LECTURE NOTES
ALGEBRA 2 FOR COMPUTER SCIENTISTS

DAVID STANOVSKÝ
TRANSLATED FROM CZECH BY MICHAEL KOMPATSCHER, FILIPPO SPAGGIARI

stanovsk@karlin.mff.cuni.cz

Date: May 22, 2023.

Contents

Homomorphisms

1. Group homomorphisms

In mathematics, the word homomorphism is often used for mappings that preserve the basic structure of mathematical objects. For example, in linear algebra, homomorphisms are maps between vector spaces that preserve the addition and multiplication with scalars (you may also know such homomorphisms as \linear maps"). In graph theory, a (graph) homomorphism is an edge-preserving map between two graphs.

In this section, we are going to de ne group homomorphisms as maps between groups that preserve their basic algebraic operations. As we will see, group homomorphisms also preserve a number of other important properties.

1.1. **Definition and basic properties.** In this whole subsection, we will use the notation $\mathbf{G} = (G; \ ; \ ^{-1}; 1)$ and $\mathbf{H} = (H; \ ; ^{0}; e)$ for two distinct groups.

Definition. Let $\mathbf{G}; \mathbf{H}$ be groups. A map $' : G \to H$ is a homomorphism from $\mathbf{G}$ to $\mathbf{H}$, if, for all $a; b \in G$:

$$'(a \ b) = '(a) \ '(b); \qquad '(a^{-1}) = '(a)^{0}; \qquad '(1) = e:$$

For short, we are going to write $' : \mathbf{G} \to \mathbf{H}$, if the map $' : G \to H$ is a homomorphism from $\mathbf{G}$ to $\mathbf{H}$.

We rst show that the second and third equations in this de nition already follow from the rst one. In many cases, this can make it signi cantly easy to check, whether a map is a homomorphism:

Lemma 1.1. Let $\mathbf{G}; \mathbf{H}$ be groups and $' : G \to H$ be a map between their domains. Then $'$ is a homomorphism between $\mathbf{G}$ and $\mathbf{H}$, if $'(a \ b) = '(a) \ '(b)$ for all $a; b \in G$.

Proof. Let us rst show that $'(1) = e$. For this note that $e \ '(1) = '(1) = '(1 \ 1) = '(1) \ '(1)$. Cancelling the right factor $'(1)$ gives us $e = '(1)$.

Next, let us show that $'(a^{-1}) = '(a)^{0}$ for all $a \in G$. For this, note that $e = '(1) = '(a \ a^{-1}) = '(a) \ '(a^{-1})$. By the uniqueness of inverse elements in $\mathbf{H}$, we get $'(a^{-1}) = '(a)^{0}$.

Let $' : \mathbf{G} \to \mathbf{H}$ be a homomorphism. Then, its image is the range of its values, i.e. the set

$$\mathrm{Im}(') = \{'(a) : a \in G\}:$$

We de ne the kernel of $'$ as the set

$$\mathrm{Ker}(') = \{a \in G : '(a) = e\}:$$

Proposition 1.2. Let $\mathbf{G}; \mathbf{H}$ be groups and $' : \mathbf{G} \to \mathbf{H}$ a homomorphisms. Then

(1) $\mathrm{Im}(')$ is a subgroup of $\mathbf{H}$;
(2) $\mathrm{Ker}(')$ is a subgroup of $\mathbf{G}$.

Proof. (1) Clearly $e \in \mathrm{Im}(')$, since $e = '(1)$. If $'(a); '(b) \in \mathrm{Im}(')$, then $'(a)^{0} = '(a^{-1}) \in \mathrm{Im}(')$ and $'(a) \ '(b) = '(a \ b) \in \mathrm{Im}(')$. So $\mathrm{Im}(')$ is closed under the group operations of $\mathbf{H}$, i.e. a subgroup of $\mathbf{H}$.

(2) Clearly $1 \in \mathrm{Ker}(')$, since $'(1) = e$. If $a; b \in \mathrm{Ker}(')$, then so are a^{-1} an $a \ b$, since $'(a^{-1}) = '(a)^{0} = e^{0} = e$ and $'(a \ b) = '(a) \ '(b) = e \ e = e$.

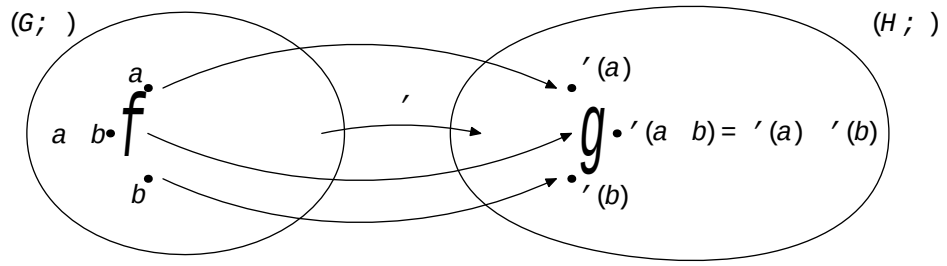


Figure 1. Homomophism $' : \mathbf{G} \to \mathbf{H}$.

Proposition 1.3. Let $\mathbf{G}, \mathbf{H}$ be groups and $' : \mathbf{G} \to \mathbf{H}$ be a homomorphism. Then $'$ is injective, if and only if $\mathrm{Ker}(') = \{1\}$.

Proof. If $'$ is injective, then it maps at most one element of G to e. Thus $\mathrm{Ker}(')$ must be equal to 1. On the other hand, assume that $'(a) = '(b)$, for two elements $a \neq b$. Then $e = '(a)\ '(b)^0 = '(a\ b^{-1})$, thus $1 \neq a\ b^{-1} \in \mathrm{ker}(')$.

Examples. Several important maps in mathematics are examples of group homomorphisms:

The map $z \mapsto |z|$ for complex numbers $z \in \mathbb{C}$ is a homomorphism between the multiplication groups $\mathbb{C}^* \to \mathbb{R}^*$, since $|ab| = |a|\,|b|$. Its kernel is the subgroup of complex numbers of absolute value 1 (the unit circle). The map is however not a homomorphism between the additive groups $\mathbb{C} \to \mathbb{R}$, since in general $|a + b| \neq |a| + |b|$.

Consider the map $z \mapsto e^z$, again on the complex numbers. This is a group homomorphism $\mathbb{C} \to \mathbb{C}^*$, since $e^{a+b} = e^a\ e^b$. Its kernel is the subgroup $\{2\pi i = \{k \in i : k \in \mathbb{Z}\} \subset \mathbb{C}$, its image is $\mathbb{C}^*$.

Consider the map $A \mapsto \det(A)$ for matrices $A \in \mathbf{T}^{n\ n}$ over a field $\mathbf{T}$. Restricted to invertible matrices, this is a group homomorphism $\mathbf{GL}_n(\mathbf{T}) \to \mathbf{T}^*$, since $\det(AB) = \det(A)\ \det(B)$. Its kernel is the special linear group $\mathbf{SL}_n(\mathbf{T})$, its image is all of $\mathbf{T}^*$.

Let's consider the map $\sigma \mapsto \mathrm{sgn}(\sigma)$ for permutations $\sigma \in \mathbf{S}_n$. This is a homomorphism $\mathbf{S}_n \to \mathbb{Z}^*$, since $\mathrm{sgn}(\quad) = \mathrm{sgn}(\ )\ \mathrm{sgn}(\ )$. Its kernel is the alternating group $\mathbf{A}_n$, its image is $\mathbb{Z}^* = \{1\}$.

Example. Let $\mathbf{G}$ be a group, and $a \in G$ an element of order n. Then, the \discrete" exponential function $\mathbb{Z}_n \to \mathbf{G}$, $k \mapsto a^k$, is a homomorphism. Its image is the subgroup $\langle a\rangle_{\mathbf{G}}$.

Example. The action of the group $\mathbf{G}$ on the set X is nothing but a homomorphism $\mathbf{G} \to \mathbf{S}_X$ (recall the definition of group actions from Algebra 1!).

Homomorphisms are determined by their values on generators: To see this let $' : \mathbf{G} \to \mathbf{H}$ be a homomorphism, let $\mathbf{G} = \langle X\rangle$ and denote the values $'(a) = h_a$ for all $a \in X$. A general element of the group $\mathbf{G}$ can be written in the form $g = a_1^{k_1} \cdots a_n^{k_n}$, where $a_1, \ldots, a_n \in X$ and $k_1, \ldots, k_n \in \mathbb{Z}$. The value of g under $'$ then must be equal to

$$'(g) = '(a_1)^{k_1} \cdots '(a_n)^{k_n} = h_{a_1}^{k_1} \cdots h_{a_n}^{k_n}.$$

However, we remark that (unlike for vector spaces) it is not possible to obtain a homomorphism by choosing arbitrary images for the elements of some minimal generating set, as the following statement shows:

Proposition 1.4 (the order of elements)**.** Let $\varphi : \mathbf{G} \to \mathbf{H}$ be a group homomorphism and $a \in G$. Then

$$\operatorname{ord}(\varphi(a)) \mid \operatorname{ord}(a) \text{ (for } \operatorname{ord}(a) < \infty\text{)}.$$

If φ is injective, then

$$\operatorname{ord}(\varphi(a)) = \operatorname{ord}(a).$$

Proof. First, assume that $\operatorname{ord}(a) = n < \infty$. Then $\varphi(a)^n = \varphi(a^n) = \varphi(1) = e$, thus there must be a $k \mid n$ with $\varphi(a)^k = e$.

If φ is injective, then $a^k \neq 1$ implies $\varphi(a)^k = \varphi(a^k) \neq e$. Thus $\operatorname{ord}(\varphi(a)) = \operatorname{ord}(a)$ (even if $\operatorname{ord}(a) = \infty$).

Exercise. Describe all homomorphisms between $Z_{10} \to \mathbf{S}_3$.

Solution. The group Z_{10} is cyclic and generated by 1. Thus every homomorphism φ is already determined $\varphi(1)$, as $\varphi(k) = \varphi(1+\ldots+1) = \varphi(1) \ldots (1) = \varphi(1)^k$. The order of 1 in Z_{10} is 10, thus, the order of $\varphi(1)$ in $\mathbf{S}_3$ must divide 10. However, in $\mathbf{S}_3$ there are only elements of order 1, 2, 3, so we only have the following four possibilities: $\varphi(1) \in \{id, (1\,2), (1\,3), (2\,3)\}$. It is easy to verify that these four options give rise to the four homomorphisms $k \mapsto id$, and $k \mapsto (1\,2)^k$, $k \mapsto (1\,3)^k$, $k \mapsto (2\,3)^k$. (Think for a moment for yourself, why $k \mapsto (1\,2\,3)^k$ is not a homomorphism, without the help of Proposition 1.4.)

Proposition 1.5. Let $\mathbf{G}, \mathbf{H}, \mathbf{K}$ be groups and $\varphi : \mathbf{G} \to \mathbf{H}$, $\psi : \mathbf{H} \to \mathbf{K}$ homomorhpisms. Then

(1) $\psi \circ \varphi$ is a homomorphism $\mathbf{G} \to \mathbf{K}$,
(2) if φ is bijective, then φ^{-1} is a homomorphism $\mathbf{H} \to \mathbf{G}$.

Proof. (1) Let us denote $\mathbf{K} = (K, +, \cdot, 0)$ (without implying that $\mathbf{K}$ is abelian). For all $a, b \in G$ we then have

$$(\psi \circ \varphi)(a \cdot b) = \psi(\varphi(a \cdot b)) = \psi(\varphi(a) \cdot \varphi(b)) = \psi(\varphi(a)) + \psi(\varphi(b)) = (\psi \circ \varphi)(a) + (\psi \circ \varphi)(b),$$

since ψ and φ are homomorphisms. Thus also $\psi \circ \varphi : G \to \mathbf{K}$ is a homomorphism.

(2) Let $u, v \in H$ such that $u = \varphi(a)$ and $v = \varphi(b)$ for some elements $a, b \in G$. Then

$$\varphi^{-1}(u \cdot v) = \varphi^{-1}(\varphi(a) \cdot \varphi(b)) = \varphi^{-1}(\varphi(a \cdot b)) = a \cdot b = \varphi^{-1}(u) \cdot \varphi^{-1}(v),$$

since φ^{-1} is a homomorphism, and $\varphi^{-1} \circ \varphi = id$. Thus also φ^{-1} is a homomorphism.

1.2. Isomorphisms.

Definition. A bijective homomorphism is called an isomorphism.

From Proposition 1.5 it follows that the compositions and inverse functions of isomorphisms are also isomorphisms.

An isomorphism can be viewed as a map that \copies" an algebraic structure: If we have a group $\mathbf{G}$ and a bijective map $\varphi : G \to H$, then we can de ne group operations on H by

$$e = \varphi(1), \qquad a^{\prime} = \varphi((\varphi^{-1}(a))^{-1}), \qquad a \cdot b = \varphi(\varphi^{-1}(a) \cdot \varphi^{-1}(b)).$$

It is not hard to see that φ^{-1} is an isomorphism from $\mathbf{H} = (H, \cdot^{\prime}, e)$ to the old group $\mathbf{G}$ (and hence φ is an isomorphism from $\mathbf{G}$ to $\mathbf{H}$). In other words, $\mathbf{G}$ and $\mathbf{H}$ are copies of each other, only modulo renaming the elements by φ (because of this $\mathbf{H}$ is also called an isomorphic copy of $\mathbf{G}$). Note that every isomorphism can be viewed in this way.

Two groups $\mathbf{G}, \mathbf{H}$ are called isomorphic, if there exists a homomorphisms $\mathbf{G} \to \mathbf{H}$; for short we then also write $\mathbf{G} \simeq \mathbf{H}$. Proposition 1.5 implies, that $\simeq$ is an equivalence relation in the class of groups:

> Re exivity: $\mathbf{G} \simeq \mathbf{G}$ is witnessed by the isomorphism $id : \mathbf{G} \to \mathbf{G}$;
> Symmetry: If $\mathbf{G} \simeq \mathbf{H}$ is witnessed by an isomorphism φ, then also $\mathbf{H} \simeq \mathbf{G}$, as φ^{-1} is an isomorphism;
> Transitivity: If $\mathbf{G} \simeq \mathbf{H}$ and $\mathbf{H} \simeq \mathbf{K}$, then there are isomorphisms $\varphi : \mathbf{G} \simeq \mathbf{H}$, and $\psi : \mathbf{H} \simeq \mathbf{K}$. Their composition $\varphi\psi$ is an isomorphism from $\mathbf{G}$ to $\mathbf{K}$, hence $\mathbf{G} \simeq \mathbf{K}$.

If we restrict the codomain of an injective homomorphism $\varphi : \mathbf{G} \to \mathbf{H}$ to its image, we obtain an isomorphism $\varphi : \mathbf{G} \to \mathbf{Im}(\varphi)$, so the image of φ is an isomorphic copy of $\mathbf{G}$ in $\mathbf{H}$. Injective homomorphisms are called embeddings.

Example. The groups Z_2 and Z are isomorphic. This can already be seen from their operation tables:

+	0	1
0	0	1
1	1	0

	1	1
1	1	1
1	1	1

By renaming the elements in operation table for + by $0 \mapsto 1$, $1 \mapsto -1$, we obtain the operation table for . Thus, this map is a group isomorphism from Z_2 and Z .

Example. The groups C and R R are isomorphic. Intuitively, complex numbers, correspond to pairs of real numbers, and addition on C is de ned coordinate-wise. So, $a + bi \mapsto (a, b)$ is a group isomorphism witnessing C R R.

Example. The groups Z_n and $C_n = \langle \ _n \rangle_C$, with $_n = e^{2 \ i=n}$, are isomorphic. It is not hard to see that $^{k_1}_n \ ^{k_2}_n = ^{k_1+k_2}_n = ^{k_1+k_2 \bmod n}_n$. Thus $k \mapsto ^k_n$ is a group isomorphism witnessing Z_n C_n.

Example. The groups Z_2 Z_2, Z_8 and $\{id, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$ $\mathbf{S}_4$ are all isomorphic. This is not immediately clear, but one can prove that all three groups are of the form $\mathbf{G} = \langle a, b \rangle$, where a and b are generators satisfying $a^2 = 1$, $b^2 = 1$, and $ab = ba$ is the third non-unity element. The formal proof is left as an exercise.

Exercises.

1. Show, that the three groups Z_2 Z_2, Z_8 and $\{id, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$ $\mathbf{S}_4$ are isomorphic.

2. Show that R R^+, where R^+ is the subgroup of the multiplication group R that consists of all positive numbers.

3. Show that $\mathbf{GL}_2(Z_2)$ $\mathbf{S}_3$. (Hint: all permutations of nonzero vectors are linear mappings.)

4. Show that C R^+ $\mathbf{S}$, where $\mathbf{S}$ is the subgroup of C consisting of the unit circle.

5. Show that $\mathbf{D}_{12}$ $\mathbf{S}_3$ Z_2.

6. Show that the map

$$i \mapsto \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}; \quad j \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

can be extended to an injective group homomorphisms $\mathbf{Q}_8 \to \mathbf{GL}_2(C)$. This means the quaternion group is isomorphic to a subgroup of 2 2-matrices over C. Try to extend the statement to an isomorphism of the entire ring of quaternions and a certain matrix ring.

1.3. **Non-isomorphic groups.** We saw that $a + bi \mapsto (a;b)$ is a group isomorphism $\mathbb{C}' \mathbb{R} \mathbb{R}$; however, it is not an isomorphism between the multiplication groups $\mathbb{C}$ and $\mathbb{R} \mathbb{R}$. But, could those two groups still be isomorphic via some other map?

For another example, recall that the Chinese remainder theorem states that, for coprime numbers $m;n$, $\mathbb{Z}_{mn}' \mathbb{Z}_m \mathbb{Z}_n$. But what is the situation for $m;n$, which are not coprime? Then, the map $x \mapsto (x \bmod m; x \bmod n)$ is neither injective, nor onto: Both 0 and the least common multiple $\operatorname{lcm}(m;n)$ are mapped to $(0;0)$. So it clearly cannot be an isomophism. But how can we exclude that there isn't another isomorphism?

Invariants are a general principle that make it possible to resolve such questions. We call a property V an invariant if for every two isomorphic groups $\mathbf{G}' \mathbf{H}$ it holds that $\mathbf{H}$ has property V if $\mathbf{G}$ has property V (and vice-versa).

An example of an invariant is the number of elements of a given order: if $'$ is an isomorphism, according to Proposition 1.4, the order of a and $'(a)$ is always the same.

Examples.

The group $\mathbb{Z}_{mn}$ contains an element of order mn. However, the order of every element $\mathbb{Z}_m \mathbb{Z}_n$ needs to divide the least common multiple of m and n. So, if m and n are not coprime, then $\mathbb{Z}_{mn}$ and $\mathbb{Z}_m \mathbb{Z}_n$ are not isomorphic.

The group $\mathbb{C}$ contains elements of any order, but the group $\mathbb{R} \mathbb{R}$ only contains elements of order $1;2;1$. So these groups cannot be isomorphic.

The quaternion group $\mathbf{Q}_8$ and the dihedral group $\mathbf{D}_8$ both have 8 elements, and contain elements of order $1;2;4$. But, since $\mathbf{Q}_8$ has six elements of order 4, while $\mathbf{D}_8$ has only two, they cannot be isomorphic.

Another example of an invariant is the minimal number of generators:

Proposition 1.6. Let $' : \mathbf{G} \to \mathbf{H}$ be a group homomorphism that is onto. If $\mathbf{G} = \langle X \rangle$, then $\mathbf{H} = \langle '(X) \rangle$.

Proof. Since $\mathbf{G}$ is generate by X, every element $a \in G$ can be written as $a = u_1^{k_1} \cdots u_n^{k_n}$ for $u_1;\ldots;u_n \in X$. Then, also $'(a) = '(u_1)^{k_1} \cdots '(u_n)^{k_n} \in \langle '(X) \rangle$. Since $'$ is surjective, $\mathbf{H} = \langle '(X) \rangle$.

Unlike vector spaces, groups can have minimal generating sets of di erent sizes, e.g. $\mathbb{Z} = \langle 1 \rangle = \langle 2;3 \rangle$. But an invariant is the smallest number of elements needed to generate a given group:

Example. The groups $\mathbb{Z}$ and $\mathbb{Z} \mathbb{Z}$ are not isomorphic, because $\mathbb{Z} = \langle 1 \rangle$ is generated by one element, but this is not true for $\mathbb{Z} \mathbb{Z}$: Every element $(a;b)$ only generates the subgroup $\langle (a;b) \rangle = \{(ka;kb) : k \in \mathbb{Z}\}$, which is not the full $\mathbb{Z} \mathbb{Z}$. A similar argument shows also that $\mathbb{Z}_{mn} \not\cong \mathbb{Z}_m \mathbb{Z}_n$, for non-coprime $m;n$.

The two invariants mentioned above make it possible to prove nonisomorphicity in many cases, but not in all. As an example, take the groups $\mathbb{Q}$ and $\mathbb{Q}^+ = \{a \in \mathbb{Q} : a > 0\} \subset \mathbb{Q}$, which are both not nitely generated and contain only elements of in nite orders in addition to their neutral element.

Example. We say that $\mathbf{G}$ has square-roots if for every $a \in G$, there exists an element $b \in G$ with $a = b^2$. The existence of square-roots is invariant under isomorphisms. To see this, let us assume $' : \mathbf{G} \to \mathbf{H}$ is an isomorphism, and that square-roots exist in $\mathbf{G}$. Then for every element $u \in H$, let $a = '^{-1}(u)$ be its pre-image. By assumption, there is a $b \in G$ with $a = b \, b$. Then $v = '(b)$ is a root of u, since $u = '(a) = '(b^2) = '(b)^2 = v^2$.

This invariant can be used to show that Q and Q^+ are not isomorphic: On one side, for every $a \in Q$, there exists a $b \in Q$ such that $a = b + b$; on the other side, there are numbers $0 < a \in Q$ such that there is no $0 < b \in Q$ with $a = b^2$.

In general, it can be said that an invariant (with respect to isomorphisms) is any property that can be formulated using the operations of the given algebraic structure, equality, logical conjunctions and quanti ers (i.e. by so-called rst-order formula). You can nd details in any textbook on mathematical logic. However, even in this most general sense, just looking at invariants may not be enough in proving that two algebras are not isomorphic. For instance, the groups Q and Q Q, cannot be distinguished by any rst-order formula, but are nevertheless not isomorphic. We further remark that (depending on how they are presented) it can be computationally very hard, or even undecidable, to check if two nitely generated groups are isomorphic or not.

Exercises.

1. Decide whether the groups $\mathbf{A}_4$ and $\mathbf{D}_{12}$ are isomorphic.

2. Decide whether the groups Z_{24} and Z_{15} are isomorphic.

3. Decide whether the groups $\mathbf{SL}_2(Z_3)$ an $\mathbf{S}_4$ are isomorphic.

4. Show that $Q \not\cong Q \quad Q$. (Hint: assume $' : Q \to Q \quad Q$ is a homomorphism, and let $'(1) = (r; s)$. Prove that this value already uniquely determines the map $'$ uniquely and that we never get a mapping onto $Q \quad Q$.)

1.4. **Classification theorems.** One of the basic goals, when studying classes of algebraic structures is to classify them, i.e. to give a a complete list of them up to isomorphism. While it can be very di cult to achieve this in general, it often is feasible for some subclasses.

Probably the easiest example is the classi cation of cyclic groups. A group is called cyclic if it has one generator. Each such group is either isomorphic to Z or to exactly one Z_n for some $n \in N$. In other words, Z and Z_n are, up to isomorphisms, all cyclic groups.

Theorem 1.7 (classi cation of cyclic groups)**.** Let $\mathbf{G}$ be a cyclic group.

(1) If $\mathbf{G}$ is in nite, it is isomorphic to Z.
(2) If $\mathbf{G}$ is of order $n < 1$, then it is isomorphic to Z_n.

Proof. Let $\mathbf{G} = \langle a \rangle$ be the cyclic group.

(1) If $\mathbf{G}$ is in nite, then $\mathrm{ord}(a) = 1$. We claim that

$$Z \to \mathbf{G}; \qquad k \mapsto a^k:$$

is an isomorphism. It is easy to see that it is a homomorphism, since $a^k \quad a^l = a^{k+l}$. The kernel of it is trivial, since $a^k \neq 1$ for all $k \neq 0$, so by Proposition 1.3 it is injective. By de nition, it is also surjective, thus it is an isomorphism.

(2) Assume that $\mathbf{G}$ is of order n, so $\mathrm{ord}(a) = n$. Then consider the map

$$Z_n \to \mathbf{G}; \qquad k \mapsto a^k:$$

This map is a again a homomorphism, since $a^k \quad a^l = a^{k+l} = a^{k+l \bmod n}$; the second equality follows from the following consideration: if $k + l < n$, the statement is trivial; if $k + l \quad n$, then $k + l \bmod n = k + l \quad n$, and thus $a^{k+l \bmod n} = a^{k+l} \quad a^{\quad n} = a^{k+l} \quad 1^{\quad 1} = a^{k+l}$. Similarly to (1), we see that the kernel is trivial and that the map is onto $\mathbf{G}$.

