = 2^n$, the degree of z is a divisor of 2^n (note that $F_0(z) \subset F_n$, so $[F_0(z) : F_0] \mid [F_n : F_0]$). □

This allows us to draw some conclusions. If we do not specify the set S in the following, we assume that $S = \{0, 1\}$.

23.8. **Theorem.** *We cannot "double the cube".*

"Doubling the cube" means to construct the side of a cube whose volume is twice that of the unit cube.

Proof. The number we have to construct is $\sqrt[3]{2}$. But this number is algebraic of degree 3 (since it has minimal polynomial $x^3 - 2$), which is not a power of 2. □

23.9. **Theorem.** *We cannot construct a regular* 7*-gon.*

Proof. If we could, we certainly could also construct the x-coordinate of one of the vertices next to 1 in a regular 7-gon centered at the origin and with one vertex at 1. Twice this number is $2\cos\frac{2\pi}{7}$ and is algebraic of degree 3 (its minimal polynomial is x^3+x^2-2x-1, which is for example seen to be irreducible via reduction mod 2). Since 3 is not a power of 2, we have a contradiction. □

23.10. **Remark.** Analyzing the situation further, one can prove that the regular n-gon can be constructed if and only if $n = 2^k p_1 \cdots p_\ell$, where the p_j are distinct *Fermat primes* of the form $2^{2^m} + 1$. Gauss was the first to prove that; he also provided a construction of the regular 17-gon. (The only known Fermat primes are 3, 5, 17, 257 and 65537.)

23.11. **Theorem.** *We cannot construct an angle of* 40 *degrees.*

Proof. If we could, we could also construct twice its cosine $2\cos\frac{2\pi}{9}$. But again, this number is algebraic of degree 3 (minimal polynomial is $x^3 - 3x + 1$). □

23.12. **Corollary.** *We cannot trisect an arbitrary angle.*

What we mean by "trisecting an arbitrary angle" is the following. Let α be some angle (in radians). We let $S = \{0, 1, e^{i\alpha}, e^{-i\alpha}\}$. To trisect the angle α then means to construct $e^{i\alpha/3}$ from S. Equivalently, we can replace the exponentials by cosines, since we can easily (re-)construct the sines.

Proof. If we could, then we could also trisect the special angle $2\pi/3$. Its cosine is $-1/2$, so can be constructed from $\{0, 1\}$. But we know that the angle $2\pi/9$ is not constructible. □

23.13. **Theorem.** *We cannot "square the circle".*

"Squaring the circle" means to construct the side of a square whose area is the same as that of the unit circle.

Proof. The number we must construct is $\sqrt{\pi}$. If this is constructible, then so is its square π. But π is not even algebraic! (The proof of this fact is unfortunately beyond the scope of this course.) □