The classi cation of nitely generated abelian groups is already much more complicated. It states that every abelian group with a nite set of generators is isomorphic to the direct product of nitely many cyclic groups. Moreover, using the

Chinese remainder theorem in the form of Proposition 4.4, it is su cient to consider only the order of the prime power of the nite cyclic group. These direct components are uniquely determined (except for the order), i.e. by choosing non-isomorphic cyclic groups we get non-isomorphic direct products.

Theorem 1.8 (Classi cation of nitely generated abelian groups)**.** Let $\mathbf{G}$ be a nitely generated abelian group with $j\mathbf{G}j > 1$. Then there exist $m; n \quad 0$, primes $p_1; :::; p_m$ (not necessarily di erent ones) and numbers $k_1; :::; k_m$ such that

$$\mathbf{G} \ ' \ Z^n \quad Z_{p_1^{k_1}} \quad Z_{p_2^{k_2}} \quad \quad Z_{p_m^{k_m}}:$$

The numbers $m; n$ are unique, and the numbers $p_1^{k_1}; :::; p_m^{k_m}$ are unique, up to their order.

The proof of Theorem 1.8 is quite long; you can nd it in every textbook on basic group theory.

Example. By Theorem 1.8 every 4-element abelian group is isomorphic to Z_4 or to $Z_2 \quad Z_2$.

The group Z_5 is of order 4 and abelian. Since ord(2) = 4, we get $Z_5 \ ' \ Z_4$. The group Z_8 is also abelian and of order 4. Since ord(3) = ord(5) = ord(7) = 2, we get $Z_8 \ ' \ Z_2 \quad Z_2$.

A favorite past-time of group theorists is to enumerate all small groups up to isomorphism, which can also be seen as a classi cation theorem. Currently, a list of all groups up to size $2047 = 2^{11} \quad 1$ is known. The following table contains the classi cation of all groups of order n for $n \quad 15$ and for $n = p; 2p; p^2$, where p is a prime number.

n	groups of order n
1	Z_1
2	Z_2
3	Z_3
4	$Z_4; \ Z_2 \quad Z_2$
5	Z_5
6	$Z_6; \ \mathbf{S}_3 = \mathbf{D}_6$
7	Z_7
8	$Z_8; \ Z_2 \quad Z_4; \ Z_2 \quad Z_2 \quad Z_2; \ \mathbf{D}_8; \ \mathbf{Q}_8$
p	Z_p
p^2	$Z_{p^2}; \ Z_p \quad Z_p$
$2p$	$Z_{2p}; \ \mathbf{D}_{2p}$
12	$Z_4 \quad Z_3; \ Z_2 \quad Z_2 \quad Z_3; \mathbf{A}_4; \mathbf{D}_{12}; \mathbf{X}$
15	$Z_3 \quad Z_5$

The case $n = p$ is a consequence of Lagrange's theorem (recall Algebra 1): a group of prime size cannot have proper subgroups, so it must be generated by any of its elements (except the unit). By the classi cation of cyclic groups, it must be isomorphic to Z_p. Other cases in the above table are already considerably more di cult.

Exercises.

1. Prove that every four-element group is isomorphic to Z_4 or $Z_2 \quad Z_2$.

2. Prove that every six-element group is isomorphic to Z_6 or $\mathbf{S}_3$.

3. Decompose the groups Z_{16}, Z_{20} and Z_{21} into a direct product of cyclic groups (i.e. write some direct product to which these groups are isomorphic, see Theorem 1.8).

4. Is there a positive integer n such that Z_n is isomorphic to (a) Z_7, (b) Z_8, (c) Z_9?

2. Quotient groups

2.1. Normal subgroups.

In this section, we discuss an essential construction in group theory, namely quotient groups (also called factor groups). For this, we first need to introduce a special type of subgroups, so-called normal subgroups. We will first get acquainted with this term.

Proposition 2.1 (equivalent definitions of normal subgroups)**.** Let $\mathbf{G}$ be a group, and $\mathbf{H} \subseteq \mathbf{G}$ be a subgroup. Then, the following statements are equivalent:

(1) $aH = Ha$ for all $a \in G$ (i.e. the left and right cosets of H are equal),
(2) $aha^{-1} \in H$ for all $h \in H$ and $a \in G$ (i.e. H is closed under conjugation by any element).

Proof. (1) $\Rightarrow$ (2). Let $h \in H$, and $a \in G$. By assumption $ah \in aH = Ha$, thus there exists an element $k \in H$, such that $ah = ka$. This implies $aha^{-1} = k \in H$.

(2) $\Rightarrow$ (1). We first prove that $aH \subseteq Ha$, so let $ah \in aH$. By (2), $k = aha^{-1} \in H$, which implies $ah = ka \in Ha$. Thus $aH \subseteq Ha$. The opposite inclusion $aH \supseteq Ha$ can be shown symmetrically, thus $aH = Ha$.

Definition. A subgroup $\mathbf{H} \subseteq \mathbf{G}$ is called a normal subgroup of $\mathbf{G}$, if one of the equivalent conditions in Proposition 2.1 holds. For short, we then also write $\mathbf{H} \trianglelefteq \mathbf{G}$.

In abelian groups, every subgroup is normal (both conditions in Proposition 2.1 are trivially satisfied). For trivial reasons, also $\{1\} \trianglelefteq \mathbf{G}$ and $\mathbf{G} \trianglelefteq \mathbf{G}$ hold in every group $\mathbf{G}$. In the following, we discuss some non-trivial examples that we already encountered:

Examples.

The special linear group $\mathbf{SL}_n(\mathbf{T})$ of matrices with determinant 1 is a normal subgroup of the general linear group $\mathbf{GL}_n(\mathbf{T})$, since condition (2) follows from the following property of determinants: $\det(AHA^{-1}) = (\det A)(\det H)(\det A)^{-1} = \det H$.

The alternating group $\mathbf{A}_n$, i.e. the subgroup of all even permutations, is a normal subgroup $\mathbf{S}_n$, by the following property of the signature: $\operatorname{sgn}(aha^{-1}) = (\operatorname{sgn} a)(\operatorname{sgn} h)(\operatorname{sgn} a)^{-1} = \operatorname{sgn} h$.

The dihedral group $\mathbf{D}_{2n}$ is not a normal subgroup of $\mathbf{S}_n$ for $n > 3$ ($(12) \in D_{2n}$, but $(1,2)\,(12)\,(1,2)^{-1} = (2134) \notin D_{2n}$).

The following fundamental observation connects normal subgroups to homomorphisms:

Proposition 2.2. The kernel of a homomorphism is a normal subgroup.

Proof. Let $\varphi : \mathbf{G} \to \mathbf{H}$ be a homomorphism. By Proposition 1.2 we already know that $\mathbf{Ker}(\varphi)$ is a subgroup of $\mathbf{G}$. The normality follows from the identity $\varphi(uau^{-1}) = \varphi(u)\,\varphi(a)\,\varphi(u)^{0} = \varphi(u)\,\varphi(u)^{0} = e$, for all $a \in \operatorname{Ker}(\varphi)$ and $u \in G$.

Exercises.

1. Find all normal subgroups of $\mathbf{D}_{12}$.

2. Show that all subgroups of the quaternion group $\mathbf{Q}_8$ are normal.

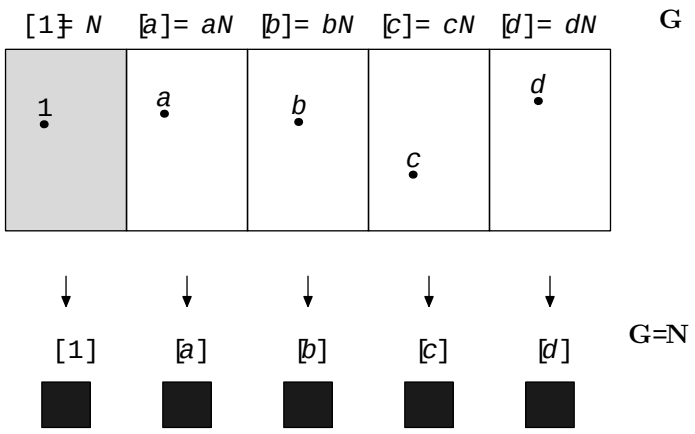


Figure 2. Construction of the quotient group G=N

3. Prove that the group $\mathbf{A}_5$ contains no proper normal subgroups. Prove that the group $\mathbf{A}_4$ contains a unique proper normal subgroup (the Klein group). (Hint: consider a normal subgroup $\mathbf{N}$ that contains a permutation of a certain cycle-type; since $\mathbf{N}$ is closed under conjugation, it then must contain every other permutations of this type (recall Algebra 1). Then think, about which subgroups are generated by the 3-cycles (:::), or by permutations of the form (::)(::).)

4. Prove that the group $\mathbf{S}_5$ contains a single proper normal subgroup, $\mathbf{A}_5$. Prove that the group $\mathbf{S}_4$ contains two proper normal subgroups, $\mathbf{A}_4$ and the Klein group. (Instructions: if $\mathbf{N}$ contains only even permutations, use the previous exercise; otherwise, show that $\mathbf{N}$ contains a transposition, and conclude that it is equal to the full group.)

5. Let $\mathbf{G}$ be a group and $\mathbf{H}$ $\mathbf{G}$, such that $[\mathbf{G} : \mathbf{H}] = 2$. Show that $\mathbf{H}$ is normal.

2.2. **The construction of quotient groups.** The idea of quotient objects can be found in various branches of mathematics. Informally speaking, we start with an object with a very ne structure (like the stars in the sky). If we move away from the object, some elements will merge (when viewed with the naked eye, we cannot distinguish stars of the same, far-away galaxy). What we see is then a quotient object (bright points in the sky) of the original object (all stars). A little more formally, we identify similar objects (those stars that are not distinguishable by our eyes). What exactly is meant by the relation of similarity depends on the speci c type of object. For groups

Definition. Let $\mathbf{G}$ be a group, and $\mathbf{N}$ a normal subgroup. We then de ne a relation on G, by

$$a \quad b \quad , \quad a \; b^{\,1} \; 2 \; N:$$

According to a theorem in Algebra 1 (see the Section on Lagrange's theorem) $a \quad b$ holds, if and only if $Na = Nb$. Therefore is an equivalence relation on G. Its blocks are the cosets of $\mathbf{N}$ in $\mathbf{G}$, and because $\mathbf{N}$ is a normal subgroup, left and right cosets are the same (Proposition 2.1), so

$$[a] = aN = Na:$$

On these blocks we de ne the following operations

$$[a] \; [b] = [a \; b] \qquad \text{and} \qquad [a]^{\,1} = [a^{\,1}]$$

(in the following lemma we will check that this is well-de ned). Together with the block $[1] = N$, these operations form a group, whose elements are the blocks of

. This group is called the quotient group (or factor group) of $\mathbf{G}$ modulo $\mathbf{N}$. For short, it is denoted by

$$\mathbf{G}/\mathbf{N} = (f[a] : a \in G\}; \; ; \; ^{-1}; [1]).$$

Lemma 2.3. Let $\mathbf{G}$ be a group and $\mathbf{N}$ a normal subgroup of $\mathbf{G}$. Then

(1) the operations and $^{-1}$ of $\mathbf{G}/\mathbf{N}$ are well-de ned, and
(2) the quotient group $\mathbf{G}/\mathbf{N}$ is indeed a group.

Proof. (1) Let us consider two blocks of given by two di erent representatives $[a] = [c]$ and $[b] = [d]$. We need to show that $[a \; b] = [c \; d]$ and $[a^{-1}] = [c^{-1}]$. Since $a \; c$ and $b \; d$, we get that $a \; c^{-1} \in N$ and $b \; d^{-1} \in N$. Using that N is closed under conjugation, we can then show that

$$(ab) \; (cd)^{-1} = abd^{-1}c^{-1} = ac^{-1}cbd^{-1}c^{-1} = \underbrace{(ac^{-1})}_{\in N} \; \underbrace{c(bd^{-1})c^{-1}}_{\in N} \in N;$$

thus $a \; b \; c \; d$. In other words $[a \; b] = [c \; d]$, which is what we wanted to prove. For the inverse, note that $ac^{-1} \in N \; , \; a^{-1}c \in N$, and thus $a^{-1} \; a^{-1} \; ac^{-1} = c^{-1}$.

(2) We next show that $\mathbf{G}/\mathbf{N}$ satis es the group axioms. The operation is associative, since $[a] \; ([b] \; [c]) = [a \; (b \; c)] = [(a \; b) \; c] = ([a] \; [b]) \; [c]$, and $[a] \; [1] = [a \; 1] = [a] = [1 \; a] = [1] \; [a] = [a]$. Similarly, it can be checked that $[a] \; [a]^{-1} = [a \; a^{-1}] = [1] = [a]^{-1} \; [a]$.

Example. Let's consider the group $\mathbf{G} = Z$, and the normal subgroup $\mathbf{H} = nZ$ for some $n > 1$. We then compute

$$a \quad b \; , \; n \mid a \quad b \; , \; a \quad b \pmod{n}.$$

The classes of this equivalence relation are of the form

$$[a] = \{k \in Z : k \quad a \pmod{n}\} = a + nZ; \quad a = 0; : : : ; n \quad 1.$$

The operations of the quotient group are $[a] + [b] = [a + b] = [a + b \bmod n]$ and $[\; a] = [n \quad a]$, i.e. like addition and subtraction on $0; : : : ; n \quad 1$ modulo n. It is not di cult to verify that $[a] \mapsto a$ is an isomorphism $Z/nZ \simeq Z_n$.

Example. Next, let us consider $\mathbf{G} = \mathbf{S}_n$ and the normal subgroup $\mathbf{H} = \mathbf{A}_n$. It holds that

$$, \qquad ^{-1} \in A_n \; , \; \operatorname{sgn}(\;) = \operatorname{sgn}(\;).$$

In other words, the equivalence relation has exactly two blocks: the set E of even permutations and the set O of odd permutations. The multiplication on the quotient group $\mathbf{G}/\mathbf{H}$ is then given by $E \quad E = O \quad O = E$ and $E \quad O = O \quad E = O$. So $\mathbf{G}/\mathbf{H}$ is isomorphic to Z .

How can we easily determine all the quotient groups of a given group? The following proposition will help us in this task, and provide a connection to homomorphic images.

Theorem 2.4 (Homomorphism Theorem)**.** let $' : \mathbf{G} \to \mathbf{H}$ be a group homomorphism.

(1) If $\mathbf{N} \quad \mathbf{Ker}(')$ is a normal subgroup of $\mathbf{G}$, then the map

$$: \mathbf{G}/\mathbf{N} \to \mathbf{H}; \qquad [a] \mapsto '(a)$$

is well-de ned, and a group homomorphism.
(2) (1st Isomorphism Theorem) $\mathbf{G}/\mathbf{Ker}(') \simeq \mathbf{Im}(')$.

Proof. (1) We first need to verify that the mapping is well defined, as it can happen that the same block is represented in two different ways, i.e. that $[a] = [b]$ for some $a \neq b$. But, in this case,

$$[a] = [b] \;,\; ab^{-1} \in N \;)\; ab^{-1} \in \mathrm{Ker}(\varphi) \;,\; \varphi(ab^{-1}) = 1 \;,\; \varphi(a) = \varphi(b);$$

thus $\Phi([a])$ does not depend on the choice of the representative a, and Φ is a well-defined mapping. Since $\Phi([ab]) = \varphi(ab) = \varphi(a)\,\varphi(b) = \Phi([a])\,\Phi([b])$, it is homomorphism.

To prove (2), we use (1) for $\mathbf{N} = \mathbf{Ker}(\varphi)$. The resulting homomorphism Φ is injective, since

$$[a] = [b] \;,\; ab^{-1} \in \mathrm{Ker}(\varphi) \;,\; \varphi(ab^{-1}) = 1 \;,\; \varphi(a) = \varphi(b).$$

If we restrict Φ to its image, it is also onto, and thus $\mathbf{G}/\mathbf{Ker}(\varphi) = \mathbf{Im}(\Phi) = \mathbf{Im}(\varphi)$.

The first isomorphism theorem is a good tool if we want to determine what a given quotient group looks like: Proving that $\mathbf{G}/\mathbf{N} \cong \mathbf{H}$ is the same, as finding a homomorphism from $\mathbf{G}$ onto $\mathbf{H}$ whose kernel is $\mathbf{N}$. We illustrate the method with several examples.

Example. What does the quotient group $\mathbb{Z}/n\mathbb{Z}$ look like? We analyzed the situation already in the examples above, so we know we should look for the homomorphism $\mathbb{Z} \to \mathbb{Z}_n$ whose kernel is the subgroup $n\mathbb{Z}$. The situation is solved by the mapping $a \mapsto a \bmod n$, which is obviously a homomorphism onto $\mathbb{Z}_n$, whose the kernel is $\{a \in \mathbb{Z} : a \bmod n = 0\} = n\mathbb{Z}$. By the 1st isomorphism theorem

$$\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n.$$

Example. What does the quotient group $\mathbf{S}_n/\mathbf{A}_n$ look like? We analyzed the situation already above and know we should look for the homomorphism $\mathbf{S}_n \to \mathbb{Z}$, whose kernel is the subgroup A_n. This homomorphism is given by $\pi \mapsto \mathrm{sgn}(\pi)$, which maps $\mathbf{S}_n$ to $\mathbb{Z}$; the kernel is given by the even permutations. By the 1st isomorphism theorem

$$\mathbf{S}_n/\mathbf{A}_n \cong \mathbb{Z}.$$

Example. What does the factor group $\mathbf{GL}_n(\mathbf{T})/\mathbf{SL}_n(\mathbf{T})$ look like? Recall that the equivalence relation corresponding to $\mathbf{SL}_n(\mathbf{T})$ is given by:

$$A \sim B \;,\; AB^{-1} \in \mathrm{SL}_n(\mathbf{T}) \;,\; \det AB^{-1} = \det A\,(\det B)^{-1} = 1 \;,\; \det A = \det B.$$

The blocks of this equivalence are thus determined by the value of the determinant, which can be any non-zero element of the field $\mathbf{T}$. At the same time, the determinant of a product is the product of determinants, i.e. $\det(AB) = \det(A)\,\det(B)$. So, the mapping $\det : \mathbf{GL}_n(\mathbf{T}) \to \mathbf{T}^*$ is a homomorphism onto the group $\mathbf{T}^*$. Its kernel is formed by the matrices of determinant 1, i.e. $\mathbf{SL}_n(\mathbf{T})$. According to the 1st isomorphism theorem,

$$\mathbf{GL}_n(\mathbf{T})/\mathbf{SL}_n(\mathbf{T}) \cong \mathbf{T}^*.$$

There are however cases, in which such an analysis does not give us any deeper insight. Other tricks can sometimes be used, such as considering the order of the quotient group and knowledge of small groups.

Example. Let us determine the quotient group $\mathbf{S}_4/\mathbf{K}$ for the normal subgroup $K = \{id; (1\,2)(3\,4); (1\,3)(2\,4); (1\,4)(2\,3)\}$. According to Lagrange's theorem, $|\mathbf{S}_4/\mathbf{K}| = [\mathbf{S}_4 : \mathbf{K}] = 24/4 = 6$. Thus, the quotient group $\mathbf{S}_4/\mathbf{K}$ is either isomorphic

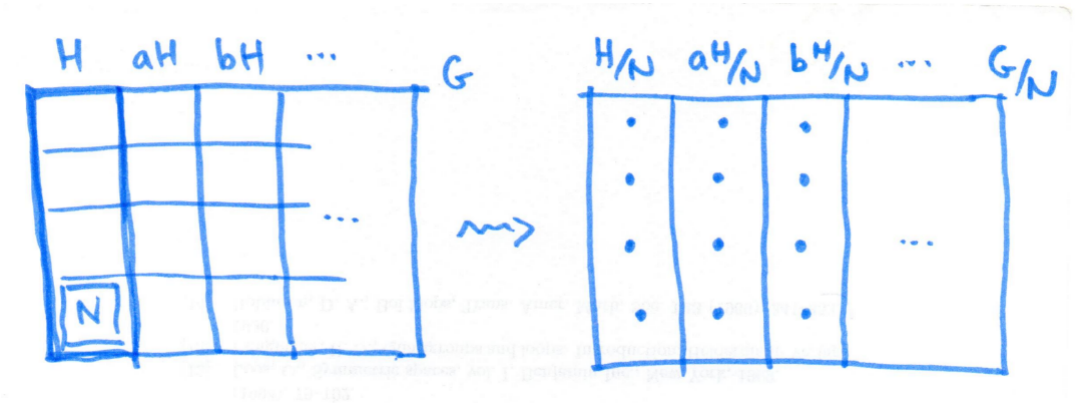


Figure 3. Illustration of the 2nd isomorphism theorem. A larger subgroup $\mathbf{H}$ determines a coarser equivalence (with larger blocks)

to the group S_3 or cyclic group Z_6. We prove that it is not abelian, which con rms the rst option:

$$[(1\ 2\ 3)]\ \ [(1\ 2\ 3\ 4)] = [(1\ 2\ 3)\ \ (1\ 2\ 3\ 4)] = [(1\ 3\ 4\ 2)];$$
$$[(1\ 2\ 3\ 4)]\ \ [(1\ 2\ 3)] = [(1\ 2\ 3\ 4)\ \ (1\ 2\ 3)] = [(1\ 3\ 2\ 4)];$$

and $[(1\ 3\ 4\ 2)] \neq [(1\ 3\ 2\ 4)]$, since $(1\ 3\ 4\ 2)\ \ (1\ 3\ 2\ 4)^{-1} = (1\ 2\ 4) \notin K$.

The 1st isomorphism theorem allows us also to give a more elegant proof of the classi cation of cyclic groups:

Alternative proof of Theorem 1.7. Let $\mathbf{G} = \langle a \rangle$ be a cyclic group and consider the mapping

$$' : Z \to \mathbf{G}; \qquad k \mapsto a^k:$$

Clearly, this map is surjective onto $\mathbf{G}$. If $'$ is also injective, then $\mathbf{G} \simeq Z$ is an isomorphism. Otherwise, $\mathbf{Ker}(') = nZ$, where $n = \mathrm{ord}(a)$, and by the 1st isomorphism theorem, $\mathbf{G} \simeq Z/nZ \simeq Z_n$:

What do quotients of quotient groups look like? This is what the 2nd theorem on isomorphism is about:

Theorem 2.5 (2nd isomorphism theorem)**.** Let $\mathbf{G}$ be a group and $\mathbf{N}$ a normal subgroup.

(1) If $\mathbf{N} \trianglelefteq \mathbf{H} \trianglelefteq \mathbf{G}$, then $\mathbf{H}/\mathbf{N}$ is a normal subgroup of $\mathbf{G}/\mathbf{N}$.
(2) If $\mathbf{K} \trianglelefteq \mathbf{G}/\mathbf{N}$, then there exists a normal subgroup $\mathbf{H} \trianglelefteq \mathbf{G}$ such that $\mathbf{K} = \mathbf{H}/\mathbf{N}$.
(3) For $\mathbf{N} \trianglelefteq \mathbf{H} \trianglelefteq \mathbf{G}$,

$$(\mathbf{G}/\mathbf{N})/(\mathbf{H}/\mathbf{N}) \simeq \mathbf{G}/\mathbf{H} \text{ holds.}$$

Proof. (1) Let $[a]; [b]$ be elements of $\mathbf{H}/\mathbf{N}$, that is, $a; b \in H$, and let $[g]$ be an element of $\mathbf{G}/\mathbf{N}$. Then $[a][b] = [ab]$ is an element of $\mathbf{H}/\mathbf{N}$ because $ab \in H$. For the same reasoning $\mathbf{H}/\mathbf{N}$ also contains $[1]$, $[a]^{-1} = [a^{-1}]$ and $[g][a][g]^{-1} = [gag^{-1}]$.

(2) Let $H = \{a \in G : [a] \in K\}$. For $a; b \in H$ and $g \in G$, $ab \in H$ holds because $[ab] = [a][b] \in K$, and for the same reason the elements of $\mathbf{H}$ also 1, a^{-1} and gag^{-1}. Obviously $\mathbf{K} = \mathbf{H}/\mathbf{N}$.

(3) Consider the homomorphism $' : \mathbf{G}/\mathbf{N} \to \mathbf{G}/\mathbf{H}$, $[a]_{\mathbf{N}} \mapsto [a]_{\mathbf{H}}$. It is well de ned because $\mathbf{N} \subseteq \mathbf{H}$ and thus $[a]_{\mathbf{N}} = [b]_{\mathbf{N}}$ implies $[a]_{\mathbf{H}} = [b]_{\mathbf{H}}$. It is a homomorphism, $'([a]_{\mathbf{N}}[b]_{\mathbf{N}}) = '([ab]_{\mathbf{N}}) = [ab]_{\mathbf{H}} = [a]_{\mathbf{H}}[b]_{\mathbf{H}} = '([a]_{\mathbf{N}})'([b]_{\mathbf{N}})$. Its

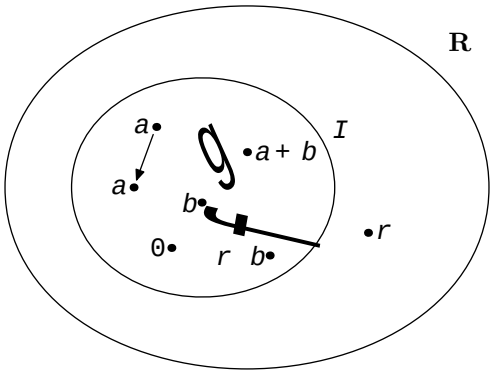


Figure 4. An ideal I of the ring $\mathbf{R}$.

image is the entire $\mathbf{G}/\mathbf{H}$ and its kernel consists of those $[a]_{\mathbf{N}}$ for which $a \in H$, i.e. $\mathbf{Ker}(\varphi) = \mathbf{H}/\mathbf{N}$. Applying the 1st isomorphism theorem, we obtain the above relation.

Exercises.

1. Using the 1st isomorphism theorem, describe what the quotient groups $\mathbb{R}/\mathbb{R}^+$, $\mathbb{R}/\{1, -1\}$ and $\mathbb{C}/\mathbf{S}$ look like, where $\mathbf{S}$ denotes the subgroup of numbers with absolute value 1.

2. Describe the groups $\mathbb{R}/\mathbb{Z}$ and $\mathbb{Q}/\mathbb{Z}$.

3. Show that $\mathbb{C}/\mathbb{C}_n \cong \mathbb{C}$.

4. Determine all quotient groups of the quaternion group $\mathbf{Q}_8$.

5. Let n be even (and think of the dihedral group $\mathbf{D}_{2n}$ as the symmetries of an n-gon). Show that the subgroup $\mathbf{N}$ of $\mathbf{D}_{2n}$, which is generated by the point reflection is normal, and that $\mathbf{D}_{2n}/\mathbf{N} \cong \mathbf{D}_n$.

3. Ideals and divisibility

Before we discuss ring homomorphisms and the construction of quotient rings, we need to introduce ideals, which fulfill a role similar to normal subgroups in group theory.

3.1. Ideals.

Definition. Let $\mathbf{R}$ be a commutative ring. An ideal of $\mathbf{R}$ is a subset $I \subset R$, such that

if $a, b \in I$, then $a \in I$ and $a + b \in I$
if $a \in I$ and $r \in R$, then $ra \in I$.

Example. The sets $n\mathbb{Z} = \{nz : z \in \mathbb{Z}\} = \{u \in \mathbb{Z} : n \mid u\}$ are ideals of the ring $\mathbb{Z}$. There are no other ideals of $\mathbb{Z}$ (proof left as an exercise; it also follows from Theorem 3.2, which we discuss later).

The above construction in the example above can be generalized to arbitrary commutative rings:

Proposition 3.1 (Definition of principal ideals)**.** Let $\mathbf{R}$ be a commutative ring and $a \in R$. Then

$$aR = \{ar : r \in R\} = \{u \in R : a \mid u\}$$

is an ideal of $\mathbf{R}$. If $\mathbf{R}$ contains a unit, then aR is the smallest ideal (with respect to inclusion), which contains a.

Proof. If u and v are divisible by a. then so are $u + v$ and u. Furthermore $a \mid ru$ for every $r \in R$. Hence aR is an ideal.

Now let I be any ideal containing the element a. Then I must contain all its multiples, i.e. $aR \subseteq I$. Since R has a unity $a \cdot 1 \in aR$, therefore aR is the smallest ideal containing the element a.

Definition. An ideal as in Proposition 3.1 is called principal. Especially $\{0\} = 0R$ and $R = 1R$ are principal ideals in every commutative ring with unity; we call them the trivial ideals.

Principal ideals nicely reflect divisibility: the transitivity of the divisibility relation immediately implies that

$a \mid b$ if and only if $bR \subseteq aR$;
$a \sim b$ if and only if $aR = bR$.

The historical motivation for the study of ideals was to solve the problem of non-unique factorisations. The idea was that the ambiguity is due to the absence of elements that separates the factors in ambiguous decompositions. For example, in $\mathbb{Z}[\sqrt{5}]$ we have $4 = 2^2 = (1 + \sqrt{5})(1 + \sqrt{5})$. If there were "ideal" irreducible elements ("ideal" meaning "hypothetical" here) p, q such that $2 = pq$, $1 + \sqrt{5} = p^2$ and $1 + \sqrt{5} = q^2$, suddenly we would have a unique, factorisation $4 = p^2q^2$. These "ideal elements" eventually turned out to be the so-called prime ideals, we will see a definition later. Modern algebraic number theory is based on the knowledge that in many rings, including $\mathbb{Z}[\sqrt{5}]$, each ideal can be uniquely decomposed into a product of prime ideals.

3.2. Principal ideal domains.

Definition. A commutative ring R, in which every ideal is a principal ideal is called a principal ideal ring; if R is additionally an integral domain, we call it a principal ideal domain (PID).

The aim of this subsection is to characterise PIDs among all integral domains by properties of divisibility.

Example. The integral domains $\mathbb{Z}$ and $\mathbf{T}[x]$, for a field $\mathbf{T}$ are principal ideal domains. More generally, every Euclidean domain has only principal ideals (Theorem 3.2).

The opposite implication does not hold, but it is not easy to find an example. Probably the easiest example of a non-Euclidean domain that has only principal ideals is $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$; the proof of this fact is rather difficult.

In Algebra 1 we showed that $\mathbb{Z}[x]$ and the multivariate polynomial rings $\mathbf{T}[x_1, \ldots, x_n]$ for a field $\mathbf{T}$ are not Euclidean, because Bézout's equality does not hold. We can also show that they are not PIDs, by constructing a non-principal ideal in them. Both examples are based on the following idea: If aR is a principal ideal that contains two coprime elements u, v, then $a \mid u$ and $a \mid v$. But this implies that $a \sim 1$, in other words, $aR = R$. In other words, any ideal containing two comprime elements must be full or not principal.

Example. The ring $\mathbb{Z}[x]$ is not a PID. To show this, let us define the subset

$$I = \{f \in \mathbb{Z}[x] : f(0) \text{ is even}\}.$$

It is not hard to see that I is an ideal of $\mathbb{Z}[x]$. At the same time, I contains the polynomials 2 and x, which are coprime. Thus I cannot be principal.

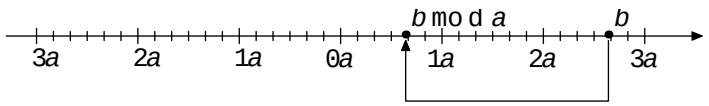


Figure 5. Illustration of Theorem 3.2 in the case of $\mathbf{R} = \mathbb{Z}$.

Example. The ring $\mathbf{R}[x_1,\dots,x_k]$ (where $\mathbf{R}$ is an arbitrary integral domain, and $k > 1$) is not a PID. For this we define

$$I = \{f \in R[x_1,\dots,x_k] : f(0,\dots,0) = 0\}.$$

It can be shown that I is an ideal of $\mathbf{R}[x_1,\dots,x_k]$. On the other hand I contains the polynomials x_1 and x_2, which are coprime. Thus I cannot be principal.

Theorem 3.2. In Euclidean domains, every ideal is principal.

Proof. Let I be an ideal of the Euclidean domain $\mathbf{R}$. If $I = \{0\}$, then $I = 0R$, so without loss of generality, let us assume that I contains an element $a \neq 0$. Let us pick such an element a for which the Euclidean norm $\delta(a)$ is minimal. We are going to prove that $I = aR$. Since I is an ideal, clearly $aR \subseteq I$; thus we only need to prove the opposite inclusion. For contradiction, let us assume that there is an element $b \in I \setminus aR$.

By the properties of Euclidean domains, there are $q, r \in R$ with $b = aq + r$ and $\delta(r) < \delta(a)$. It holds that $r \neq 0$ since b is not divisible by a, and thus $0 < \delta(r) < \delta(a)$. On the other hand

$$r = \underbrace{b}_{\in I} - \underbrace{aq}_{\in I} \in I,$$

which contradicts to the choice of a as the element of I with the smallest norm $\delta(a)$.

Proposition 3.3 (ideals in fields)**.** Let $\mathbf{R}$ be a commutative ring with unity. Then $\mathbf{R}$ is a field if and only if $\mathbf{R}$ has only trivial ideals.

Proof. ($\Rightarrow$) Every field is a Euclidean domain, so all of its ideals are principal. As $\mathbf{R}$ is a field $a \neq 0$ implies $a \mid 1$, and thus for every non-zero ideal aR we have $aR = 1R = R$.

($\Leftarrow$) For every principal ideal aR, $a \neq 0$, by assumption $aR = R = 1R$ holds, thus every element $a \neq 0$ is invertible.

We next prove an auxiliary statement about general ideals.

Proposition 3.4 (intersection, sum and union of ideals)**.** let $\mathbf{R}$ be a commutative ring.

(1) If I, J are ideals of $\mathbf{R}$, then also $I \cap J$ is an ideal $\mathbf{R}$.
(2) If I, J are ideals of $\mathbf{R}$, then so is their sum $I + J = \{a + b : a \in I, b \in J\}$. Furthermore $I + J$ is the smallest ideal containing $I \cup J$.
(3) If I_j, $j \in \mathbb{N}$, are ideals of $\mathbf{R}$ such that $I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots$, then so is their union $\bigcup_{j\in\mathbb{N}} I_j$.

Proof. (1) Let $a, b \in I \cap J$ and $r \in R$. Then $a + b,\ -a, ra$ belongs to both ideals I, J, i.e. also to their intersection.

(2) Let $a + b, c + d \in I + J$, with $a, c \in I$ and $b, d \in J$. Then clearly also $(a + b) + (c + d) = (a + c) + (b + d) \in I + J$, $-(a + b) = (-a) + (-b) \in I + J$. For any element $r \in R$ also $r(a + b) = ra + rb \in I + J$. Thus $I + J$ is an ideal. Both ideals I, J are a subset of $I + J$. Conversely, if an ideal K contains both I and J,

then it surely also contains all sums of elements from I and elements from J, i.e. $I + J \quad K$.

(3) Let $a; b \; 2 \bigcup_{j \; 2\mathbb{N}} I_j$ and $r \; 2 \; R$. Then there exist $j; k \; 2 \; \mathbb{N}$ such that $a \; 2 \; I_j$ and $b \; 2 \; I_k$, i.e. $a; b \; 2 \; I_{\max(j;k)}$, and thus $a + b; \quad a; ra \; 2 \; I_{\max(j;k)} \quad \bigcup_{j \; 2\mathbb{N}} I_j$.

By the above, the smallest ideal containing two elements $a; b$ is the ideal

$$aR + bR = far + bs : \; r; s \; 2 \; Rg;$$

and further, by induction, the smallest ideal containing the elements $a_1; : : : ; a_n$ (the so-called ideal generated by the elements $a_1; : : : ; a_n$) is

$$a_1R + \quad + a_nR = f\sum a_i r_i : \; r_1; : : : ; r_n \; 2 \; R:g$$

The elements $a_1; : : : ; a_n$ are then also called an ideal basis of $a_1R + \quad + a_nR$ (not to be confused with e.g. bases in linear algebra, as we do not assume independence). The non-principal ideals we discussed for $\mathbb{Z}[x]$ and $\mathbf{R}[x_1; : : : ; x_k]$ for example, can be written as $2\mathbb{Z}[x] + x\mathbb{Z}[x]$, respectively $x_1\mathbf{R}[x_1; : : : ; x_k] + : : : + x_k\mathbf{R}[x_1; : : : ; x_k]$.

We next discuss how PIDs can be characterized among all rings, by properties of the divisibility relation:

Theorem 3.5. Principal ideal domains are unique factorization domains. Furthermore, Bézout's identity holds in them.

Proof. Let $\mathbf{R}$ be a PID. In order to show that $\mathbf{R}$ is a UFD (recall Algebra 1), it su ces to show that in it (1) the greatest common divisor of two elements always exists and (2) there are no in nite sequences of proper divisors. Recall that for any $u; v \; u \; j \; v \; , \; vR \quad uR$ holds.

(1) Let $a; b \; 2 \; R$ be two arbitrary elements and let us de ne $I = aR + bR$. Since every ideal is principal, there exists a $c \; 2 \; R$ such that $I = cR$. Since $aR; bR \quad cR$, we have both $c \; j \; a$ and $c \; j \; b$. Further, if d is a common divisor of $a; b$, then $aR \quad dR$ and $bR \quad dR$, so $I = cR \quad dR$ and we get $d \; j \; c$. We see that $c = \gcd(a; b)$ and additionally $c \; 2 \; aR + bR$, so $c = ar + bs$ for some $r; s \; 2 \; R$, which is exactly the statement of Bézout's identity.

(2) For the sake of argument, assume that $\mathbf{R}$ has an in nite sequence of proper divisors $a_1; a_2; : : :$ (i.e., $a_{i+1} \; j \; a_i$ and $a_i \; - \; a_{i+1}$). In other words, $a_1R \quad a_2R \quad a_3R \quad : : :$. Let us de ne $I = \bigcup_{i \; 2\mathbb{N}} a_iR$. This set also forms an ideal, and hence $I = bR$ for some $b \; 2 \; I$. However, this b must be an element of some a_iR, for some $i \; 2 \; R$. But then $bR \quad a_iR \quad a_{i+1}R \quad : : : \quad I = bR$, which is a contradiction.

Summary:

Euclidean domain $=)$ principal ideal domain $=)$ unique factorization domain

The basic properties of these classes are summarized in the following table:

ring	irreducible decompositions	existence of gdc	Bézout's identity	Euclidean algorithm
Euclidean domain	X	X	X	X
PID	X	X	X	
UFD	X	X		
general				

And nally, a few examples that are worth remembering.

Euclidean domain	elds, $\mathbb{Z}$, $\mathbf{T}[x]$ ($\mathbf{T}$ eld), $\mathbb{Z}[i]$, $\mathbb{Z}[\sqrt{2}]$, $\mathbb{Z}[i\sqrt{2}]$
PID, not Euclidean	$\mathbb{Z}[\frac{1+i\sqrt{19}}{2}]$
UFD, not PID	$\mathbb{Z}[x]$, $\mathbf{R}[x_1; x_2; : : :]$ ($\mathbf{R}$ UFD)
not PID	$\mathbb{Z}[\sqrt{5}]$, $\mathbb{Z}[i\sqrt{3}]$

Exercises.

1. In Z, nd (small) generators of the following ideals: 15Z + 24Z, 15Z \ 24Z, (100Z + 60Z + 16Z) \ 21Z \ 9Z.

2. Let $\mathbf{R}$ be a PID. Prove that $aR \setminus bR = cR$, where c is the least common multiple of the elements $a; b$.

Analogous to the notion of a polynomial, we de ne a formal power series of the variable x over a commutative ring $\mathbf{R}$ with unit to be an expression $\sum_{i=0}^{1} a_i x^i$; where $a_0; a_1; : : : 2 R$ (we are not interested in convergence, only the expressions; polynomials are those power series in which there are only nitely many non-zero coe cients). We de ne the operations + and on these series using the same formulas as for polynomials, only that the upper limit $n = \deg f$ is replaced by 1 (the coe cients of the product are still de ned using a nite sum, so it is well-de ned). We denote the ring of power series over $\mathbf{R}$ by $\mathbf{R}[[x]]$.

3. Show that $\mathbf{R}[[x]]$ is indeed a commutative ring with unity, and if $\mathbf{R}$ is an integral domain, then $\mathbf{R}[[x]]$ is also an integral domain. (Attention, the second part of the problem is not a trivial generalization of the analogous statement for polynomials!)

4. Consider the integral domain $\mathbf{T}[[x]]$ of formal power series over a eld $\mathbf{T}$. Prove that

(a) the series $f = \sum_{i=0}^{1} a_i x^i \; 2 \; T[[x]]$ is invertible if and only if the coe cient a_0 is nonzero,

(b) for each series $f \; 2 \; T[[x]]$ there is an n such that $f \; k \; x^n$.

5. Is $\mathbf{T}[[x]]$ a UFD? Is it Euclidean?

4. Ring homomorphisms and quotient rings

4.1. **Homomorphisms.** Throughout this subsection, $\mathbf{R}$ and $\mathbf{S}$ will denote two arbitrary rings. Before discussing ring homomorphism, we need to de ne ideals in general, so even for rings that are not commutative.

Definition. A non-empty subset $I \quad R$ is called a (two-sided) ideal of a ring $\mathbf{R}$ if

for all $a; b \; 2 \; I$: $\quad a \; 2 \; I$ and $a + b \; 2 \; I$, and
for all $a \; 2 \; I$, $r \; 2 \; R$: $r \quad a \; 2 \; I$ and $a \quad r \; 2 \; I$.

Ring homomorphisms are naturally de ned as maps that preserve ring operations. Most of the facts in this section are direct analogies to what we discussed for group homomorphisms.

Definition. A map $' : R \; ! \; S$ is called a homomorphisms between the rings $\mathbf{R}$ and $\mathbf{S}$, if for all $a; b \; 2 \; R$ it holds that

$$'(a + b) = '(a) + '(b) \quad \text{and} \quad '(a \quad b) = '(a) \quad '(b):$$

For short, we write $' : \mathbf{R} \; ! \; \mathbf{S}$ if $'$ is a homomorphism from $\mathbf{R}$ to $\mathbf{S}$. It follows directly from Lemma 1.1 that $'(\quad a) = \quad '(a)$ for all $a \; 2 \; R$, and $'(0) = 0$.

The image of $'$ is de ned as the set

$$\mathrm{Im}(') = fb \; 2 \; S : \; b = '(a) \text{ for some } a \; 2 \; Rg:$$

The kernel of $'$ is de ned as

$$\mathrm{Ker}(') = fa \; 2 \; R : \; '(a) = 0g:$$

We next show that ideals play a similar role as normal subgroups with respect to homomorphisms.

Proposition 4.1 (image and kernel of ring homomorphisms)**.** Let $\mathbf{R}; \mathbf{S}$ be rings and $' : \mathbf{R} \; ! \; \mathbf{S}$ be a homomorphism. Then

(1) $\mathrm{Im}(')$ is a subring of $\mathbf{S}$;
(2) $\mathrm{Ker}(')$ is an ideal of $\mathbf{R}$.

Proof. A ring homomorphism is also a group homomorphism with respect to the operations $+; \ ; 0$. That is, we can use Theorem 1.2 and immediately get that the kernel and the image are closed under the operations $+; \ ; 0$. It is also easy to see that the image is closed under multiplication. It remains to complete the proof of (2): if $'(a) = 0$ and $r \in R$ is arbitrary, then $'(r \ a) = '(r) \ '(a) = '(r) \ 0 = 0$. The analogous statement holds for the product $a \ r$, i.e. $\mathrm{Ker}('')$ forms an ideal in $\mathbf{R}$.

The following statement immediately follows from Proposition 1.3:

Proposition 4.2. Let $\mathbf{R}; \mathbf{S}$ be rings and $' : \mathbf{R} \to \mathbf{S}$ a homomorphism. Then $'$ is injective, if and only if $\mathrm{Ker}('') = \{0\}$.

Similarly to groups, the following statement also holds (prove it as an exercise!).

Proposition 4.3. Let $\mathbf{R}; \mathbf{S}; \mathbf{T}$ be rings, and $' : \mathbf{R} \to \mathbf{S}$, $: \mathbf{S} \to \mathbf{T}$ homomorphisms. Then

(1) $ \ '$ is a homomorphism $\mathbf{R} \to \mathbf{T}$,
(2) if $'$ is bijective, then $'^{-1}$ is a homomorphism $\mathbf{S} \to \mathbf{R}$.

Example. Several important homomorphisms can be described by taking the remainder modulo some element. For example, in $\mathbb{Z}$, for any number $m > 0$, the map

$$'_m : \mathbb{Z} \to \mathbb{Z}_m; \qquad a \mapsto a \bmod m$$

is a homomorphism. In every polynomial ring $\mathbf{T}[x]$ ($\mathbf{T}$ eld), also every element $0 \neq m \in \mathbf{T}[x]$ gives rise to a homomorphism

$$'_m : \mathbf{T}[x] \to \mathbf{T}[x]/(m); \qquad f \mapsto f \bmod m.$$

It is not di cult to check that these maps are indeed homomorphisms and that their kernels are $m\mathbb{Z}$, or $m\mathbf{T}[x]$. Similar homomorphisms exist for every ring, in which division with a remainder is de ned.

Example. Another important family are so-called substitution homomorphisms. Consider the commutative ring $\mathbf{R} \ \mathbf{S}$ and an element $a \in S$. Then we de ne the map

$$'_a : \mathbf{R}[x] \to \mathbf{S}; \qquad f \mapsto f(a).$$

It is not di cult to verify that this is a homomorphism. If $\mathbf{R} = \mathbf{S}$, its kernel is the principal ideal $(x \ a)\mathbf{R}[x]$ and the image of the whole $\mathbf{R}$ (due to constant polynomials). In general, this is not true, e.g. for $\mathbf{R} = \mathbb{Z}$, $\mathbf{S} = \mathbb{C}$, $a = i$ we get the kernel $(x^2 + 1)\mathbb{Z}[x]$ and the image $\mathbb{Z}[i]$.

Example. Both of the above types can be combined, for example, the mapping

$$' : \mathbb{Z}[x] \to \mathbb{Z}_2; \qquad f \mapsto f(0) \bmod 2$$

is also a homomorphism, its kernel consists of those polynomials whose the absolute term is even, which is not a principal ideal.

Exercises.

1. Prove Proposition 4.3.

2. Prove that the mapping $\mathbf{R}[x; y] \to \mathbf{R}$, $f \mapsto f(0; 0)$ is a ring homomorphism. Compute its kernel and image. Is the kernel a principal ideal?

4.2. **Isomorphisms.** Bijective homomorphisms are called isomorphisms. We call two rings isomorphic, if there is an isomorphism between them; we then also write $\mathbf{R} \simeq \mathbf{S}$ for short. Everything that was said in section 1.2 about group isomorphisms applies also to ring isomorphisms:

An isomorphism can be regarded as a map that \copies" a ring: if we have a ring $\mathbf{R}$, and a bijective map $\varphi : R \to S$, then we can de ne a ring on the set S by the operations

$$a \oplus b = \varphi(\varphi^{-1}(a) + \varphi^{-1}(b));$$

for $\in \{+, \cdot\}$.

The map φ^{-1} then is an isomorphism between the new ring $\mathbf{S}$ and the old ring $\mathbf{R}$. The rings $\mathbf{R}$ and $\mathbf{S}$ are \isomorphic copies" of the other since only a \ renaming" of the elements by φ has occurred. Every isomorphism can be viewed in this way.

Example. It is easy to see that the set of matrices

$$S = \left\{ \begin{pmatrix} a & b \\ b & a \end{pmatrix} : a, b \in \mathbb{R} \right\}$$

forms a subring of $\mathbf{M}_2(\mathbb{R})$, the ring of all 2×2 matrices over the real numbers. The map

$$\varphi : \mathbb{C} \to \mathbf{S}, \qquad a + bi \mapsto \begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

is an isomorphism, since it is bijective and for all $a, b, c, d \in \mathbb{R}$:

$$\begin{aligned} \varphi((a + bi) + (c + di)) &= \varphi((a + c) + (b + d)i) = \begin{pmatrix} a+c & b+d \\ b+d & a+c \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ b & a \end{pmatrix} + \begin{pmatrix} c & d \\ d & c \end{pmatrix} = \varphi(a + bi) + \varphi(c + di) \end{aligned}$$

and

$$\begin{aligned} \varphi((a + bi) \cdot (c + di)) &= \varphi((ac - bd) + (ad + bc)i) = \begin{pmatrix} ac-bd & ad+bc \\ ad+bc & ac-bd \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ b & a \end{pmatrix} \cdot \begin{pmatrix} c & d \\ d & c \end{pmatrix} = \varphi(a + bi) \cdot \varphi(c + di). \end{aligned}$$

Thus, the eld $\mathbb{C}$ is isomorphic to the matrix ring $\mathbf{S}$; the isomorphism identi es the number $a + bi$ with the corresponding matrix $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$.

Example. Let $\mathbf{R}$ be an arbitrary ring of characteristic $n > 0$. It is easy to see that the elements of the form $1 + \cdots + 1$ form a subring $\mathbf{P}$, which is sometimes called prime ring. It is easy to see that mapping

$$\varphi : \mathbb{Z}_n \to \mathbf{P}, \qquad 0 \mapsto 0, \quad k \mapsto \underbrace{1 + \cdots + 1}_{k};$$

is a ring isomorphism. An analogous statement applies to rings of characteristic 0. Then the prime ring also includes elements of the form $-(1 + \cdots + 1)$, and is isomorphic to $\mathbb{Z}$.

Another important example of a ring isomorphism, which we saw in Algebra 1, is the modular map in the proof of the Chinese remainder theorem. It is not only a ring isomorphism but also a group isomorphism when restricted to the corresponding multiplication groups:

Proposition 4.4 (algebraic version of the Chinese remainder theorem)**.** Let $m_1, \ldots, m_n$ be pairwise coprime integers, and let $M = m_1 \cdots m_n$. Then, the map

$$\begin{aligned} \varphi : \mathbb{Z}_M &\to \mathbb{Z}_{m_1} \times \cdots \times \mathbb{Z}_{m_n} \\ a &\mapsto (a \bmod m_1, \ldots, a \bmod m_n). \end{aligned}$$

is an isomorphism. It's restriction $\varphi|_{\mathbb{Z}_M^\times}$ is a group isomorphism

$$\mathbb{Z}_M^\times \simeq \mathbb{Z}_{m_1}^\times \times \cdots \times \mathbb{Z}_{m_n}^\times.$$

Proof. We already showed in Algebra 1 that $'$ is bijective. We next verify that it is a homomorphism: for both operations, $\ast \in \{+, \cdot\}$:

$$\begin{aligned}'(a) \ast '(b) &= (a \bmod m_1, \ldots, a \bmod m_n) \ast (b \bmod m_1, \ldots, b \bmod m_n) \\ &= ((a \ast b) \bmod m_1, \ldots, (a \ast b) \bmod m_n) = '(a \ast b \bmod M),\end{aligned}$$

where the last equality uses the fact that all m_i divide M.

We leave the proof of the second part as an exercise: prove that invertible elements modulo M are mapped onto invertible elements modulo an individual m_i, i.e. $'|_{Z_M}$ is a bijection onto the set $Z_{m_1} \times \cdots \times Z_{m_n}$.

Analogously, the Chinese remainder theorem for polynomials can be phrased using isomorphisms (left as exercise).

Exercises.

Decide whether $\mathbb{Q}(i) \cong \mathbb{Q}(\sqrt{2})$, or not. Think about the answer for the general pair $\mathbb{Q}(\sqrt{r})$, $\mathbb{Q}(\sqrt{s})$. (Hint: homomorphism preserves the solution of the equation $x^2 = r$.)

4.3. **The construction of quotient rings.**

In Algebra 1, we already met a special case of the construction of quotient rings: we declared two polynomials in $\mathbf{R}[x]$ to be equivalent if they have the same remainder modulo a given polynomial m, and we defined the quotient ring $\mathbf{R}[x]/(m)$ on the resulting equivalence classes, by computing $+$ and $\cdot$ modulo m. This construction works in a more general setting: we can replace the polynomial m with any ideal I and declare two elements similar if their difference is in I. We will see that, by the properties of ideals, this construction is again a ring. As every ideal is also a normal subgroup of the additive group $(R, +, 0, -)$, we only have to focus on the multiplication in the following.

Definition. Let $\mathbf{R}$ be a ring, and I an ideal. Then we define an equivalence relation on R by

$$a \sim b \quad :\Leftrightarrow \quad a - b \in I.$$

Then $a \sim b$ holds, if and only $a + I = b + I$, i.e. the equivalence classes are of the form $[a] = a + I$. We define the following operations on these blocks:

$$[a] + [b] = [a + b]; \qquad -[a] = [-a]; \qquad [a] \cdot [b] = [a \cdot b].$$

(in the following lemma, we verify that they are well-defined). The set of equivalence blocks, together with the above operations is called the quotient ring (or factor ring) of $\mathbf{R}$ modulo I,

$$\mathbf{R}/I = (\{[a] : a \in R\}, +, -, \cdot, [0]).$$

Lemma 4.5. Let $\mathbf{R}$ be a ring and I an ideal of $\mathbf{R}$. Then:

(1) the operations in Definition 4.3 are well-defined,
(2) the quotient ring $\mathbf{R}/I$ is indeed a ring.

Proof. We already know from the construction of the quotient group $(\{[a] : a \in R\}, +, -, [0])$ that addition and subtraction are well-defined. For the multiplication, let us assume $[a] = [c]$ and $[b] = [d]$. We then need to show that $[ab] = [cd]$. by our assumption $a - c \in I$ and $b - d \in I$ holds, and we compute

$$ab - cd = a\underbrace{(b - d)}_{\in I} + \underbrace{(a - c)}_{\in I} d \in I,$$

hence $ab \sim cd$, i.e. $[ab] = [cd]$. Similar to groups, it turns out that the operations defined in this way satisfy all ring axioms, including commutativity and the existence of a unit if these hold in the original ring $\mathbf{R}$ (note however, that the property of being an integrity domain does not have to be preserved, see Section 4.4).

Example. Consider the commutative ring $\mathbf{R} = Z$ and the ideal $I = nZ$. It holds that

$$a \equiv b \iff n \mid a - b \iff a \equiv b \pmod{n}.$$

As with groups, it is not hard to show that $Z/nZ \cong Z_n$.

Similar to the group case, a version of the homomorphism theorem and first isomorphism theorem holds for rings:

Theorem 4.6 (Homomorphism Theorem)**.** Let $\varphi : \mathbf{R} \to \mathbf{S}$ be a ring homomophism.

(1) If $I \subseteq \mathrm{Ker}(\varphi)$ is an ideal of $\mathbf{R}$, then the map

$$\psi : \mathbf{R}/I \to \mathbf{S}, \qquad [a] \mapsto \varphi(a)$$

is well-defined and a ring homomorphism.

(2) [1st isomorphism theorem] $\mathbf{R}/\mathrm{Ker}(\varphi) \cong \mathbf{Im}(\varphi)$.

Proof. The proof is the same as in the group version (Theorem 2.4), all you need to check that all defined mappings are additionally also ring homomorphisms.

Example. The homomorphism $Z \to Z_n$, $a \mapsto a \bmod n$, induces a ring isomorphism $Z/nZ \cong Z_n$.

If $\mathbf{R}$ is a commutative ring and $I = mR$ is a principal ideal of it, then

$$a \equiv b \iff m \mid a - b \iff a \equiv b \pmod{m}.$$

In general, if $\mathbf{R}$ admits a division with remainder (e.g. $\mathbf{R} = Z$ or $\mathbf{R} = \mathbf{T}[x]$, $\mathbf{T}$ is a field), the elements of $\mathbf{R}/mR$ can be represented as all possible remainders after division by the element m and the operations in $\mathbf{R}/mR$ are the operations in the original ring modulo m, since

$$[a] + [b] = [a + b] = [a + b \bmod m]; \qquad [a] \cdot [b] = [a \cdot b] = [a \cdot b \bmod m].$$

In particular, for $\mathbf{R} = \mathbf{T}[x]$ we see that the quotient ring construction from Algebra 1, and the quotient ring construction from this section, are essentially identical: The 1st isomorphism theorem applied to the modular mapping $\mathbf{T}[x] \to \mathbf{T}[x]/(m)$ yields the isomorphism

$$\mathbf{T}[x]/mT[x] \cong \mathbf{T}[x]/(m);$$

in which a polynomial f of degree $< \deg m$ uniquely corresponds to the corresponding block $[f]$. The notation is usually mixed, both notations $\mathbf{T}[x]/mT[x]$ and $\mathbf{T}[x]/(m)$ are used for the two formally different but isomorphic constructions.

Example. What does the quotient group $\mathbf{T}[x]/(x - a)$ for $a \in T$ look like? Let us consider the map

$$\mathbf{T}[x] \to \mathbf{T}; \qquad f \mapsto f(a).$$

Its image is $\mathbf{T}$ and its kernel is

$$\{f \in T[x] : f(a) = 0\} = \{f \in T[x] : x - a \mid f\} = (x - a)T[x].$$

By the first isomorphism theorem $\mathbf{T}[x]/(x - a) \cong \mathbf{T}$.

For polynomials of a higher degree, the situation is more complicated.

Example. What does the quotient ring $Q[x]/(x^2 + 1)$ look like? Let us consider the homomorphism

$$Q[x] \to Q(i); \qquad f \mapsto f(i).$$

Its image is $\mathbb{Q}(i)$ and its kernel is

$$\begin{aligned}\{f \in \mathbb{Q}[x] : f(i) = 0\} &= \{f \in \mathbb{Q}[x] : f(i) = f(-i) = 0\} \\ &= \{f \in \mathbb{Q}[x] : x - i \mid f,\ x + i \mid f\} \\ &= \{f \in \mathbb{Q}[x] : (x - i)(x + i) = x^2 + 1 \mid f\} \\ &= (x^2 + 1)\mathbb{Q}[x].\end{aligned}$$

By the first isomorphism theorem $\mathbb{Q}[x]/(x^2 + 1) \simeq \mathbb{Q}(i)$.

Example. What does the quotient ring $\mathbb{Q}[x]/(x^2 - 1)$ look like? Let us consider the homomorphism

$$\mathbb{Q}[x] \to \mathbb{Q} \times \mathbb{Q}, \qquad f \mapsto (f(1), f(-1)).$$

This is a map to $\mathbb{Q} \times \mathbb{Q}$, and its kernel is

$$\begin{aligned}\{f \in \mathbb{Q}[x] : f(1) = f(-1) = 0\} &= \{f \in \mathbb{Q}[x] : x - 1 \mid f,\ x + 1 \mid f\} \\ &= \{f \in \mathbb{Q}[x] : (x - 1)(x + 1) = x^2 - 1 \mid f\} \\ &= (x^2 - 1)\mathbb{Q}[x].\end{aligned}$$

By the first isomorphism theorem $\mathbb{Q}[x]/(x^2 - 1) \simeq \mathbb{Q} \times \mathbb{Q}$.

Finally, let us also look at an example with a non-principal ideal:

Example. What does the quotient ring $\mathbb{Z}[x]/I$ look like, where $I = \{f \in \mathbb{Z}[x] : m \mid f(0)\}$? Two polynomials f, g are equivalent modulo I, if $f - g \in I$, i.e. if $m \mid f(0) - g(0)$. This is equivalent ot $f(0) \equiv g(0) \pmod{m}$. Thus, there are exactly m equivalence classes. It is not hard to see that the map

$$\mathbb{Z}[x] \to \mathbb{Z}_m, \qquad f \mapsto f(0) \bmod m$$

is a homomorphism, its kernel is I, and thus $\mathbb{Z}[x]/I \simeq \mathbb{Z}_m$.

Similar to the group case, a 2nd isomorphism theorem also holds for rings (proof left as exercise!).

Proposition 4.7 (2nd isomorphism theorem)**.** Let $\mathbf{R}$ be a ring, and I an ideal of it.

(1) If also $I \subset J$ is an ideal of $\mathbf{R}$, then $J/I = \{[a] : a \in J\}$ is an ideal of $\mathbf{R}/I$.
(2) If K is an ideal of $\mathbf{R}/I$, then there exists an ideal J of $\mathbf{R}$ such that $K = J/I$.
(3) In both cases it holds that

$$(\mathbf{R}/I)\big/(J/I) \simeq \mathbf{R}\big/J.$$

Exercises.

1. What do the quotient rings $\mathbb{Z}[x]/I$ look like, where (a) $I = (x - 1)\mathbb{Z}[x]$, (b) $I = (x^2 + 1)\mathbb{Z}[x]$, (c) $I = (x^2 - 1)\mathbb{Z}[x]$? Note that the last one is not isomorphic to $\mathbb{Z} \times \mathbb{Z}$: think carefully about what the image of the substitution homomorphism is!

2. What does the quotient ring $\mathbf{T}[x]/(x^4 - 4)$ look like for $\mathbf{T} = \mathbb{Q}, \mathbb{R}, \mathbb{C}$?

3. What does the quotient ring $\mathbf{R}[x, y]/I$ look like, where (a) $I = yR[x, y]$, (b) $I = (x + y)R[x, y]$, (c) $I = \{f : f(0, 0) = 0\}$?

4. Prove the second isomorphism theorem for rings.

4.4. **Quotient rings modulo maximal and prime ideal.** In this section, we will show a generalization of the result that a ring $\mathbf{T}[\ ]/(m)$ is a field if $\mathbf{T}$ is an integral domain and m is an irreducible element in $\mathbf{T}[\ ]$. In general, the situation is more complicated: it may happen that the quotient ring according to the ideal is an integral domain, but not a field.

Definition. We say an ideal I of a ring $\mathbf{R}$ is

a prime ideal, if for all $a, b \in R$ with $ab \in I$, either $a \in I$ or $b \in I$ holds;
maximal, if I is maximal with respect to the inclusion relation, i.e. there is no ideal J such that $I \subsetneq J \subsetneq R$.

Example. Left as an exercise:

An ideal $n\mathbb{Z}$ of $\mathbb{Z}$ is maximal if and only if it is a prime ideal, if and only if n is prime.
An ideal $fT[x]$ of $\mathbf{T}[x]$ is maximal if and only if it is a prime ideal, if and only if f is irreducible in $\mathbf{T}[x]$

An analogous statement holds in all principal ideal domains. However, in general, maximal ideals and prime ideals are different concepts. For example, the ideal $I = x\mathbb{Z}[x]$ in the domain $\mathbb{Z}[x]$

is a prime ideal, since the polynomial x is irreducible,
is not maximal, since the ideal $\{f \in \mathbb{Z}[x] : 2 \mid f(0)\}$ is bigger.

Theorem 4.8 (quotient rings modulo prime and maximal ideals)**.** Let $\mathbf{R}$ be a commutative ring with unity, and I an ideal. Then

(1) $\mathbf{R}/I$ is an integral domain, if and only if I is prime;
(2) $\mathbf{R}/I$ is a field, if and only if I is maximal.

Proof. (1) By definition, the quotient ring $\mathbf{R}/I$ is an integral domain, if for all $a, b \in R$ such that $[a][b] = [0]$ either $[a] = [0]$ or $[b] = [0]$ holds. Since $[0] = I$, this is equivalent to the statement that $ab \in I$ implies $a \in I$ or $b \in I$, which is the definition of I being a prime ideal.

(2) By Proposition 3.3 $\mathbf{R}/I$ is a field if and only if it contains no proper ideals. The second isomorphism theorem says that $\mathbf{R}/I$ contains a proper ideal K, if and only if $K = J/I$, for an ideal $I \subsetneq J \subsetneq R$ of $\mathbf{R}$. Thus $\mathbf{R}/I$ is a field, if and only if I is maximal.

Example. Recall the examples of $\mathbb{Q}[x]/(f)$ from Section 4.3:

the polynomial $x^2 + 1$ is irreducible, so the ideal $(x^2 + 1)\mathbb{Q}[x]$ is maximal, and hence $\mathbb{Q}[x]/(x^2 + 1) \cong \mathbb{Q}(i)$ is a field;
the polynomial $x^2 - 1$ is not irreducible, so the ideal $(x^2 - 1)\mathbb{Q}[x]$ is not maximal (e.g. the ideal $(x - 1)\mathbb{Q}[x]$ is larger), and indeed, $\mathbb{Q}[x]/(x^2 - 1) \cong \mathbb{Q} \times \mathbb{Q}$ is not a field.

Example. Recall the example of the ideal $I = x\mathbb{Z}[x]$ in the ring $\mathbb{Z}[x]$. I is a prime ideal that is not maximal, and indeed, the factor ring $\mathbb{Z}[x]/I \cong \mathbb{Z}$ (given by substitution homomorphism) is an integral domain, but not a field.

Exercises.

1. Explain why $\mathbb{R}[\ ]/(\ ^2 + 1) \cong \mathbb{C}$ but $\mathbb{C}[\ ]/(\ ^2 + 1) \cong \mathbb{C} \times \mathbb{C}$.

2. Using the 2nd isomorphism theorem, prove that $\mathbb{Z}[i]/p\mathbb{Z}[i] \cong \mathbb{Z}_p[\ ]/(\ ^2 + 1)$. For which p do we get a field? Find an argument for $\mathbb{Z}[i]/p\mathbb{Z}[i]$ and $\mathbb{Z}_p[\ ]/(\ ^2 + 1)$ and deduce which primes are irreducible elements of $\mathbb{Z}[i]$.

3. In $\mathbb{Z}[\sqrt{5}]$, the element 4 has the two distinct decompositions $4 = 2^2 = (1+\sqrt{5})(\ 1+\sqrt{5})$. Show that $I = 2\mathbb{Z}[\sqrt{5}] + (1 + \sqrt{5})\mathbb{Z}[\sqrt{5}]$ is a prime ideal, but not principal and that $4 \in \mathbb{Z}[\sqrt{5}] = I \cdot I$. Compare this with the historical remark at the end of Section 3.1.

Algebraic number fields and roots of polynomials

5. Ring and field extensions

5.1. **Definition.**
In this section, we de ne the general notion of ring extension and show that its elements can be expressed in terms of polynomial operations.

Definition. Let $\mathbf{R}$ $\mathbf{S}$ be commutative rings and let a $a_1;:::;a_n$ 2 S. We de ne
$\mathbf{R}[a_1;:::;a_n]$ to be the smallest subring of $\mathbf{S}$ containing R and $a_1;:::;a_n$;
called the extension ring of $\mathbf{R}$ in $\mathbf{S}$ generated by the elements $a_1;:::;a_n$.

If $\mathbf{R};\mathbf{S}$ are elds, we instead de ne
$\mathbf{R}(a_1;:::;a_n)$ to be the smallest sub eld of $\mathbf{S}$ containing R and $a_1;:::;a_n$;
called the extension eld of $\mathbf{R}$ in $\mathbf{S}$ generated by the elements $a_1;:::;a_n$

Example (Gaussian integers)**.** The ring of Gaussian integers can be written as an extension ring of Z in C generated by i, that is Z[i]. Analogously, Gaussian rational numbers can be written as the smallest subring of C containing both Q and i, that is Q[i]. Observe that Q[i] is indeed a eld, since

$$(a + bi)^{\ 1} = \frac{a}{a^2 + b^2} \quad \frac{b}{a^2 + b^2} i \ 2 \ \mathrm{Q}[i];$$

hence Q[i] = Q(i). In the same way, we have R[i] = R(i) = C.

The construction of Gaussian integers and rationals can be naturally generalized in two di erent ways: extending Z either by means of a square root of other numbers, instead of $i = \sqrt{\ 1}$, or by higher complex roots of one.

Example (Quadratic extensions)**.** For any s 2 Z consider the quadratic extensions

$$\mathrm{Z}[\sqrt{s}] = fa + b\sqrt{s} : \ a;b \ 2 \ \mathrm{Z}g \quad \mathrm{C};$$
$$\mathrm{Q}[\sqrt{s}] = \mathrm{Q}(\sqrt{s}) = fa + b\sqrt{s} : a;b \ 2 \ \mathrm{Q}g \quad \mathrm{C}:$$

It is easy to see that the right-hand side is a subring (resp. a sub eld) of C.

Example (Cyclotomic extensions)**.** For $_n = e^{2 \ i=n}$ (the so-called n-th root of unity) consider the n-th cyclotomic extensions

$$\mathrm{Z}[\ _n] = fa_0 + a_1 \ _n + a_2 \ _n^{\ 2} + ::: + a_{n \ 1} \ _n^{\ n \ 1} : \ a_0;:::;a_{n \ 1} \ 2 \ \mathrm{Z}g \quad \mathrm{C};$$
$$\mathrm{Q}[\ _n] = \mathrm{Q}(\ _n) = fa_0 + a_1 \ _n + a_2 \ _n^{\ 2} + ::: + a_{n \ 1} \ _n^{\ n \ 1} : \ a_0;:::;a_{n \ 1} \ 2 \ \mathrm{Q}g \quad \mathrm{C}:$$

For n = 3 we obtain the Eisenstein numbers, for n = 4 the Gaussian numbers. We remark that, in general, di erent such sums can describe the same element. For instance, for n = 3 we have $_3^{\ 2} = \ 1 \ _3$. The proof that Q[$_n$] = Q($_n$) is not as simple as for quadratic extensions. Later we will show a general statement (Proposition 6.4), from which this fact immediately follows.

Example. We can also consider extension rings generated by more than one element. For instance

$$\mathrm{Z}[\sqrt{2};\sqrt{3}] = fa + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} : \ a;b;c;d \ 2 \ \mathrm{Z}g:$$

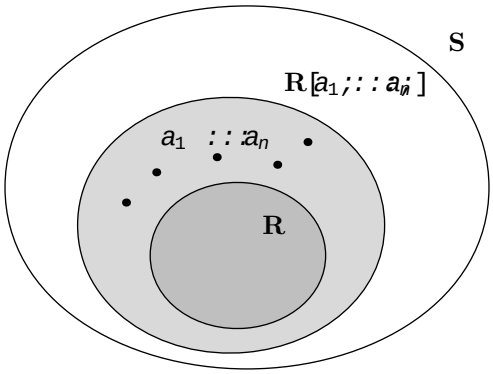


Figure 6. Extension ring $\mathbf{R}[a_1;:::;a_n]$ $\mathbf{S}$.

Proposition 5.1 (Ring extension structure theorem)**.** Let $\mathbf{R}$ $\mathbf{S}$ be commutative rings with unity and let $a \in S$. Then

$$\mathbf{R}[a] = \{f(a) : f \in R[x]\}$$
$$= \{u_0 + u_1 a + ::: + u_n a^n : n \in \mathbb{N};\ u_0;:::;u_n \in R\}.$$

If $\mathbf{R}$ $\mathbf{S}$ are fields, then

$$\mathbf{R}(a) = \left\{ \frac{f(a)}{g(a)} : f;g \in R[x];\ g(a) \neq 0 \right\}.$$

Proof. Let us denote by $M = \{f(a) : f \in R[x]\}$. We have to prove that the set M

(1) is a subring of $\mathbf{S}$,
(2) contains $R \cup \{a\}$,
(3) is the smallest subring $\mathbf{S}$ satisfying the previous conditions.

(1) Let $f(a);g(a) \in M$, where $f;g \in R[x]$. Their sum $f(a) + g(a) = (f + g)(a)$ is also in M, because $f + g \in R[x]$, in a similar way we conclude for the product and for the elements $f(a)$. Moreover, by choosing $f = 0$ we obtain $0 \in M$.

(2) By choosing constant polynomials, we get R M. By choosing $f = x$ we also obtain $a \in M$.

(3) Consider any subring $\mathbf{U}$ containing $R \cup \{a\}$. This subring must contain all the powers a^i, their arbitrary multiples by elements of R, and also arbitrary sums of these multiples. That is, it must contain all elements of the form $u_0 + u_1 a + ::: + u_n a^n$, where $u_0;:::;u_n \in R$, therefore M U. The second statement involving fields can be done analogously, however, particular attention must be paid to the inverses (exercise!).

Example. Consider the quadratic extensions $\mathbb{Z}[\sqrt{s}]$. Since $\sqrt{s}^2 = s$, $\sqrt{s}^3 = s\sqrt{s}$, etc., the value of every polynomial $f \in \mathbb{Z}[x]$ on the element $\sqrt{s}$ will be equal to $a + b\sqrt{s}$, $a;b \in \mathbb{Z}$. Hence $\mathbb{Z}[\sqrt{s}] = \{f(\sqrt{s}) : f \in \mathbb{Z}[x]\} = \{a + b\sqrt{s} : a;b \in \mathbb{Z}\}$.

Analogously, for cyclotomic extensions we get the expression $\mathbb{Z}[\varepsilon_n] = \{a_0 + a_1 \varepsilon_n + ::: + a_{n-1} \varepsilon_n^{n-1} : a_0;:::;a_{n-1} \in \mathbb{Z}\}$, because $\varepsilon_n^n = 1$, $\varepsilon_n^{n+1} = \varepsilon_n$, etc.

For every field $\mathbf{T}$, clearly the inclusion $\mathbf{T}[a]$ $\mathbf{T}(a)$ holds. But under which conditions do we get equality $\mathbf{T}[a] = \mathbf{T}(a)$? In Section 6 we will show that this happens when a is a so-called algebraic element, that is, when it is the root of some nonzero polynomial of $\mathbf{T}[x]$. One implication can be proven right away:

Proposition 5.2. Let $\mathbf{T}$ $\mathbf{S}$ be fields and $a \in S$ an element that is not a root of any nonzero polynomial in $\mathbf{T}[x]$. Then $\mathbf{T}[a] \neq \mathbf{T}(a)$.

Proof. From Theorem 5.1 we have $\mathbf{T}[a] = \{f(a) : f \in T[x]\}$. If there was an element a^{-1} in this set, then there would be a polynomial $f \in T[x]$ such that $f(a) = a^{-1}$, that is $af(a) = 1$, and thus a would be a root of the non-zero polynomial $xf - 1 \in T[x]$, which is impossible.

Exercises.

1. Describe the elements of the following rings and order them by inclusion
(a) $Z[\sqrt{6}]$, $Z[\sqrt{24}]$, $Z[\sqrt{2}, \sqrt{3}]$.
(b) $Q[\sqrt{6}]$, $Q[\sqrt{24}]$, $Q[\sqrt{2}, \sqrt{3}]$.

2. Do the rings $Q[\sqrt{2}, \sqrt{3}]$ and $Q[\sqrt{2} + \sqrt{3}]$ coincide? What is the answer for $Z[\sqrt{2}, \sqrt{3}]$ and $Z[\sqrt{2} + \sqrt{3}]$?

3. Describe the elements of $Q[\sqrt[3]{s}]$ and $Q(\sqrt[3]{s})$. Do they coincide?

4. For which $s, t \in Z$ does $\sqrt{s} \in Z[\sqrt{t}]$ hold? Consider s, t, such that they are not divisible by the square of a prime number.

5. Let $\mathbf{R}$ be a subring of $\mathbf{S}$, and let $u_1, \ldots, u_n \in S$. Show that

$$\mathbf{R}[u_1, \ldots, u_n] = \{f(u_1, \ldots, u_n) : f \in R[x_1, \ldots, x_n]\}.$$

If they are fields, show that

$$\mathbf{R}(u_1, \ldots, u_n) = \left\{ \frac{f(u_1, \ldots, u_n)}{g(u_1, \ldots, u_n)} : f, g \in R[x_1, \ldots, x_n],\ g(u_1, \ldots, u_n) \neq 0 \right\}.$$

5.2. **Field extensions as vector spaces.**

In this section, we study the so-called degree of an extension, that is the dimension of a larger field as a vector space over its subfield. But what do we mean here, exactly, by dimension?

A field extension is a pair of fields $\mathbf{T}, \mathbf{S}$ such that $\mathbf{T} \subset \mathbf{S}$. We say $\mathbf{T}$ is a subfield of $\mathbf{S}$, or that $\mathbf{S}$ is an extension of $\mathbf{T}$.

The main point of this section is to understand that the field $\mathbf{S}$ can be considered as a vector space over the field $\mathbf{T}$: the addition and subtraction of this vector space are simply the addition and subtraction of $\mathbf{S}$. Instead of the multiplication as an operation $S \times S \to S$ we only consider the restriction $T \times S \to S$. If we consider the elements of the larger field $\mathbf{S}$ to be vectors, the elements of the smaller field $\mathbf{T}$ to be scalars, this describes the multiplication of a scalar times a vector. We will denote the resulting vector space by $\mathbf{S}_\mathbf{T}$.

Note that this is indeed a vector space: the additive structure $(S, +, -, 0)$ is an Abelian group, and for all $a, b \in T$ (scalars), $v, w \in S$ (vectors) each of the axioms of a vector spaces holds: $a(bv) = (ab)v$ follows from the associativity of multiplication, $1v = v$ from the unit property, and $a(v + w) = av + aw$ and $(a + b)v = av + bv$ from distributivity.

Definition. The dimension of the vector space $\mathbf{S}_\mathbf{T}$ is called the degree of the extension, and is denoted by

$$[\mathbf{S} : \mathbf{T}] = \dim \mathbf{S}_\mathbf{T}.$$

If the degree of $[\mathbf{S} : \mathbf{T}]$ is finite, we say that it is an extension of finite degree.

Examples.

$[C : R] = 2$. Every complex number can be written uniquely as $a + bi$, $a, b \in R$, that is the elements $1, i$ form a basis of the space C_R.
Analogously, for a non-square integer s, the degree $[Q(\sqrt{s}) : Q] = 2$. Hence, the elements $1, \sqrt{s}$ form a basis of the space $Q(\sqrt{s})_Q$.
$[Q(\sqrt{2}, \sqrt{3}) : Q] = 4$, the basis of the space $Q(\sqrt{2}, \sqrt{3})_Q$ consists, for example, of the elements $1, \sqrt{2}, \sqrt{3}, \sqrt{6}$.

Be careful, for $\omega_3 = e^{2\pi i/3}$ the degree $[\mathbb{Q}(\omega_3) : \mathbb{Q}] = 2$ and not 3: the elements $1, \omega_3, \omega_3^2$ are linearly dependent because $\omega_3^2 = -1 - \omega_3$.

If u is a transcendental number (i.e. not algebraic, like the constants e or π), the degree of $[\mathbb{Q}(u) : \mathbb{Q}]$ is infinite (countable): a linearly independent set consists, for example, of elements $1, u, u^2, \ldots$ (see Theorem 6.7).

The degree of $[\mathbb{R} : \mathbb{Q}]$ is even uncountable: spaces of countable dimension over a countable field are countable, while real numbers are uncountable.

It is natural to define the notion of the prime ring. For any ring with unity $\mathbf{R}$ consider the mapping

$$\mathbb{Z} \to \mathbf{R}, \qquad n \mapsto \underbrace{1 + \ldots + 1}_{n}.$$

It is easy to see that it is a homomorphism; its image is called prime ring of the ring $\mathbf{R}$, and its kernel is the ideal $n\mathbb{Z}$, where n is the characteristic of the ring $\mathbf{R}$. From the First Isomorphism Theorem, we conclude that the prime ring of any ring is isomorphic either to the ring $\mathbb{Z}$, and in this case, it has characteristic 0, or to the ring $\mathbb{Z}/n\mathbb{Z} \simeq \mathbb{Z}_n$ when the characteristic is n.

Now, repeat the construction with the field $\mathbf{T}$. Then, the prime field of $\mathbf{T}$ is its smallest subfield. It certainly contains the prime ring, but it also must contain the inverse of every non-zero element in it. The characteristic of the field is either 0 or the prime number p. In the second case, the prime ring is already a field (isomorphic to $\mathbb{Z}_p$), so the two concepts agree. In the case of characteristic 0, the prime field consists of all the fractions ab^{-1}, where a, b are elements of the prime ring, that is, the prime field is isomorphic to $\mathbb{Q}$.

Naturally, each field is an extension of its prime field. In the particular case of finite fields, we obtain very interesting consequences of the vector space construction of a field over its prime field.

Proposition 5.3. The number of elements of a finite field is a power of a prime number.

Proof. The finite field $\mathbf{T}$ of characteristic p is an extension of its prime field $\mathbf{P} \simeq \mathbb{Z}_p$. That is, the vector space $\mathbf{T}_{\mathbf{P}}$ is isomorphic to $(\mathbb{Z}_p)^k$, where $k = [\mathbf{T} : \mathbf{P}]$, that is, it has p^k elements.

6. Algebraic elements and extensions of finite degree

6.1. Algebraic and transcendental numbers.

One of the main mathematical problems in the 18th and 19th centuries involved the following two questions:

How to find the roots of a given polynomial? Can we express those roots by a formula involving only arithmetic operations on the coefficients?

Given a (real or complex) number. Is there any polynomial with integer coefficients of which that number is a root? How to find it?

The answer to the first question is addressed by the Galois Theorem, which characterizes polynomials whose roots can be expressed by formulas. For degree 2 you know such formulas from school. For polynomials of degree 4 there are the so-called Cardano formulas. However, for some polynomials of degree 5 there are provably no such formulas, and in practice the roots can only be found approximately, using numerical methods. In this subsection, we analyze in detail the second question.

Definition. A complex number a is called algebraic if there exists a non-zero polynomial f with integer coefficients such that $f(a) = 0$. Otherwise, a is said to be transcendental.

Examples. Several \well-known" numbers are algebraic.

Rational numbers are algebraic, the rational number $\frac{a}{b}$ is a root of the polynomial $bx - a$.

The n-root of an integer is algebraic, for instance $\sqrt[n]{s}$ is a root of the polynomial $x^n - s$.

Some other irrational numbers are also algebraic, even if it is not obvious at a rst glance. For example, $\sqrt{2} + \sqrt{3}$ is a root of $x^4 - 10x^2 + 1$.

Using the theory of eld extensions, we will prove that the sum, di erence, product, and quotient of algebraic numbers is still an algebraic number (Theorem 6.11).

Example. Apparently, Leonhard Euler was already assuming the fact that not every number is algebraic, but the rst proof of the existence of transcendental numbers was given much later.

Joseph Liouville proved that irrational algebraic numbers cannot be e - ciently approximated by rational numbers, in a certain sense. For that reason, for example, the number $\sum_{i=1}^{1} 10^{-i!}$ cannot be algebraic (that is, a number that has a one in the decimal expansion precisely at positions of the form $i!$, otherwise zeros).

In 1873, Charles Hermite proved that the number e is transcendental, and it was not until 1882 that Ferdinand von Lindemann found a proof of the transcendence of the number . It is however still open for $e +$, e , and several other well-known irrational numbers.

Georg Cantor surprised mathematicians in 1874 when he proved that almost all real numbers are transcendental (in the sense of probability given by a uniform distribution, i.e. a random real number is transcendental with probability 1).

Each of the known proofs of the transcendence of particular numbers is quite complicated. Except for Cantor's argument: he proved in a relatively simple way that there are many transcendental numbers without having to nd any. His argument is based on counting: there are many more transcendental numbers (uncountably many) than algebraic numbers (only countably many). We will now show Cantor's proof, which was one of the main motivations for the development of set theory.

Recall that an in nite set is called countable if its elements can be ordered into a sequence indexed by natural numbers (i.e., it is a set as large as N). We call all other (i.e. larger than N) in nite sets uncountable.

First, note that the union of two countable sets is countable: If $A = \{a_1; a_2; \ldots\}$ and $B = \{b_1; b_2; \ldots\}$, then $A \cup B = \{a_1; b_1; a_2; b_2; \ldots\}$.

So the set Z is countable (the union of 0, all positive and all negative numbers). Even the set Q is countable: order the positive rational numbers by the sum of the numerator and denominator (order those with the same sum arbitrarily), do the same for the negative ones, take the unity and add a zero to the beginning.

Proposition 6.1. The set of algebraic numbers is countable.

Proof. Let's de ne the index of a polynomial $f = a_0 + a_1x + \ldots + a_nx^n \neq 0$ as the sum $|a_0| + |a_1| + \ldots + |a_n| + n$. Note that there are only nitely many integer polynomials of a given index (for instance, for index 1: $f = 1$; index 2: $f = 2$, $f = x$; index 3: $f = 3$, $f = 2x$, $f = x - 1$, $f = x^2$), that is, all polynomials can be arranged in a sequence according to the increasing index. At the same time, every non-zero polynomial has only nitely many roots, i.e. by substituting the polynomial for its roots we obtain a sequence containing all algebraic numbers.

Theorem 6.2. The set of real numbers is uncountable.

Proof. If the set of real numbers was countable, the interval $[0;1)$ would certainly also be countable, and therefore we could arrange the numbers from this interval into a sequence

$$\begin{aligned} a_1 &= 0;a_{11}a_{12}a_{13}\ldots \\ a_2 &= 0;a_{21}a_{22}a_{23}\ldots \\ a_3 &= 0;a_{31}a_{32}a_{33}\ldots \\ &\vdots \end{aligned}$$

Now, let's define the number $b = 0;b_1b_2b_3\ldots$ so that $b_1 \neq a_{11}$, $b_2 \neq a_{22}$, etc. This number cannot be in the list, because it differs in the i-th position from the i-th number, for every $i \in \mathbb{N}$. This is a contradiction with the fact that all numbers from the interval $[0;1)$ should have been there. (For this argument to be correct, decimal expansions ending in nines must be avoided or readapted.)

Every real number is either algebraic or transcendental. There are only a few of the former, so the latter must be uncountable. So, not only must transcendental numbers exist, but there are many more of them than rational ones.

In the following, we will relate the algebraic nature of a given number to the degree of a certain field extension and we will also discuss how to find polynomials whose roots are algebraic numbers.

6.2. **Minimal polynomial and degree of simple extension.**

In this section, we relate the notion of an algebraic number to the properties of so-called simple extensions, i.e. extensions of the form $\mathbf{T}(a)$, determined by one element. The main goal of this section is to develop a theory that ends up with the following characterization: a number a is algebraic if and only if the degree of $[\mathbb{Q}(a) : \mathbb{Q}]$ is finite, moreover, this degree is equal to the degree of any irreducible polynomial, of which a is the root.

We start with a general definition of the algebraic property, for an element of an arbitrary field.

Definition. Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension and let $a \in S$. We say that a is algebraic over $\mathbf{T}$ if there exists a nonzero polynomial in $\mathbf{T}[x]$ of which a is a root. Otherwise, the element a is called transcendental over $\mathbf{T}$.

Note that a number is algebraic in the sense of Section 6.1 if and only if it is an algebraic element over the field $\mathbb{Q}$: the given number is a root of some rational polynomial if and only if it is a root of some integer polynomial, just multiply coefficients.

Definition. Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension and let $a \in S$ be algebraic over $\mathbf{T}$. By a minimal polynomial of a over $\mathbf{T}$ we mean an irreducible monic polynomial $m_{a;\mathbf{T}}$ of $\mathbf{T}[x]$, which has a as a root.

Proposition 6.3 (Properties of the minimal polynomial)**.** Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension and let $a \in S$ be algebraic over $\mathbf{T}$. Then

(1) The minimal polynomial $m_{a;\mathbf{T}}$ exists and it is uniquely determined;
(2) The element a is a root of the polynomial $f \in T[x]$ if and only if $m_{a;\mathbf{T}} \mid f$.

Proof. The set $I = \{f \in T[x] : f(a) = 0\}$ is an ideal of $\mathbf{T}[x]$, and since $\mathbf{T}[x]$ is a principal ideal domain (Theorem 3.2), there exists a monic polynomial $m \in T[x]$ such that $I = mT[x]$. We thus showed that $f(a) = 0$ if and only if $m \mid f$. If the polynomial m were not irreducible in $\mathbf{T}[x]$, i.e. if $m = fg$, where $f;g \notin \{1, m\}$, then $0 = m(a) = f(a)g(a)$, i.e. the element a would be a root of at least one of the

polynomials f, g, but $m \nmid f, g$, contradiction. That is, m is the minimal polynomial of the element a over $\mathbf{T}$. For any other monic irreducible polynomial $\tilde{m} \in T[x]$ of which a is a root, we have $m \mid \tilde{m}$ holds and from irreducibility and monicity we get $\tilde{m} = m$.

Example. It is easy to see that

$$m_{1;\mathbb{Q}} = x - 1; \quad m_{i;\mathbb{Q}} = x^2 + 1; \quad m_{\sqrt[3]{2};\mathbb{Q}} = x^3 - 2;$$

because they are irreducible polynomials that have the given element as a root.

Example. Attention, because for $\zeta_3 = e^{2\pi i/3}$ the minimal polynomial is $m_{\zeta_3;\mathbb{Q}}$ and not $x^3 - 1$, because the latter polynomial is not irreducible. It holds that $x^3 - 1 = (x-1)(x^2 + x + 1)$, the ζ_3 is a root of the second factor, which is irreducible, and thus $m_{\zeta_3;\mathbb{Q}} = x^2 + x + 1$.

Example. Let's compute the minimal polynomial of $a = \sqrt{2} + \sqrt{3}$. We have

$$a^2 = 5 + 2\sqrt{6}; \quad a^3 = 11\sqrt{2} + 9\sqrt{3}; \quad a^4 = 49 + 20\sqrt{6}$$

and we see that $a^4 = 10a^2 - 1$. Therefore, a is a root of the polynomial $x^4 - 10x^2 + 1$. This polynomial is irreducible: thanks to the criterion of the existence of a rational root, we see that it does not have a rational root, and it cannot be decomposed into the product of two polynomials of degree 2, since $\sqrt{2} + \sqrt{3}$ is not a solution of any quadratic equation.

Proposition 6.4 (Structure of simple extensions)**.** Let $\mathbf{T} \subset \mathbf{S}$ be a field extension, and let $a \in S$ be an algebraic element over $\mathbf{T}$. Then

$$\mathbf{T}(a) = \mathbf{T}[a].$$

Proof. From Theorem 5.1 we get

$$T[a] = \{f(a) : f \in T[x]\}.$$

We prove that these elements form a subfield. Let $0 \neq f(a) \in T[a]$, we are looking for its inverse, i.e. a polynomial $g \in T[x]$ such that $f(a)g(a) = 1$. Since $f(a) \neq 0$, the polynomial $m_{a;\mathbf{T}}$ does not divide f. The irreducibility of $m_{a;\mathbf{T}}$ implies $\gcd(m_{a;\mathbf{T}}, f) = 1$ i.e., according to Bézout's identity, there are polynomials $u, g \in T[x]$ such that $1 = um_{a;\mathbf{T}} + gf$. By substituting the element a we get that

$$1 = u(a)m_{a;\mathbf{T}}(a) + g(a)f(a) = u(a) \cdot 0 + g(a)f(a) = f(a)g(a);$$

therefore $g(a)$ is the inverse element of $f(a)$.

Alternative proof. Consider the homomorphism $\varphi : \mathbf{T}[x] \to \mathbf{T}[a]$, $f \mapsto f(a)$. It is clearly surjective and its kernel is the ideal $m_{a;\mathbf{T}}T[x]$. For the First Isomorphism Theorem we conclude that $\mathbf{T}[x]/(m_{a;\mathbf{T}}) \cong \mathbf{T}[a]$. Since $m_{a;\mathbf{T}}$ is irreducible, this ideal is maximal, therefore $\mathbf{T}[a]$ is a field, and then $\mathbf{T}[a] = \mathbf{T}(a)$.

Example. The number $\sqrt{s}$, for $s \in Z$, is algebraic over Q, therefore $\mathbb{Q}(\sqrt{s}) = \mathbb{Q}[\sqrt{s}]$. Indeed,

$$(a + b\sqrt{s})^{-1} = \frac{a}{a^2 - b^2 s} - \frac{b}{a^2 - b^2 s}\sqrt{s} \in \mathbb{Q}[\sqrt{s}].$$

For extensions by more than one element, the formulas are not that neat (try it!) and Proposition 6.4 does not hold in general.

Next recall that in Proposition 5.2 we proved that, if a is transcendental over $\mathbf{T}$, then $\mathbf{T}[a] \neq \mathbf{T}(a)$. Together with Proposition 6.4 this implies:

Observation 6.5. Let $\mathbf{T} \subset \mathbf{S}$ be a field extension and $a \in S$. Then a is algebraic over $\mathbf{T}$ if and only if $\mathbf{T}(a) = \mathbf{T}[a]$.

Proposition 6.6 (Degree of simple extensions)**.** Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension and let $a \in S$ be an algebraic element over $\mathbf{T}$. Then

$$[\mathbf{T}(a) : \mathbf{T}] = \deg m_{a;\mathbf{T}}.$$

Proof. Let us denote by $n = \deg m_{a;\mathbf{T}}$. We prove that the elements $1, a, a^2, \ldots, a^{n-1}$ form the basis of the vector space $\mathbf{T}(a)_{\mathbf{T}}$, and thus that its dimension is n.

If the elements $1, a, a^2, \ldots, a^{n-1}$ were linearly dependent, then $\sum_{i=0}^{n-1} t_i a^i = 0$ would hold for some $t_i \in T$ of which would at least one was non-zero. Thus, the element a would be the root of the (non-zero) polynomial $\sum_{i=0}^{n-1} t_i x^i \in T[x]$ with degree less than that of $m_{a;\mathbf{T}}$, which would be a contradiction with the minimality. We now prove that the elements $1, a, a^2, \ldots, a^{n-1}$ generate the vector space $\mathbf{T}(a)_{\mathbf{T}}$. Consider the element $f(a)$ of the field $\mathbf{T}(a) = \mathbf{T}[a]$. We want to express it as a linear combination. Let $q, r \in T[x]$ be such that $f = q\, m_{a;\mathbf{T}} + r$ and $\deg r < \deg m_{a;\mathbf{T}} = n$. Then

$$f(a) = q(a)\ m_{a;\mathbf{T}}(a) + r(a) = q(a)\ 0 + r(a) = r(a),$$

and since the degree of r is less than n, we have $f(a) = r(a) = \sum_{i=0}^{n-1} t_i a^i$, where $t_i \in T$ are the coefficients of the r polynomial.

Example. Using Proposition 6.6, one can determine the degree of a simple extension.

$[C : R] = [R(i) : R] = \deg m_{i;R} = \deg(x^2 + 1) = 2$.
$[Q(\sqrt[n]{p}) : Q] = \deg(x^n - p) = n$ for any $n \in N$ and a prime number p, since the given polynomial is irreducible according to Eisenstein's criterion. (If p is not prime, then the situation is more complicated.)
$[Q(e^{2\pi i/n}) : Q] = \varphi(n)$ (Euler's totient function), but this is not easy to prove, the theory of cyclotomic polynomials is used for this. If n is prime, the minimal polynomial is $x^{n-1} + x^{n-2} + \ldots + 1 = \frac{x^n - 1}{x - 1}$, whose irreducibility can be shown from Eisenstein's criterion after substitution.

Corollary 6.7. Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension, and let $a \in S$. The element a is algebraic over $\mathbf{T}$ if and only if the degree $[\mathbf{T}(a) : \mathbf{T}]$ is finite).

Proof. If a was transcendental, then $1, a, a^2, \ldots$ would form an infinite linearly independent set: if $\sum_{i=0}^{n} t_i a^i = 0$ for some coefficients $t_i \in T$, at least one nonzero, would be a the root of the nonzero polynomial $\sum_{i=0}^{n} t_i x^i$ of $\mathbf{T}[x]$, contratiction. The opposite implication follows from Proposition 6.6.

Example. We will show the structure of so-called quadratic extensions, i.e. extensions of degree 2. We prove that if $\mathbf{T} < \mathbf{S} \subseteq C$ and $[\mathbf{S} : \mathbf{T}] = 2$, then

$$\mathbf{S} = \mathbf{T}(\sqrt{s}) \text{ for some } s \in T.$$

Let $1, a$ be a basis of the space $\mathbf{S}_{\mathbf{T}}$. Then $\mathbf{S} = \mathbf{T}(a)$ and by Proposition 6.6 a is a root of some polynomial in $\mathbf{T}[x]$ of degree 2. A well-known formula for computing the roots of a quadratic polynomial says that $a = u + v\sqrt{s}$ for some $u, v, s \in T$, and thus $\mathbf{S} = \mathbf{T}(u + v\sqrt{s}) = \mathbf{T}(\sqrt{s})$.

Exercises.

Recall the notation $\varepsilon_n = e^{2\pi i/n}$.

1. Compute the minimal polynomial $m_{a;Q}$, where a is $1 + \sqrt{5}$, $1 - \sqrt{2}$, $\frac{3+\sqrt{2}}{\sqrt{2}-1}$, $i + \sqrt{5}$, $\varepsilon_7 + \varepsilon_7^{-1}$.

2. Compute the dimension and find a basis of the vector space $Q(\varepsilon_n)_Q$ for $n = 5, 6, 8$.

3. Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension and $a \in S$ an algebraic element. Describe the polynomial $m_{a^{-1};\mathbf{T}}$ using the coefficients of the polynomial $m_{a;\mathbf{T}}$.

4. Let $\mathbf{T} \subset \mathbf{S}$ be a field extension and $a \in S$ an algebraic element. Prove that if $[\mathbf{T}(a) : \mathbf{T}]$ is odd, then $\mathbf{T}(a) = \mathbf{T}(a^2)$.

5. The splitting field of a polynomial is the smallest field extension, in which the polynomial decomposes into linear factors.

Compute the degree of the splitting field of $x^5 - 3x + 3$ over Q (Hint: show irreducibility).

6. Compute the degree of the splitting field of $x^4 + x^3 + 2x^2 + x + 1$ over Q.

6.3. Extensions by multiple elements.

The following general rule is used to calculate the degree of extensions by more than one element.

Proposition 6.8 (degree of non-simple extensions)**.** Let $\mathbf{T} \subset \mathbf{S} \subset \mathbf{U}$ be field extensions. Then

$$[\mathbf{U} : \mathbf{T}] = [\mathbf{U} : \mathbf{S}] \cdot [\mathbf{S} : \mathbf{T}].$$

Proof. Let us choose a basis A of the vector space $\mathbf{S}_{\mathbf{T}}$ and a basis B of the vector space $\mathbf{U}_{\mathbf{S}}$. We are then going to prove that

$$C = \{ab : a \in A, b \in B\}$$

is a basis of the vector space $\mathbf{U}_{\mathbf{T}}$.

First, we are going to prove that C is a generating set of $\mathbf{U}_{\mathbf{T}}$ (clearly $C \subset U$, and thus C generates a subspace of $\mathbf{U}_{\mathbf{T}}$). If $u \in U$, then $u = \sum_j s_j b_j$ for some $s_j \in S$ and $b_j \in B$. Each s_j can be written as $s_j = \sum_i t_{ij} a_i$ for some $t_{ij} \in T$ and $a_i \in A$, and by substituting the second equality into the first we get

$$u = \sum_j \Big(\sum_i t_{ij} a_i\Big) b_j = \sum_{i,j} t_{ij} \cdot a_i b_j.$$

Thus, u is a linear combination of elements of C with coefficients from the field $\mathbf{T}$.

Next, we prove linear independence. Assume that $\sum_{i,j} t_{ij} \cdot a_i b_j = 0$ for some $t_{ij} \in T$ and $a_i b_j \in C$. We break the sum down into

$$0 = \sum_{i,j} t_{ij} a_i b_j = \sum_j \underbrace{\Big(\sum_i t_{ij} a_i\Big)}_{\in S} b_j.$$

The linear independence of elements b_j over the field $\mathbf{S}$ gives us $\sum_i t_{ij} a_i = 0$ for each j and from the linear independence of elements a_i over the field $\mathbf{T}$ we get $t_{ij} = 0$ for all i, j.

It follows that C is a basis of $\mathbf{U}_{\mathbf{T}}$, and thus

$$[\mathbf{U} : \mathbf{T}] = |C| = |A| \cdot |B| = |A| \cdot |B| = [\mathbf{S} : \mathbf{T}] \cdot [\mathbf{U} : \mathbf{S}].$$

Proposition 6.6 and 6.8 can be applied to compute the degree of expansions of the type $\mathbf{T}(a_1, a_2, \ldots)$: We can split the double extension $\mathbf{T} \subset \mathbf{T}(a, b)$ into two simple extensions $\mathbf{T} \subset \mathbf{T}(a) \subset \mathbf{T}(a, b)$ and compute

$$[\mathbf{T}(a, b) : \mathbf{T}] = [\mathbf{T}(a, b) : \mathbf{T}(a)] \cdot [\mathbf{T}(a) : \mathbf{T}] = \deg m_{b;\mathbf{T}(a)} \cdot \deg m_{a;\mathbf{T}} \le \deg m_{b;\mathbf{T}} \cdot \deg m_{a;\mathbf{T}}.$$

But attention! When expressing the degree $[\mathbf{T}(a, b) : \mathbf{T}(a)]$ we have to use the minimal polynomial of the element b over the field $\mathbf{T}(a)$, which may be of smaller degree than the minimal polynomial over by the body $\mathbf{T}$. By repeatedly applying the described procedure, we can easily prove the following corollary:

Corollary 6.9. Let $\mathbf{T} \subset \mathbf{S}$ be an extension of fields and let $a_1, \ldots, a_n \in S$ be algebraic elements over $\mathbf{T}$. Then $\mathbf{T}(a_1, \ldots, a_n)$ is a finite degree extension over $\mathbf{T}$.

Example. Let us show

$$Q(\sqrt{2}+\sqrt{3}) = Q(\sqrt{2},\sqrt{3}).$$

using the above statements. Clearly $Q(\sqrt{2}+\sqrt{3}) \subseteq Q(\sqrt{2},\sqrt{3})$. So if we can prove that both spaces have the same dimension, they must be identical. Let's calculate the minimal polynomials:

$$m_{\sqrt{2}+\sqrt{3},Q} = x^4 - 10x^2 + 1,$$
$$m_{\sqrt{2},Q} = x^2 - 2,$$
$$m_{\sqrt{3},Q(\sqrt{2})} = x^2 - 3 \text{ (note that this is irreducible in } Q(\sqrt{2})[x] \text{ !).}$$

By Proposition 6.6 and 6.8 we get $[Q(\sqrt{2}+\sqrt{3}) : Q] = 4$ and $[Q(\sqrt{2},\sqrt{3}) : Q] = [Q(\sqrt{2},\sqrt{3}) : Q(\sqrt{2})]\ [Q(\sqrt{2}) : Q] = 2 \cdot 2 = 4$.

If $\mathbf{T} \subseteq \mathbf{S}$ is a field extension and every element of the field $\mathbf{S}$ is algebraic over $\mathbf{T}$, we speak of an algebraic extension. All finite degree extensions have this property.

Proposition 6.10. Finite degree extensions are algebraic.

Proof. Let us denote $n = [\mathbf{S} : \mathbf{T}]$. For any element $a \in S$, we can prove that it is algebraic over $\mathbf{T}$. The elements $1, a, a^2, \ldots, a^{n-1}, a^n$ are linearly dependent because there are more of them than the dimension of the vector space $\mathbf{S}_{\mathbf{T}}$. Thus there is a non-trivial linear combination of them equal to 0, such that $\sum_{i=0}^{n} t_i a^i = 0$. That is, the element a is the root of the non-zero polynomial $\sum_{i=0}^{n} t_i x^i \in T[x]$.

We can use Proposition 6.10 to prove algebraicity in a non-constructive way: to prove that an element a is algebraic over $\mathbf{T}$, it suffices to find an extension $\mathbf{S} \supseteq \mathbf{T}$ of finite degree that contains a. A typical example is the proof that the sum, difference, product, and quotient of two algebraic elements are algebraic element:

Theorem 6.11 (algebraic elements form a subfield)**.** Let $\mathbf{T} \subseteq \mathbf{S}$ be a field extension. The elements of $\mathbf{S}$, which are algebraic over $\mathbf{T}$ form a subfield of $\mathbf{S}$.

Proof. Let us consider elements $a, b \in S$ which are algebraic over $\mathbf{T}$. The extension $\mathbf{T} \subseteq \mathbf{T}(a,b)$ is of finite degree (Corollary 6.9), and therefore algebraic (Proposition 6.10). That is, all the elements of $\mathbf{T}(a,b)$ are algebraic over $\mathbf{T}$, especially the elements $a+b$, $a-b$, $-a$ and a^{-1} (for $a \neq 0$). Thus, the algebraic elements form a subfield of the field $\mathbf{S}$.

Exercises.

1. Compute the degree of the extension $[Q(\sqrt{3},\sqrt[3]{3}) : Q]$.

2. Show that $[Q(\sqrt{p_1},\ldots,\sqrt{p_n}) : Q] = 2^n$, where $p_1,\ldots,p_n$ are pairwise different primes.

3. Let $\mathbf{T} \subseteq \mathbf{S} \subseteq \mathbf{U}$ be a field extension, $\mathbf{U}$ algebraic over $\mathbf{S}$, and $\mathbf{S}$ algebraic over $\mathbf{T}$. Is $\mathbf{U}$ necessarily algebraic over $\mathbf{T}$? If so, prove it. If not, provide a counterexample.

4. Let $\mathbf{S}$ be the splitting field of the polynomial $f \in T[x]$ of degree n. Prove that $[\mathbf{S} : \mathbf{T}]$ divides $n!$.

5. Complete the details of the proof that every extension $\mathbf{T} \subseteq \mathbf{S}$ of finite degree can be written as $\mathbf{S} = \mathbf{T}(a_1,\ldots,a_n)$, where $a_1,\ldots,a_n$ are some algebraic elements over $\mathbf{T}$.

7. Problems that cannot be solved with a ruler and a compass

In ancient Greece, mathematics mainly meant to do geometry. Thus, several classical math problems ask to construct some geometrical object by using only a ruler and compass. Some of these problems are easy enough to be taught in primary schools, such as doubling a square or bisecting an angle. There are however problems that have been unresolved for millennia: for example, the construction of a regular heptadecagon (17-gon) was only discovered by Gauss in 1796. Already

at that time, it was suspected that some problems could not be solved, but it was only the development of algebra at the beginning of the 19th century that made it possible to prove it. Among the most famous such tasks were:

doubling the cube: construct a cube that has double the volume of a given cube. This is equivalent to, for a given line segment, construct one that is $\sqrt[3]{2}$ times longer

angle trisection: construct the third of a given angle;

squaring the circle: for a given line segment, construct one that is times longer (original formulation: for a given circle k, construct a segment such that its square has the same are as the circle k; or construct a line segment that is the same length as the circumference of k; both problems can easily be converted to the construction of a line segment that is times longer).

construction of a regular n-gon, for arbitrary n.

In this section, we will show the algebraic proof method invented by Pierre Wantzel in 1837. It can be used to prove the unsolvability of all the problems mentioned (in some cases with the help of another theory, such as the proof of the transcendence of the number).

Before we start, we need to clarify what we actually mean by a ruler-and-compass construction. Initially, a nite set of M_0 points in the plane is given. From this set we can construct a new point as an intersection of straight lines or circles determined by already constructed points; this procedure can be repeated several times.

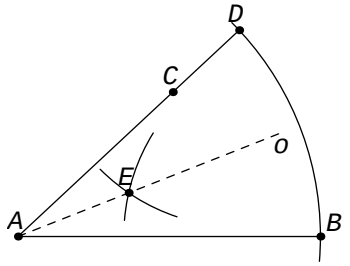
Formally a construction by ruler and compass is a sequence $M_0 \quad M_1 \quad \dots \quad M_n$ of sets, such that $M_{i+1} = M_i \cup \{X\}$, where the point X can be obtained by

(1) intersecting lines AB and CD;
(2) intersecting a line AB and a circle $k(C; |DE|)$ with center C and radius $|DE|$;
(3) intersecting circles $k(A; |BC|)$ and $k(D; |EF|)$

for some points $A, B, C, D, E, F \in M_i$.

The principle of Wantzel's method is the translation of ruler-and-compass constructions into the language of algebra: instead of sets of points, we will consider coordinates. So let us choose coordinates in the plane and consider the smallest eld $\mathbf{T}_i \quad \mathbb{R}$, which contains the x and y coordinates of all points from M_i. That is, if M_i contains points $A_1, \dots, A_k$ with coordinates $(a_1, b_1), \dots, (a_k, b_k)$, then $\mathbf{T}_i = \mathbb{Q}(a_1, b_1, \dots, a_k, b_k)$. Adding the point X with coordinates (u, v) gives $\mathbf{T}_{i+1} = \mathbf{T}_i(u, v)$. The result is a series of eld extensions $\mathbf{T}_0 \quad \mathbf{T}_1 \quad \mathbf{T}_2 \quad \dots \quad \mathbf{T}_n$.

Example (Bisecting an angle)**.** Let's see us formalize the task of bisecting an angle. For this let us have an angle given by three points A, B, C (where A is the vertex of the angle).



We construct the points

$$D = k(A; |AB|) \cap AC \quad \text{and} \quad E = k(B; |BD|) \cap k(D; |BD|);$$

and the resulting angle will be given by points $A; B; E$. So

$$M_0 = fA; B; Cg; \quad M_1 = M_0 [fDg; \quad M_2 = M_1 [fEg:$$

Let's choose coordinates such that $A = (0; 0)$, $B = (1; 0)$ and $C = (a; b)$. It is not di cult to check that $D = (\frac{a}{a^2+b^2}; \frac{b}{a^2+b^2})$ and $E = (\frac{1}{2} + \frac{a}{2}\sqrt{\frac{b^2}{a^2+b^2}}; \frac{\sqrt{3}}{2} + \frac{b+a\sqrt{3}}{2}\sqrt{\frac{b^2}{a^2+b^2}})$, thus

$$\mathbf{T}_0 = \mathrm{Q}(a; b); \quad \mathbf{T}_1 = \mathbf{T}_0(\sqrt{a^2 + b^2}); \quad \mathbf{T}_2 = \mathbf{T}_0(\sqrt{a^2 + b^2}; \sqrt{3}):$$

The key part in Wantzel's method is the following lemma:

Lemma 7.1. For all constructions with ruler and circle we have $[\mathbf{T}_{i+1} : \mathbf{T}_i] \; 2 \; f1; 2g$, for all i.

Proof. We will check that the lemma holds for all expansions corresponding to the three di erent types of constructions:

(1) If we intersect two straight lines, we obtain the coordinates of the new point by solving a system of two linear equations with two variables over the eld $\mathbf{T}_i$. More precisely, the line de ned by points $A; B$ with coordinates $(a; b)$, $(c; d)$, where $a; b; c; d \; 2 \; T_i$, has the equation

$$(b \quad d)x + (c \quad a)y = bc \quad ad$$

and we see that all three coe cients are in the eld $\mathbf{T}_i$. The solution of the system of linear equations of two variables over the eld $\mathbf{T}_i$ is the pair $(u; v)$ of elements of the eld $\mathbf{T}_i$, so that $\mathbf{T}_{i+1} = \mathbf{T}_i(u; v) = \mathbf{T}_i$ and

$$[\mathbf{T}_{i+1} : \mathbf{T}_i] = 1:$$

(2) If we intersect a straight line and a circle, we obtain the coordinates of the new point by solving a system of one linear and one quadratic equation with two variables over the eld $\mathbf{T}_i$. We discussed straight lines above; a circle $k(A; jBCj)$ determined by the points $A = (a; b)$, $B = (c; d)$, $C = (e; f)$, with $a; b; c; d; e; f \; 2 \; T_i$, satis es the equation

$$(x \quad a)^2 + (y \quad b)^2 = (c \quad e)^2 + (d \quad f)^2:$$

Note that all coe cients are from $\mathbf{T}_i$. If we express y from the linear equation and substitute it into the quadratic, we get a quadratic equation for x, whose coe cients are from $\mathbf{T}_i$ and the solution is $x' = u + v\sqrt{s}$ for some $u; v; s \; 2 \; T_i$. Substituting into the linear equation, we also obtain that $y' = u' + v'\sqrt{s}$ for some $u'; v' \; 2 \; T_i$. That is, $\mathbf{T}_{i+1} = \mathbf{T}_i(x'; y') = \mathbf{T}_i(\sqrt{s})$, from which it follows that

$$[\mathbf{T}_{i+1} : \mathbf{T}_i] \; 2 \; f1; 2g$$

depending on whether $\sqrt{s} \; 2 \; T_i$ or not. (Exercise: Perform the described calculation in detail and verify that indeed both solutions belong to $\mathbf{T}_i(\sqrt{s})$!)

(3) If we take the intersection of two circles, we obtain the coordinates of the new point by solving a system of two quadratic equations with two variables over the eld $\mathbf{T}_i$. By subtracting the equations from each other, we get rid of squares of variables (they all have a coe cient of 1) and thus obtain an equivalent system consisting of one linear and one quadratic equation, all over the eld $\mathbf{T}_i$. By the same argument as in (2), we get

$$[\mathbf{T}_{i+1} : \mathbf{T}_i] \; 2 \; f1; 2g:$$

(Perform the described calculation in detail yourself!)

Proposition 7.2 (degree of a eld extension for ruler and compass constructions)**.** For every eld extension obtained by a construction with ruler and compass, we have $[\mathbf{T}_n : \mathbf{T}_0] = 2^k$ for some $k \quad n$.

Proof. By Theorem 6.8 we have

$$[\mathbf{T}_n : \mathbf{T}_0] = [\mathbf{T}_n : \mathbf{T}_{n-1}] \cdot \ldots \cdot [\mathbf{T}_2 : \mathbf{T}_1] \cdot [\mathbf{T}_1 : \mathbf{T}_0];$$

which is a power of 2, by Lemma 7.1.

Example (doubling the cube)**.** Starting with a line segment of length 1, the goal is to construct a line segment of length $\sqrt[3]{2}$. By picking e.g. the line segment between $(0;0)$ and $(0;1)$, we can assume that $\mathbf{T}_0 = \mathbb{Q}$. Also, without loss of generality, we can assume that the resulting line segment has endpoints $(0;0)$ and $(\sqrt[3]{2};0)$. So, if there is a ruler-and-compass construction $\mathbf{T}_0;\mathbf{T}_1;\ldots;\mathbf{T}_n$, then $\sqrt[3]{2}$ must belong to the field $\mathbf{T}_n$, i.e. $\mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2}) \subset \mathbf{T}_n$. According to Proposition 6.8 we get

$$[\mathbf{T}_n : \mathbf{T}_0] = [\mathbf{T}_n : \mathbb{Q}(\sqrt[3]{2})] \cdot [\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3 \cdot [\mathbf{T}_n : \mathbb{Q}(\sqrt[3]{2})];$$

which contradicts Proposition 7.2.

(More generally, we can prove that no line segment of length a, whose minimal polynomial $m_{a;\mathbb{Q}}$ has a degree that is not a power of two, can be constructed from the unit segment.)

Example (squaring the circle)**.** Let's choose the coordinates so that the endpoints of the specified segment (indicating the center and radius of the circle) are $(0;0)$ and $(1;0)$; so $\mathbf{T}_0 = \mathbb{Q}$. The goal now is to construct a segment of length $\sqrt{\pi}$ (or $2\sqrt{\pi}$ and π in the original assignment). If that's true, the transcendental number π would be an element of some extension field $\mathbf{T}_n$, obtained by ruler-and-compass construction. But according to Proposition 7.2 $\mathbf{T}_n$ is an extension of finite degree, and therefore according to Theorem 6.10 contains only algebraic numbers, which contradicts to π being transcendental.

(More generally, we can prove that no line segment of transcendental length cannot be constructed from the unit line segment.)

Example (trisection of an angle)**.** To show that we cannot trisect angles by ruler and compass, it is enough to find one specific angle, for which it is impossible. So let us consider the angle 60° given by the points $(0;0)$, $(1;0)$ and $(\frac{1}{2};\frac{\sqrt{3}}{2}) = (\cos 60^\circ;\sin 60^\circ)$; so $\mathbf{T}_0 = \mathbb{Q}(\sqrt{3})$. We prove that it is not possible to construct the point

$$(\cos 20^\circ;\sin 20^\circ):$$

(This is enough, since if we have a line of angle 20°, we can get this one by intersection with the unit circle.) If we can prove that

$$[\mathbb{Q}(\sqrt{3};\cos 20^\circ) : \mathbb{Q}(\sqrt{3})] = 3;$$

then the rest of the proof works as for doubling the cube. So, according to Theorem 6.6 it suffices to find the minimal polynomial of the number $\cos 20^\circ$ over the field $\mathbb{Q}(\sqrt{3})$, i.e. some irreducible polynomial whose number is $\cos 20^\circ$ by the root. If we go through any collection of trigonometric formulas, we find the relationship

$$\cos 3\alpha = 4(\cos\alpha)^3 - 3\cos\alpha;$$

from which it follows that $\cos 20^\circ$ is a root of the polynomial $4x^3 - 3x - \frac{1}{2} \in \mathbb{Q}[x]$. This polynomial is irreducible in $\mathbb{Q}(\sqrt{3})[x]$ because it has no root in $\mathbb{Q}(\sqrt{3})$ (as we can easily find out by substituting $x = a + b\sqrt{3}$). So

$$m_{\cos 20^\circ;\mathbb{Q}(\sqrt{3})} = x^3 - \frac{3}{4}x - \frac{1}{8}$$

and we get $[\mathbb{Q}(\sqrt{3};\cos 20^\circ) : \mathbb{Q}(\sqrt{3})] = \deg m_{\cos 20^\circ;\mathbb{Q}(\sqrt{3})} = 3$.

Exercises.

1. Complete the details in the proof of the Lemma 7.1. Specifically, prove that

if the points $A; B$ have coordinates in the field $\mathbf{T}$, then the line AB is given by the equation with coefficients in $\mathbf{T}$,
if the points $A; B; C$ have coordinates in the field $\mathbf{T}$, then the circle $k(A; jBCj)$ is given by equation with coefficients in $\mathbf{T}$,
a system of one linear and one quadratic equation over the field $\mathbf{T}$ has at most two solutions, which are all in the field $T(\sqrt{s})$, for some $s \in T$.

We call a real number a constructible if a line segment of length 1 can be constructed into a line segment of length a.

2. Prove that the constructible numbers form a subfield $\mathbf{K}$ of the field R such that $\sqrt{a} \in K$ for every $a \in K$.

3. Prove that a regular n-gon can be constructed with a ruler and compass if and only if the constructible number is $\cos(2\pi/n)$.

4. Prove that one cannot construct a regular k-gon for any k divisible by nine.

5. Let p be prime. Prove that if a regular p-gon can be constructed with a ruler and compass, then $p - 1$ is a power of two.

6. Let p be prime. Prove that if a regular p-gon can be constructed with a ruler and compass, then $p = 2^{2^k} + 1$ for some k. (Hint: according to the previous exercise, $p = 2^m + 1$; if odd n divides m, then $2^{m/n} + 1$ divides p.)

7. Prove that if a regular n-gon can be constructed with a ruler and a compass, then a regular $2n$-gon can also be constructed.

8. Which regular n-gons for $n < 17$ can be constructed with a ruler and compass?

8. Isomorphisms of rupture fields and splitting fields

Definition. Let $\mathbf{T}$ be a field and let f be a polynomial of $\mathbf{T}[x]$ of degree ≥ 1.

A rupture field $\mathbf{S}$ of f is a minimal extension field of $\mathbf{T}$ in which f has a root. In other words, it is an extension field $\mathbf{S}$ where there is an $a \in S$ such that $\mathbf{S} = \mathbf{T}(a)$ and $f(a) = 0$.
The splitting field of f over $\mathbf{T}$ is the minimal extension field in which f decomposes into linear factors. In other words, it is an extension field $\mathbf{S}$ where there exist $a_1; \ldots; a_n \in S$ such that $\mathbf{S} = \mathbf{T}(a_1; \ldots; a_n)$ and $f = k(x - a_1) \ldots (x - a_n)$.

In the Algebra 1, we already proved that splitting fields exist for every field $\mathbf{T}$ and every polynomial f (one root is constructed in the factor ring $\mathbf{T}[\ ]/(f(\ ))$, then we proceed recursively with the polynomial $f/(x - \ )$). In this section, we prove the uniqueness of splitting fields, except for isomorphism.

Example. By the fundamental theorem of algebra, we know that the rupture and splitting fields of a polynomial f over the field Q are subfields of C: we can obtain any rupture field as $Q(a)$, where a is a complex root of f, and the splitting field is $Q(a_1; \ldots; a_m)$, where $a_1; \ldots; a_m$ are all complex roots of f.

Consider the polynomial $x^2 + 1$. The only rupture field of it, within C, is the field $Q(i) = Q(-i)$, which contains both roots $\pm i$, and is therefore also the splitting field.

Let us denote by $\varepsilon = e^{2\pi i/3}$.

Consider the polynomial $x^3 - 1$. This polynomial has two different rupture fields in C, namely $Q = Q(1)$ and $Q(\varepsilon) = Q(\varepsilon^2)$. These fields are clearly not isomorphic. The larger one is a splitting field, as it contains all three roots.
Consider the polynomial $x^3 - 2$. This polynomial has two different rupture fields, $Q(\sqrt[3]{2})$ and $Q(\varepsilon\sqrt[3]{2})$ (the latter contains both imaginary roots). Although it is not clear at first glance, these fields are isomorphic. The splitting field is the field $Q(\sqrt[3]{2}; \varepsilon\sqrt[3]{2}) = Q(\sqrt[3]{2}; \varepsilon)$.

As we saw in the examples, the rupture fields of a reducible polynomial are in general not isomorphic (e.g. the roots of the irreducible factors $x - 1$ and $x^2 + x + 1$ of $x^3 - 1$ lead to different extensions). But, perhaps somewhat surprisingly, for irreducible polynomials all rupture fields are isomorphic. We are also going to prove isomorphism for splitting field, this time without the assumption of irreducibility. Before we start with the proof, we need the following helpful definition.

Definition. Let $\mathbf{T} \subseteq \mathbf{S}, \mathbf{U}$ be fields. By a $\mathbf{T}$-isomorphism $\mathbf{S} \to \mathbf{U}$ we mean an isomorphism φ for which $\varphi(t) = t$ holds for every $t \in \mathbf{T}$.

Theorem 8.1 (Uniqueness of rupture fields and splitting fields)**.** Let $\mathbf{T}$ be a field and $f \in \mathbf{T}[x]$ be of degree ≥ 1.

(1) If f is irreducible, then every two rupture fields of f over $\mathbf{T}$ are $\mathbf{T}$-isomorphic.
(2) Every two splitting fields of f over $\mathbf{T}$ are $\mathbf{T}$-isomorphic.

Theorem 8.1 is a special case of the slightly more general Lemmas 8.2 and 8.3 on the extension of partial isomorphisms, which we are going to prove below. The following remarks and observations are needed to formulate them.

Let $\mathbf{T} \subseteq \mathbf{T}_1$, $\mathbf{T} \subseteq \mathbf{T}_2$ be extensions of fields and $\varphi : \mathbf{T}_1 \to \mathbf{T}_2$ be a $\mathbf{T}$-isomorphism. The map φ can be extended to the $\mathbf{T}$-isomorphism of domains of polynomials over these fields (we will again denote it by φ):

$$\varphi : \mathbf{T}_1[x] \to \mathbf{T}_2[x]; \qquad \sum a_i x^i \mapsto \sum \varphi(a_i) x^i.$$

Let us denote $f = \sum a_i x^i$, $g = \sum b_i x^i$. The coefficients of the sum $f + g$ are $a_i + b_i$, the coefficients of the sum $\varphi(f) + \varphi(g)$ are $\varphi(a_i) + \varphi(b_i) = \varphi(a_i + b_i)$ and we see that $\varphi(f + g) = \varphi(f) + \varphi(g)$. The coefficients of the product fg are $\sum_{i+j=k} a_i b_j$, the coefficients of the product $\varphi(f)\varphi(g)$ are $\sum_{i+j=k} \varphi(a_i)\varphi(b_j) = \varphi(\sum_{i+j=k} a_i b_j)$ and we see that $\varphi(fg) = \varphi(f)\varphi(g)$. The bijectivity of the map is obvious. An immediate consequence of the product property is that

> $f \mid g$ in $\mathbf{T}_1[x]$ if and only if $\varphi(f) \mid \varphi(g)$ in $\mathbf{T}_2[x]$;
> The polynomial f is irreducible in $\mathbf{T}_1[x]$ if and only if $\varphi(f)$ is irreducible in $\mathbf{T}_2[x]$.

Lemma 8.2. Let $\mathbf{T} \subseteq \mathbf{T}_1$, $\mathbf{T} \subseteq \mathbf{T}_2$ be extensions of fields and $\varphi : \mathbf{T}_1 \to \mathbf{T}_2$ be a $\mathbf{T}$-isomorphism. Let $f \in T_1[x]$ be an irreducible polynomial, $\mathbf{T}_1(a)$ be a rupture field of f over $\mathbf{T}_1$ and $\mathbf{T}_2(b)$ be a rupture field of $\varphi(f)$ over $\mathbf{T}_2$. Then there exists a $\mathbf{T}$-isomorphism $\psi : \mathbf{T}_1(a) \to \mathbf{T}_2(b)$ such that $\psi(a) = b$ and $\psi|_{T_1} = \varphi$.

Proof. By Theorem 6.4, $T_1(a) = T_1[a] = \{g(a) : g \in T_1[x]\}$ and $T_2(b) = T_2[b] = \{g(b) : g \in T_2[x]\}$. So consider the mapping

$$\psi : T_1(a) \to T_2(b); \qquad g(a) \mapsto \varphi(g)(b).$$

First, it is necessary to prove that this is well-defined. Note that $f = m_{a;\mathbf{T}_1}$ since f is an irreducible polynomial and a is its root, and likewise $\varphi(f) = m_{b;\mathbf{T}_2}$, since $\varphi(f)$ is an irreducible polynomial and b is its root. We have

$$g(a) = h(a) \quad , \quad (g - h)(a) = 0 \quad , \quad f \mid g - h$$

and analogously

$$\varphi(g)(b) = \varphi(h)(b) \quad , \quad \varphi(g - h)(b) = 0 \quad , \quad \varphi(f) \mid \varphi(g - h).$$

The equivalence of the two statements on the right-hand side follows from the observation above. We have proved that ψ is a well-defined mapping and, moreover, injective. Obviously, this is a bijection and it is easy to verify that it is a ring homomorphism: for every $g, h \in T_1[x]$ we have $\psi(g(a) + h(a)) = \psi((g + h)(a)) = \varphi(g + h)(b) = \varphi(g)(b) + \varphi(h)(b) = \psi(g(a)) + \psi(h(a))$ and analogously for multiplication. The elements of the field $\mathbf{T}_1$ correspond to the choice of a constant polynomial

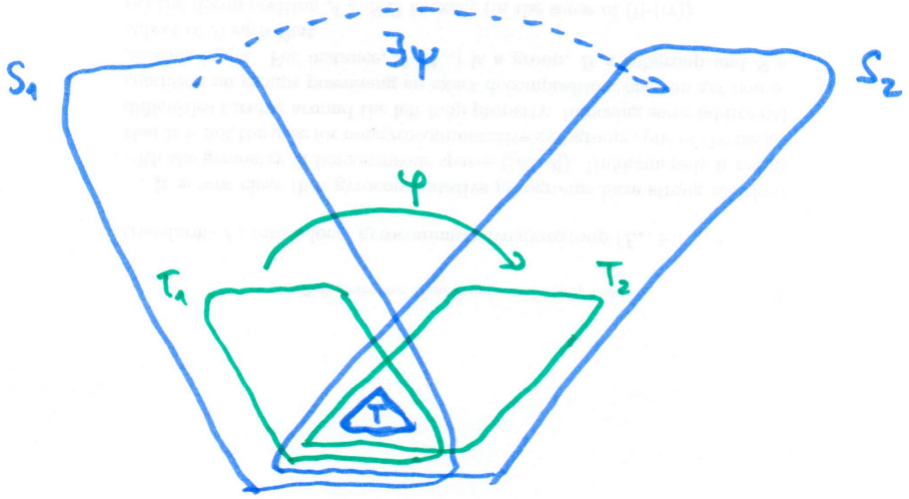


Figure 7. Illustration of the proof of the uniqueness of a splitting field.

c, for such a polynomial $(c) = (c(a)) = '(c)(b) = '(c)$ holds, so $j_{\mathbf{T}_1} = '$. By choosing $g = x$ we verify that $(a) = b$.

Lemma 8.3. Let $\mathbf{T} \quad \mathbf{T}_1$, $\mathbf{T} \quad \mathbf{T}_2$ be extensions of fields and $' : \mathbf{T}_1 \to \mathbf{T}_2$ be a $\mathbf{T}$-isomorphism. Let $f \in T_1[x]$ be a polynomial of degree ≥ 1 and denote by $\mathbf{S}_1$ the splitting field of the polynomial f over $\mathbf{T}_1$ and $\mathbf{S}_2$ the splitting field of the polynomial $'(f)$ over $\mathbf{T}_2$. Then there exists a $\mathbf{T}$-isomorphism $: \mathbf{S}_1 \to \mathbf{S}_2$ such that $j_{T_1} = '$.

Proof. We proceed by induction on the degree of the polynomial f. If $\deg f = 1$, then $\mathbf{S}_1 = \mathbf{T}_1$, $\mathbf{S}_2 = \mathbf{T}_2$, and $= '$. In the induction step, consider the irreducible divisor g of the polynomial f and its root a in $\mathbf{S}_1$. Then $'(g)$ is an irreducible divisor of the polynomial $'(f)$ and consider its root b in $\mathbf{S}_2$. By Lemma 8.2 there is a homomorphism $: \mathbf{T}_1(a) \to \mathbf{T}_2(b)$ such that $(a) = b$ and $j_{T_1} = '$. Let's write $f = (x \quad a)\ h$ for some $h \in T_1[x]$, i.e. also $(f) = (x \quad b) \quad (h)$. Then $\mathbf{S}_1$ is the splitting field of the polynomial h over $\mathbf{T}_1(a)$ and $\mathbf{S}_2$ is the splitting field of the polynomial (h) over $\mathbf{T}_2(b)$. Since $\deg h < \deg f$, by the induction assumption there exists a $\mathbf{T}$-isomorphism $: \mathbf{S}_1 \to \mathbf{S}_2$ such that $j_{T_1(a)} = $, i.e. also $j_{T_1} = '$.

By choosing $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{T}$ and $' = id$ in both lemmas, we get Theorem 8.1.

9. Classification of finite fields

9.1. Frobenius endomorphism.

We begin with an example of a homomorphism that plays a crucial role in rings with characteristic different from zero.

Theorem 9.1 (Frobenius endomorphism)**.** Let $\mathbf{R}$ be a commutative ring with unit and of prime characteristic p. Define the mapping

$$'_p : \mathbf{R} \to \mathbf{R}; \qquad a \mapsto a^p.$$

(1) The mapping $'_p$ is a homomorphism.
(2) If $\mathbf{R}$ is a field, then $'_p$ is injective.
(3) If $\mathbf{R}$ is a finite field, then $'_p$ is an automorphism.

The map φ_p is called Frobenius endomorphism, or Frobenian automorphism in the case when it is bijective. A special case of point (1) is equality

$$(a+b)^p = a^p + b^p;$$

which is the dream of every high school student, but is valid only under the assumption of a prime characteristic p.

Proof. (1) Clearly $(a\ b)^p = a^p\ b^p$ and, according to the binomial theorem,

$$(a+b)^p = \sum_{i=0}^{p} \binom{p}{i} a^i b^{p\ i} = a^p + b^p;$$

for p divides all the binomial coefficients $\binom{p}{i}$, $i = 1;\dots;p\ \ 1$, since all the prime numbers contained in the denominator are smaller.

(2) The condition $\varphi_p(a) = a^p = 0$ implies that $a = 0$. Thus, the kernel of the homomorphism φ_p is trivial, i.e. it is injective (Theorem 4.2).

(3) An injective mapping on a finite set is bijective.

9.2. **Derivatives and multiple roots.**

Let $\mathbf{R}$ be an integral domain. Let $a \in R$ and $f \in R[x]$. Recall that an element a is a root of f if and only if $x\ \ a \mid f$. We call a root a multiple if $(x\ \ a)^2 \mid f$.

By the derivative of the polynomial $f = \sum_{i=0}^{n} a_i x^i$ we mean the polynomial $f' = \sum_{i=0}^{n\ 1} (i+1) a_{i+1} x^i$. It is not difficult to verify (exercise!) that the derivative defined in this way satisfies all the basic formulas you know from the analysis course. The field $\mathbf{R}$ can be any, no properties of real numbers are used for this, the idea of the derivative as a tangent line is not needed.

To prove the classification of finite fields, the following observation will be useful.

Lemma 9.2. Let the polynomial f have a multiple root a. Then $f'(a) = 0$.

Proof. Write $f = (x\ \ a)^2 g$ and use the product formula to calculate the derivative: $f' = 2(x\ \ a)g + (x\ \ a)^2 g'$, i.e. $f'(a) = 0$.

9.3. **Classification of finite fields.**

By applying the theorems on the existence and uniqueness of splitting fields, we show that for every power of a prime number p^k there exists, except for isomorphism, exactly one field of size p^k. The principle of the proof is that a field has exactly p^k elements if and only if it is a splitting field of the polynomial $x^{p^k}\ \ x$ over the field Z_p. From the existence and uniqueness of splitting fields, it follows the existence and uniqueness of finite fields.

Lemma 9.3. The splitting field of the polynomial $x^{p^k}\ \ x$ over the field Z_p has exactly p^k elements.

Proof. Let us denote $q = p^k$. Let $\mathbf{T}$ be the splitting field of the polynomial $f = x^q\ \ x$ over Z_p. We show that the roots of f form a subfield in $\mathbf{T}$. The Frobenius Endomorphism Proposition 9.1 says that the mapping $\varphi : a \mapsto a^p$ is a homomorphism of $\mathbf{T} \to \mathbf{T}$. Its k-fold composition, φ^k, is also a homomorphism and shows $a \mapsto (((a^p)^p)\dots)^p = a^{p^k} = a^q$, that is

$$(a+b)^q = a^q + b^q \quad \text{and} \quad (a\ b)^q = a^q\ b^q$$

for each $a, b \in T$. Thus, if a, b are the roots of the polynomial f, i.e. $a^q = a$ and $b^q = b$, then $(a+b)^q = a^q + b^q = a + b$ is also a root of f and likewise $(a\ b)^q = a^q\ b^q = a\ b$, $(\ a)^q = \ a^q = \ a$ and $(a^{\ 1})^q = (a^q)^{\ 1} = a^{\ 1}$. So the roots form the subfield. It then follows from the minimality requirement that the splitting field $\mathbf{T}$ consists precisely of the roots of f and therefore has at most $\deg f = q$ elements.

To prove that $\mathbf{T}$ has exactly q elements, it suffices to verify that the polynomial f does not have multiple roots. If the element a were a multiple root, according to Lemma 9.2 $f'(a) = 0$ would hold. However, $f' = qx^{q-1} - 1 = -1$, so it has no roots.

Lemma 9.4. Let $\mathbf{T}$ be a finite field, $|T| = p^k$. Then $\mathbf{T}$ is the splitting field of the polynomial $x^{p^k} - x$ over the field Z_p and in $\mathbf{T}[x]$

$$x^{p^k} - x = \prod_{a \in T} (x - a).$$

Proof. Let us denote $q = p^k$. First, note that every element $a \in T$ is a root of the polynomial $f = x^q - x$. This applies trivially for 0, and for a non-zero element a we use Lagrange's theorem: $\mathrm{ord}(a) \mid |T| = q - 1$, i.e. $a^{q-1} = 1$ and $a^q = a$. So $\prod_{a \in T}(x - a) \mid f$, and from the equality of degrees and leading coefficients we get the equality of these polynomials. We have shown that f decomposes into linear factors in $\mathbf{T}$. But is it a minimal expansion? Yes, it is, because according to the previous lemma, the decomposition superfield of the polynomial f has exactly q elements, i.e. $\mathbf{T}$ is this field.

Theorem 9.5 (Classification of finite fields)**.**

(1) A finite field of size n exists if and only if $n = p^k$ for some prime p and natural number k.
(2) Finite fields of the same size are isomorphic.

Proof. (1) ($\Rightarrow$) follows from Assertion 5.3, ($\Leftarrow$) follows from Lemma 9.3 and the theorem on the existence of splitting fields. (2) follows from Lemma 9.4 and Theorem 8.1 on the uniqueness of splitting fields.

In the winter semester, we presented finite fields in the form of $Z_p[\,]/(m)$ factor rings. Can every finite field be represented in this way?

Proposition 9.6 (Representation of finite fields)**.** For every prime p and natural number k there exists an irreducible polynomial $m \in Z_p[\,]$ of degree k and

$$F_{p^k} \simeq Z_p[\,]/(m).$$

Proof. According to Theorem 9.5, there exists a field $\mathbf{T} \supset Z_p$ of size p^k. In the chapter on cyclic groups, we proved that the group $\mathbf{T}^*$ is cyclic. Let a denote some generator and consider the minimal polynomial $m_{a;Z_p}$. The latter is certainly irreducible and its degree is

$$\deg m_{a;Z_p} = [Z_p(a) : Z_p] = [\mathbf{T} : Z_p] = k;$$

while the first equality follows from Assertion 6.6, the second from the fact that $\mathbf{T} = Z_p(a)$ since $\mathbf{T}$ consists of powers of the element a, and the third from the fact that a vector space with p^k elements has dimension k. It follows from the uniqueness in Theorem 9.5 that $\mathbf{T} \simeq Z_p[\,]/(m_{a;Z_p})$.

Note the way in which we proved the existence of an irreducible polynomial of degree k in $Z_p[x]$: we first proved the existence of some field of size p^k in order to take the generator of its multiplicative group and its minimal polynomial. A direct proof of the existence of these polynomials is possible, but it would be much more technical and give less insight of the whole situation.

Algorithms in polynomial arithmetic

10. Modular representations

The goal of this section is to understand modular representations and to become familiar with an algorithm that computes such representations quickly, the so-called fast Fourier transformation (FFT).

The base principle of modular representation is the following: instead of a single computation in a large ring (such as the integers, or a polynomial ring $\mathbf{T}[x]$), we perform several small computations in smaller rings (integers modulo m, or the ring of the coefficients $\mathbf{T}$) and reconstruct from the result the solution of the original problem. Clever use of this idea leads to surprisingly fast algorithms for some tasks.

Let $\mathbf{R}$ be a domain, which admits a division with remainder. Then, a modular representation of $\mathbf{R}$, is an isomorphism

$$\mathbf{R}/(m) \to \mathbf{R}/(m_1) \times \cdots \times \mathbf{R}/(m_n)$$
$$a \mapsto (a \bmod m_1, \ldots, a \bmod m_n)$$

for pairwise coprime elements $m_1, \ldots, m_n \in R$, such that $m = m_1 \cdots m_n$. The fact that this map is an isomorphism follows from the generalized Chinese remainder theorem. (Recall that, in Algebra 1, we did not prove the Chinese remainder theorem in full generality, but for two important special cases: $\mathbf{R} = Z$ and $\mathbf{R} = \mathbf{T}[x]$ for some field $\mathbf{T}$.)

Note that a modular representation does not faithfully represent the entire ring, but only its elements up to multiples of m (in Z this allows for a faithful representation of the numbers $0, 1, \ldots, m-1$; in $\mathbf{T}[x]$ for polynomials of degree less than $\deg m$). In practice, this does not matter, as long as we choose as many elements $m_1, \ldots, m_n$ as are needed to faithfully execute a computation.

There is an obvious algorithm for converting an element to its modular representation: dividing by the remainder. But, this naive algorithm has quadratic complexity, which may be too slow for some applications.

Converting a modular representation back to the original element means solving a system of linear congruences. We already discussed a general algorithm in Algebra 1 and showed some examples in the practicals. It is not difficult to check that this is also an algorithm with quadratic complexity.

Can we find faster algorithms? In general, the answer is no, but, in special cases, the answer is yes. We are going to focus on polynomial rings $\mathbf{T}[x]$ and the special case of linear polynomials $m_i = x - u_i$, where the roots u_i are two different elements. In that case, the map computing a modular representation is the substitution map $f \mapsto (f(u_1), \ldots, f(u_n))$.

Thus converting an element quickly to a modular representation and back corresponds to finding fast algorithms for substituting values in polynomials / interpolating values by polynomials. To achieve this faster than in quadratic time is seemingly impossible: after all, we have n values and the polynomial has n terms, so how do we get under quadratic time? The solution is to choose the root elements u_i in a smart way.

10.1. **Discrete Fourier transformations.** In this whole section, let $\mathbf{T}$ be a fixed field. We say that an element $\omega \in T$ is a primitive n-th root of unity if its order is $\operatorname{ord}(\omega) = n$ in the group $\mathbf{T}^\times$. In other words

(1) $!^n = 1$,
(2) $!^i \neq 1$ for all $i = 1; 2; :::; n \quad 1$.

Note that, by Lagrange's theorem, it is su cient to test condition (2) only for $i \mid n$; this condition also implies that $!^i \neq !^j$ for all $i \neq j < n$.

To represent polynomials of degree $< n$ over the eld $\mathbf{T}$, we consider their values at points $1; !; !^2; :::; !^{n \quad 1}$, where $!$ is a primitive n-th root of unity in $\mathbf{T}$. Such a representation is called a discrete Fourier transform and a fast algorithm for its calculation is called a fast Fourier transform. Under certain assumptions, we get a fast Fourier transform algorithm of time complexity $O(n \log n)$.

Examples.

The eld C contains a primitive n-th root of unity for every $n \geq 2$ N, e.g.

$$! = e^{\frac{2 i}{n}} = \cos(2 =n) + i \sin(2 =n):$$

Note that $!$ generates a cyclic group can be seen as the vertices of a regular n-gon on the unit circle in the complex plane. Each of its generators is also a primitive n-th root.

In Q, $! = \quad 1$ is a primitive square root of unity. There are no other primitive roots of unity.

The nite eld F_q contains the primitive n-th root of unity if and only if $n \mid q \quad 1$: in the chapter on cyclic groups (Algebra 1) we proved that the group F_q is cyclic, and it has $q \quad 1$ elements. Thus it contains an element of order n if and only if $n \mid q \quad 1$.

Definition. The Discrete Fourier transform at the point $!$ is the mapping $DFT_! : T^n \to T^n$ de ned by

$$DFT_!(a_0; :::; a_{n \quad 1}) = (f(!^0); f(!^1); :::; f(!^{n \quad 1}));$$

where $f = \sum_{i=0}^{n \quad 1} a_i x^i$.

The value of the polynomial $f = \sum_{i=0}^{n \quad 1} a_i x^i$ at the point can be written as a matrix product

$$f( ) = \sum_{i=0}^{n \quad 1} a_i \quad ^i = \begin{pmatrix} 1; & ; :::; & ^{n \quad 1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n \quad 1} \end{pmatrix}:$$

In general, if we substitute n values of $_1; :::; _n$, we get the expression

$$\begin{pmatrix} f(_1) \\ f(_2) \\ \vdots \\ f(_n) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 & ::: & n \quad 1 \\ 1 & 1 & 1 & ::: & 1 \\ 0 & 1 & 2 & ::: & n \quad 1 \\ 2 & 2 & 2 & ::: & 2 \\ & & & \vdots & \\ 0 & 1 & 2 & ::: & n \quad 1 \\ n & n & n & ::: & n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n \quad 1} \end{pmatrix}:$$

Thus, the modular representation at the points $_1; :::; _n$ is actually a linear map (an endomorphism of the vector space $\mathbf{T}^n$)

$$' : \mathbf{T}^n \to \mathbf{T}^n; \qquad u \mapsto A \quad u;$$

where $u = (a_0; a_1; :::; a_{n \quad 1})^T$ is the column vector of coe cients and A is the so-called Vandermond matrix mentioned above. This matrix is regular if and only if the elements of $_i$ are two di erent. Thus, the discrete Fourier transform at the point $!$ is a linear mapping

$$DFT_! : \mathbf{T}^n \to \mathbf{T}^n; \qquad u \mapsto A_! \quad u;$$

where A_ω is the Vandermond matrix corresponding to the elements $\omega^0, \omega^1, \ldots, \omega^{n-1}$, i.e.

$$A_\omega = (\omega^{ij})_{i,j=0}^{n-1}.$$

Since DFT_ω is a bijection, we can consider the inverse of DFT_ω^{-1}, which is called inverse DFT, also denoted by IDFT_ω. By the properties of DFT_ω, the inverse corresponds to the interpolation of values at the mentioned points $\omega^0, \omega^1, \ldots, \omega^{n-1}$ by a polynomial of degree n. Note that

$$\mathrm{IDFT}_\omega(u) = A_\omega^{-1} \cdot u.$$

The following statement shows that the IDFT is a special case of the DFT. This will allow us to focus on just one algorithm for converting between normal and modular representations.

Proposition 10.1. If ω is a primitive n-th root of unity in $\mathbf{T}$, then

$$(A_\omega)^{-1} = \frac{1}{n} \cdot A_{\omega^{-1}}.$$

For the statement to make sense at all, we must first check that the characteristic of the field $\mathbf{T}$ does not divide n. This can be deduced from the fact that there is a primitive n-th root of unity in $\mathbf{T}$ (exercise!).

Proof. It is enough to show that $A_\omega \cdot (\frac{1}{n} \cdot A_{\omega^{-1}})$ is the identity matrix. Since

$$A_\omega = \left(\omega^{ij}\right)_{i,j=0}^{n-1} \quad \text{and} \quad A_{\omega^{-1}} = \left(\omega^{-ij}\right)_{i,j=0}^{n-1},$$

we obtain the following matrix-product:

$$A_\omega \cdot A_{\omega^{-1}} = \left(\sum_{k=0}^{n-1} \omega^{ik}\omega^{-kj}\right)_{i,j=0}^{n-1}.$$

For $i = j$ we obtain the value

$$\sum_{k=0}^{n-1} \omega^{ik}\omega^{-ki} = \sum_{k=0}^{n-1} 1 = n;$$

and for $i \neq j$ we obtain

$$\sum_{k=0}^{n-1} \omega^{ik}\omega^{-kj} = \sum_{k=0}^{n-1} \omega^{k(i-j)} = \sum_{k=0}^{n-1} (\omega^{i-j})^k;$$

which is a geometric sum with base $\omega^{i-j} \neq 1$ (since ω is a primitive n-th root and $|i-j| < n$). Thus

$$\sum_{k=0}^{n-1} (\omega^{i-j})^k = \frac{(\omega^{i-j})^n - 1}{\omega^{i-j} - 1} = \frac{1-1}{\omega^{i-j} - 1} = 0;$$

as $\omega^n = 1$. In conclusion, we proved that there are n elements on the diagonal and zeros off the diagonal, so after multiplying $\frac{1}{n}$ we get an identity matrix.

10.2. **Fast Fourier transform.** The fast Fourier transform (FFT) is a fast algorithm for calculating the DFT. It relies on the divide and conquer method. The idea is as follows: if we substitute the value of x in the polynomial $f = \sum_{i=0}^{n-1} a_i x^i$ of odd degree (i.e. n is even), we can write

$$f(x) = \underbrace{(a_0 + a_2 x^2 + \ldots + a_{n-2} x^{n-2})}_{g(x^2)} + x \cdot \underbrace{(a_1 + a_3 x^2 + a_5 x^4 + \ldots + a_{n-1} x^{n-2})}_{h(x^2)};$$

i.e.

$$f(x) = g(x^2) + x \cdot h(x^2);$$

where $g; h$ are polynomials of at most half the degree of f, de ned by

$$g = \sum_{i=0}^{\frac{n}{2}\ 1} a_{2i} x^i \quad \text{and} \quad h = \sum_{i=0}^{\frac{n}{2}\ 1} a_{2i+1} x^i:$$

Thus, we divided the task of substituting the value of into a polynomial with n coe cients into two tasks of substituting the value of ^{2} into half size polynomials.

Now recall that, in DFT, we need to substitute the n-many values $!^0; \ldots; !^{n\ 1}$. The second part of the trick consists in the observation that we actually only need to substitute the values $!^0; \ldots; !^{n=2\ 1}$. For the primitive n-th root of one, it holds

$$!^{n=2+i} = !^{n=2}\ \ !^i = (\ 1)\ \ !^i = \ \ !^i;$$

so both $!^i$ and $!^{n=2+i}$ have the same square. Thus, in g and h we only need to substitute half of them. So, we have divided the original task into two smaller subtasks: we need to substitute half the values into half the polynomials.

Algorithm 1 (fast Fourier transform, FFT)**.**
Input: $n = 2^k$, $!$ primitive n-th roof of unity, $a_0; a_1; \ldots; a_{n\ 1}$
Output: $\mathrm{DFT}_!(a_0; a_1; \ldots; a_{n\ 1})$
0. **if** $n = 1$ **then return** a_0
1. $(b_0; \ldots; b_{\frac{n}{2}\ 1}) := \mathtt{FFT}(n{=}2; !^2; a_0; a_2; \ldots; a_{n\ 2})$
   $(c_0; \ldots; c_{\frac{n}{2}\ 1}) := \mathtt{FFT}(n{=}2; !^2; a_1; a_3; \ldots; a_{n\ 1})$
2. $d_i := b_i + !^i c_i, \quad d_{\frac{n}{2}+i} := b_i \ \ !^i c_i$ for all $i = 0; \ldots; \frac{n}{2}\ \ 1$
   return $(d_0; \ldots; d_{n\ 1})$

Proposition 10.2. Algorithm 1 is correct.

Proof. We perform the proof by induction on n. For $n = 1$, $\mathrm{DFT}_!$ outputs a_0, which is correct.

We next look at an induction step $\frac{n}{2}\ !\ n$. First, note that $!^2$ is a primitive $\frac{n}{2}$-th root of one: clearly $(!^2)^{n=2} = !^n = 1$ and further for all $i = 1; 2; \ldots; \frac{n}{2}\ \ 1$ $(!^2)^i = !^{2i} \neq 1$ because $2i < n$ and $!$ is a primitive square root.

So let $f = \sum_{i=0}^{n\ 1} a_i x^i$ and de ne the polynomials $g; h$ as above. By induction assumption

$$(b_0; \ldots; b_{\frac{n}{2}\ 1}) = (g(1); g(!^2); g(!^4); \ldots; g(!^{n\ 2}));$$
$$(c_0; \ldots; c_{\frac{n}{2}\ 1}) = (h(1); h(!^2); h(!^4); \ldots; h(!^{n\ 2})):$$

We want to prove that for $i = 0; 1; \ldots; \frac{n}{2}\ \ 1$

$$d_i = f(!^i) \qquad \text{and} \qquad d_{i+\frac{n}{2}} = f(!^{i+n=2}):$$

The rst identity follows directly from the formula derived above:

$$f(!^i) = g(!^{2i}) + !^i h(!^{2i}) = b_i + !^i c_i = d_i:$$

The second identity is similarly obtained

$$f(!^{i+n=2}) = g(!^{2i+n}) + !^{i+n=2} h(!^{2i+n}) = b_i \ \ !^i c_i = d_{i+\frac{n}{2}}$$

using the easy observation that $!^{2i+n} = !^{2i} !^n = !^{2i}$ and that $!^{i+n=2} = !^i !^{n=2} = \ !^i$. Here $!^{n=2} = \ 1$, because it is the square root of one, and there are only two: 1 and $\ 1$.

Proposition 10.3. Algorithm 1 has running time $O(n \log n)$.

Proof. We will follow a proven scheme in analyzing divide-and-conquer algorithms: Let us denote by $T(n)$ the number of operations in the eld $\mathbf{T}$ that the algorithm performs on an input of length n. Then $T(1) = 0$ and

$$T(n) = 2T(n{=}2) + O(n);$$

therefore $T(n) \quad 2T(n=2) + cn$ for some c and by substituting $n = 2^k$ we get

$$\begin{aligned} T(2^k) \quad & 2T(2^k \ \ 1) + c2^k \quad 2(2T(2^k \ \ 2) + c2^k \ \ 1) + c2^k \\ & 4T(2^k \ \ 2) + c(2^k + 2^k) \\ & \vdots \\ & 2^k T(2^k \ \ k) + ck2^k = 2^k T(1) + ck2^k = O(k2^k): \end{aligned}$$

Hence $T(n) = O(n \log n)$:

Example. For the field $\mathbf{T} = Z_{41}$, calculate the modular representation of the polynomial

$$5x^3 + x + 1:$$

First, we need to find the primitive fourth root of 1: e.g. $! = \ 9$ (we see that $!^2 = \ 1$, $!^3 = 9$ and $!^4 = 1$). Next, we calculate DFT $_9(1; 1; 0; 5)$. We divide the task into

$$\text{DFT} \ _1(1; 0) = (1; 1) \qquad \text{and} \qquad \text{DFT} \ _1(1; 5) = (6; \ \ 4):$$

The result is

$$(1 + (\ \ 9)^0 \ \ 6; 1 + (\ \ 9)^1 \ \ (\ \ 4); 1 \quad (\ \ 9)^0 \ \ 6:1 \quad (\ \ 9)^1 \ \ (\ \ 4)) = (7; \ \ 4; \ \ 5; 6):$$

10.3. **Primitive roots of unity.** The existence of a primitive root of unity is essential for FFT to work. But, if it does not exist in the field $\mathbf{T}$, we can sometimes work in an alternative field instead.

For finite fields F_q the situation is clear: the primitive n-th root in F_q exists if and only if $n \ j \ q \quad 1$ and we obtain it as a power of $! = a^{(p \ \ 1)=n}$, where a is the generator of the cyclic group F_q. The generator a can be found by random selection | the probability of success is $'(q \quad 1)=(q \quad 1)$, which is typically a relatively large fraction. (See Algebra 1.)

We also discuss the case of the rational field Q in detail. Unfortunately, there are no primitive roots of one (except for the element $\ \ 1$). But, FFT is still used, and there are basically two ways to do it. In both cases, it is a good idea to convert to polynomials over Z first: we simply multiply the input values by a large enough integer so that all denominators disappear (and revert this, after we are done with our computations).

The first option is to work in the complex numbers with approximate values for

$$! = e^{\frac{2 \ i}{n}} = \cos(2 \ =n) + i \sin(2 \ =n):$$

This is a numeric method, which requires to use floating-point arithmetic. There are rigorous estimates of how precisely we need to approximate $!$ to get the desired precision of the result. If we know that the result is an integer, absolute precision can be achieved: just take $!$ so that the resulting error is less than $\frac{1}{2}$ and simply round the resulting values. Compared to the modular method below, however, this strategy turns out to be more time-consuming.

If we have an integer polynomial as an input, the following trick can be used: instead of Z we do computations in the field Z_p, where p is chosen so large that modular arithmetic does not affect the result. The specific implementation depends on the given problem, we will illustrate the method in the next section using the example of multiplying integer polynomials. Let us note at this point that we will mainly be interested in those prime numbers p for which $p \quad 1$ is divisible by a sufficiently large power of two { large enough that Z_p contains a primitive root suitable for FFT. Such primes are sometimes called FFT-primes, a relatively large power of two divides $p \quad 1$, e.g. for primes 17, 41, 97. Where to get them? The existence of FFT-primes follows from Dirichlet's theorem: it says that for every

coprime $a; m$ there are in nitely many primes $\equiv a \pmod m$, i.e., in particular in nitely many primes $\equiv 1 \pmod{2^k}$ for any k. However, Dirichlet's theorem does not say anything about how these prime numbers are distributed, i.e. how long we will search for some. The exact answer to this question is not yet known (the Riemann hypothesis, a problem for which a million dollar reward is o ered, implies that we don't have to look for long).

Exercises.

1. Prove that if there is a primitive n-th root of unity in $\mathbf{T}$, then the characteristic of the eld $\mathbf{T}$ does not divide n.

2. Let $(2; 0; 2; 0)$ be the modular representation of the polynomial f given by DFT_4 in Z_{17}. Compute the polynomial f using the FFT.

3. Compute the time complexity of one attempt of the probabilistic algorithm to nd the primitive n-th root of unity in the eld Z_p, where $n \mid p - 1$. (Recall that we choose a at random, compute $\omega = a^{(p-1)/n}$ and verify that $\omega^1; \ldots; \omega^{n-1} \neq 1$.) Why is this algorithm ine cient at the task of nding the generator of the group Z_p ?

4. The non-existence of the n-th primitive root of unity in the eld $\mathbb{Q}$ could be solved by working in the splitting eld $\mathbb{Q}(\omega)$ of the polynomial $x^n - 1$, which apparently contains such a root. Why is this not a good idea?

11. Fast polynomial multiplication and division

11.1. **Fast multiplication.**

Perhaps the best known application of modular representations is a fast algorithm for multiplying polynomials. This algorithm is practically used in all real-world applications since, even for relatively small polynomials, it is usually faster than the \long multiplication" from school (depending on the eld $\mathbf{T}$).

Recall that a modular representation is a ring homomorphism, i.e. it preserves the basic operations of addition and multiplication. Speci cally for polynomials, we consider maps

$$\begin{aligned} \varphi : \mathbf{T}[x]/m &\to \mathbf{T}^n \\ f &\mapsto (f(\alpha_1); \ldots; f(\alpha_n)); \end{aligned}$$

where $m = (x - \alpha_1) \cdots (x - \alpha_n)$. It then holds that

$$\varphi(f \cdot g \bmod m) = \varphi(f) \cdot \varphi(g);$$

here the multiplication in $\mathbf{T}^n$ is understood component-wise. When choosing

$$n > \deg(f \cdot g) = \deg f + \deg g$$

we get $\varphi(f \cdot g) = \varphi(f) \cdot \varphi(g)$, i.e. products in the modular representation corresponds exactly to the product in the standard representation of polynomials. The advantage of the modular representation is that we only need to perform n operations in the eld $\mathbf{T}$ to compute the product.

The algorithm for fast multiplication in $\mathbf{T}[x]$ is based on this principle. First, we choose a suitable modular representation for the given polynomials { except for the condition $n > \deg f + \deg g$, we also need to choose a representation for which we have a fast conversion algorithm, such as FFT. Then we compute the modular representations $a; b$ of the polynomials $f; g$, calculate their (componentwise) product $c = a \cdot b$ and use again FFT to compute the inverse discrete Fourier transform of c, i.e. $f \cdot g$.

Algorithm 2 (modular multiplication of polynomials)**.**
Input: $f; g \in T[x]$
Output: $f \cdot g$
0. choose $n > \deg f + \deg g$ and suitable points $\alpha_1, \ldots, \alpha_n \in T$
1. $a = (f(\alpha_1), \ldots, f(\alpha_n))$
   $b = (g(\alpha_1), \ldots, g(\alpha_n))$
2. $c = a \cdot b$
3. **return** polynomial h of degree $< n$ such that $(h(\alpha_1), \ldots, h(\alpha_n)) = c$

Example. We compute the product of the polynomials $f = \frac{1}{2}x^2 + \frac{1}{2}x + 1; g = x^3 - \frac{1}{3}x \in \mathbb{Q}[x]$:

0. we choose, for example $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = -1$, $\alpha_4 = 2$, $\alpha_5 = -2$, $\alpha_6 = 3$,
1. $a = (1; 2; 1; 4; 2; 7)$, $b = \frac{1}{3}(0; 2; -2; 22; -22; 78)$,
2. $c = a \cdot b = \frac{1}{3}(0; 4; -2; 88; -44; 546)$,
3. by interpolation we find $h = f \cdot g = \frac{1}{2}x^5 + \frac{1}{2}x^4 + \frac{5}{6}x^3 - \frac{1}{6}x^2 - \frac{1}{3}x$.

If we leave aside the choice in step 0., the algorithm involves two transformations: one to the modular representation of polynomials of degree $< n$, one from this modular representation, and n multiplications in the field $\mathbf{T}$. By randomly choosing $\alpha_1, \ldots, \alpha_n$, using simple value substitution and standard interpolation algorithms, we get a time complexity of $3 \cdot O(n^2) + O(n) = O(n^2)$, which is good for nothing. With the smart choice of the DFT representation and the use of the FFT algorithm, we get a time complexity of

$$3 \cdot O(n \log n) + O(n) = O(n \log n).$$

(Formally speaking, this is the complexity expressed with respect to the parameter n, which we choose in step 0. However, this n typically depends linearly on the sum of the degrees of the given polynomials – e.g. in the case of FFT, we choose $n = 2^k > \deg f + \deg g$, i.e. $n < 2(\deg f + \deg g)$, so the complexity has the same asymptotic behavior also with respect to the sum of the degrees of the input polynomials.)

In the rest of the section, we will discuss a concrete implementation of step 0. If there is a primitive n-th root of unity in $\mathbf{T}$, we can straightforwardly apply FFT. If $\mathbf{T} = Z_p$ and the corresponding root does not exist, we can instead solve the problem in Z and then reduce the result modulo p. Multiplication of rational polynomials can also be easily converted to multiplication over Z: we multiply the specified polynomials by numbers $u; v$ that the denominators in the coefficients disappear, perform the product of integer polynomials and divide the result by the product uv. So we will now solve the multiplication in $Z[x]$ problem.

As we mentioned in the previous section, one option is numerical computation with the complex root $\omega = e^{2\pi i/n}$. Since the result is an integer, it is enough to choose such precision that the resulting error is $< \frac{1}{2}$ (specific estimates can be found in the literature).

Example. We compute the product of the polynomials $f = \frac{1}{2}x^2 + \frac{1}{2}x + 1$ and $g = x^3 - \frac{1}{3}x$, or, equivalently, the product of $2f = x^2 + x + 2$ and $3g = 3x^3 - x$.

0. We choose $n = 8 = 2^3$ and the representation DFT_ω for $\omega = e^{2\pi i/8} \doteq 0{:}71 + 0{:}71i$. We will perform computations up to two decimal points.
1. $a = \mathrm{DFT}_\omega(2; 1; 1; 0; 0; 0; 0; 0) =$

$$(4; 2{:}71 + 1{:}72i; 0{:}98 + 1{:}01i; 1{:}28 - 0{:}31i; 2{:}02; 1{:}28 + 0{:}32i; 0{:}95 - 1{:}02i; 2{:}73 - 1{:}79i);$$

$b = \mathrm{DFT}_\omega(0; -1; 0; 3; 0; 0; 0; 0) =$

$$(2; -2{:}86 + 1{:}44i; -4{:}08i; 2{:}92 + 1{:}48i; -2{:}13; 2{:}98 - 1{:}53i; 4{:}25i; -3{:}04 - 1{:}58i);$$

2. $c = a \cdot b =$

$$(8; -10{:}21 - 1{:}01i; -4{:}12 - 4{:}02i; -4{:}20 + 1{:}01i;$$
$$-4{:}30; -4{:}30 - 1{:}01i; -4{:}36 + 4{:}04i; -11{:}12 + 1{:}11i);$$

3. $\frac{1}{8}\mathrm{DFT}_{\omega^{-1}}(c) = (0{:}03; -2{:}00; -1{:}00; 4{:}98; 3{:}00; -2{:}98; 0{:}00; 0{:}00)$. After rounding, we get the (correct) result

$$h = 2f \cdot 3g = 3x^5 + 3x^4 + 5x^3 - x^2 - 2x.$$

So $f \quad g = \frac{1}{6} \quad h = \frac{1}{2}x^5 + \frac{1}{2}x^4 + \frac{5}{6}x^3 \quad \frac{1}{6}x^2 \quad \frac{1}{3}x$.

In practice, the modular method turns out to be more e ective. Instead of in $Z[x]$, we will compute the product in $Z_p[x]$, where we choose the prime p large enough such that

$$f \quad g = (f \quad g) \bmod p;$$

i.e., so large that no coe cient of the product exceeds $p=2$ in absolute value. We interpret the elements of Z_p here as $dp=2e; : : : ; \quad 1; 0; 1; : : : ; bp=2c$. How big p do we need? Let's recall the formula

$$\left(\sum_{i=0}^{n} a_i x^i\right) \quad \left(\sum_{i=0}^{m} b_i x^i\right) = \sum_{i=0}^{m+n} \left(\sum_{j=0}^{i} a_j b_{i \quad j}\right) x^i$$

(assume $n \quad m$). Each coe cient of the polynomial $f \quad g$ is the sum of at most $n + 1$ products of $a_j \quad b_{i \quad j}$. If we denote $r = \max ja_i j$ and $s = \max jb_i j$, then the absolut value of each coe cient $f \quad g$ is bounded by $(n + 1)rs$. For p, we therefore choose a prime number greater than

$$2(n + 1)rs;$$

which also satis es the FFT condition $2^k \mid p \quad 1$, for some $2^k > m + n$.

Example. We compute the product of the polynomials $f = \frac{1}{2}x^2 + \frac{1}{2}x + 1$ and $g = x^3 \quad \frac{1}{3}x$. In fact, we will compute the product of $2f = x^2 + x + 2$ and $3g = 3x^3 \quad x$.

0. We see that $r = 2$, $s = 3$, so we need a prime number $p > 2 \quad 4 \quad 2 \quad 3 = 48$ satisfying $n = 2^3 \mid p \quad 1$. We will therefore compute in Z_{97} and choose the $\mathrm{DFT}_!$ representation, e.g. for $! = 50$.
1. $a = \mathrm{DFT}_!(2; 1; 1; 0; 0; 0; 0; 0) = (4; 30; 76; 88; 2; 27; 23; 57)$;
   $b = \mathrm{DFT}_!(0; \quad 1; 0; 3; 0; 0; 0; 0) = (2; 45; 88; 86; 95; 52; 9; 11)$,
2. $c = a \quad b = (8; 89; 92; 2; 93; 46; 13; 45)$,
3. $\frac{1}{8}\mathrm{DFT}_{! \quad 1}(c) = (0; \quad 2; \quad 1; 5; 3; 3; 0; 0)$, so $h = 3x^5 + 3x^4 + 5x^3 \quad x^2 \quad 2x$.

So $f \quad g = \frac{1}{6} \quad h = \frac{1}{2}x^5 + \frac{1}{2}x^4 + \frac{5}{6}x^3 \quad \frac{1}{6}x^2 \quad \frac{1}{3}x$.

Various variants of the above principle are used in practice. E.g. the following two enhancements are implemented in the NTL library (Number Theory Library for C++):

(Chinese Remainder Theorem) A element of $Z_p[x]$ is represented by $Z_{p_1}[x]$, ..., $Z_{p_N}[x]$, with the result reconstructed using the Chinese remainder theorem performed on each coe cient of the resulting polynomial separately. Of course, all p_i must be FFT primes, in the sense of $2^k \mid p_i \quad 1$. If $p_1 \quad : : : \quad p_N > 2(n + 1)rs$, we are guaranteed to get the correct result. The advantage of this method is that all p_i usually t into one machine word, so operations in Z_{p_i} are very fast.

(Schönhage-Strassen's trick) Instead of a prime number p, we pick a value $M = 2^{2^k \quad 1 u} + 1$, where u is large enough so that $M > 2(n + 1)rs$. The ring Z_M is not a eld, but this does not matter for FFT to work (left as exercise). Importantly, the element 2^u is the 2^k-th primitive root of unity (exercise!). The advantage of these rings is that computations modulo a number of the form $M = 2^e + 1$ are very fast (linear, compared to the classic quadratic): because $2^e \quad 1 \pmod M$, to reduce the number a modulo M it su ces to perform $b\log_M ac$ operations of addition and subtraction modulo M; multiplication and division by the power of two are implemented as a bit shift, i.e. cost (almost) nothing.

Exercises.

1. Compute the product of the polynomials $f = x + 1$ and $g = x^2 \quad 1$ by all the mentioned methods (random point interpolation, numerical FFT, modular FFT).

2. Why is it not possible to compute the quotient and the remainder of a polynomial division in an analogous way?

3. Let us have two polynomials of degree n as input and assume that we know the corresponding primitive root (i.e., step 0. is trivial). Compute the exact (not asymptotic) time complexity of fast multiplication using algorithm 1 for the FFT. Compare it with long multiplication.

4. Formulate an e cient (subquadratic) algorithm for dividing the number $a \in \mathbb{N}$ by the number $M = 2^e + 1$ and estimate its time complexity with respect to $\ell(a)$ and $e = O(\ell(M))$ (ℓ denotes the length of the binary expansion). Compute the time complexity of multiplication in the ring $\mathbb{Z}_M$ depending on $\ell(M)$.

11.2. **Fast polynomial division.** Unfortunately, modular representations cannot be used directly to divide polynomials. But, we are going to present a more sophisticated algorithm that is based on fast multiplication and the computation of inverse elements in the ring of formal power series. We already know how to multiply quickly. For the fast computation of inverse elements, we will present a variant of Newton's method.

Let us brie y repeat the notion of formal power series over the commutative ring $\mathbf{R}$. Its elements are formal expressions of the form $\sum_{i=0}^{\infty} a_i x^i$, where a_i are coe cients from R. The addition and multiplication on these expressions are similarly de ned as for polynomials:

$$\left(\sum_{i=0}^{\infty} a_i x^i\right) \quad \left(\sum_{i=0}^{\infty} b_i x^i\right) = \sum_{i=0}^{\infty} (a_i \quad b_i) x^i;$$

$$\left(\sum_{i=0}^{\infty} a_i x^i\right) \left(\sum_{i=0}^{\infty} b_i x^i\right) = \sum_{i=0}^{\infty} \left(\sum_{j=0}^{i} a_j b_{i \quad j}\right) x^i.$$

We denote the ring of formal power series over a ring $\mathbf{R}$ by $\mathbf{R}[[x]]$. If $\mathbf{R}$ is an integral domain, then $\mathbf{R}[[x]]$ is also an integral domain. The polynomials over $\mathbf{R}$ always form a subring of $\mathbf{R}[[x]]$. We will further need the following important property.

Proposition 11.1. Let $\mathbf{R}$ be an integral domain and $f = \sum a_i x^i \in R[[x]]$. Then f is invertible in $\mathbf{R}[[x]]$ if and only if a_0 is invertible in $\mathbf{R}$.

The proof of this statement is not di cult and follows from the considerations in Section 11.3, where we will deal with the e cient computation of inverse power series. For now, let's see how polynomial division can be converted to this problem.

For the purposes of this section, we introduce a technical notation: for the polynomial f we de ne

$$f^* = x^{\deg f} \quad f(x^{\quad 1}).$$

In other words, f^* is a polynomial that arises from f when we write its terms in reverse order. E.g. for $f = 3x^3 + 2x^2 \quad 1$ is $f^* = x^3 \quad (3x^{\quad 3} + 2x^{\quad 2} \quad 1) = 3 + 2x \quad x^3$.

Let $\mathbf{T}$ be a eld and consider the polynomials $f, g \in T[x]$, $g \neq 0$. Denote $n = \deg f$, $m = \deg g$ and assume $n \geq m$. We want to compute the quotient and the remainder, i.e. we are looking for polynomials $q, r \in T[x]$ satisfying

$$f = gq + r, \qquad \deg r < m.$$

For such polynomials clearly also

$$f(x^{\quad 1}) = g(x^{\quad 1})q(x^{\quad 1}) + r(x^{\quad 1})$$

holds, and after multiplying x^n we get the condition

$$f^* = g^* q^* + x^{n \quad \deg r} r^*.$$

Next, let's work in the ring of power series $\mathbf{T}[[x]]$. Since g^* is guaranteed to have a nonzero absolute term, there exists an inverse $(g^*)^{\quad 1} \in \mathbf{T}[[x]]$. After multiplying by this series we get the expression

$$q^* = f^* \quad (g^*)^{\quad 1} \quad x^{n \quad \deg r} \quad r^* \quad (g^*)^{\quad 1}.$$

The left-hand side of the equality is a polynomial of degree at most $n - m$, so on the right-hand side, all coefficients of powers x^{n-m+1} and bigger are equal to 0. But the coefficients of the power series $x^{n-\deg r} r(g)^{-1}$ are 0 for the first $n - \deg r > n - m$ terms, so q is in fact equal to the first $n - m + 1$ terms of the power series $f(g)^{-1}$!

Algorithm 3 (fast polynomial division).
Input: $f, g \in T[x]$, $g \neq 0$
Output: f div g, f mod g
0. $n := \deg f$, $m := \deg g$, **if** $n < m$ **then return** $0, f$
1. $h :=$ the first $n - m + 1$ terms of $(g)^{-1}$
2. $w := fh \bmod x^{n-m+1}$
3. let q be the polynomial of degree $n - m$, for which $q = w$
4. **return** $q, f - gq$

Note that in step 3, we cannot simply set $q = w$, because the degrees may not fit: e.g. for $f = x^3$, $g = x$, $w = 1$, but $q = x^2 \neq w$.

The correctness of the algorithm follows from the analysis above it. In terms of complexity, we converted division with a remainder to two multiplications and one inverse power series computation. Specifically, for an input of degree n, m, we perform two multiplications of polynomials of degree $< n$ and look for $n - m + 1 - n$ terms of the power series $(g)^{-1}$. We can solve the multiplication in time $O(n \log n)$. If we could search for the first n terms of the inverse power series with the same complexity, we would get the time complexity of the division

$$O(n \log n).$$

(In practice, the division of integer polynomials is roughly five times slower than multiplication, see the exercise.)

Example. We compute the quotient and remainder of

$$f = x^4 + x^3 + x^2 + x + 1; \qquad g = x^2 + x - 1.$$

1. For $g = -x^2 + x + 1$ we get $(g)^{-1} = 1 - x + 2x^2 + \ldots$, thus $h = 2x^2 - x + 1$.
2. $w = 1 + 2x^2 + 2x^3 + \ldots \bmod x^3 = 2x^2 + 1$.
3. $q = w - 1 = x^2 + 2$.
4. The answer is f div $g = x^2 + 2$ and f mod $g = -x + 3$.

11.3. **The computation of inverses of power series.**

To complete the fast algorithm for dividing polynomials, it remains to find an algorithm for computing the initial terms of the inverse power series of a given polynomial. So let us consider the power series $f = \sum a_i x^i \in T[[x]]$, we then look for the series $g = \sum b_i x^i \in T[[x]]$ such that $f \cdot g = 1$.

Recall the Proposition 11.1. If $a_0 = 0$, then $x \mid f$, and thus there can be no such series g. So suppose $a_0 \neq 0$. The following conditions follow from the formula for multiplying power series:

$$\begin{aligned} a_0 b_0 &= 1 \\ a_0 b_1 + a_1 b_0 &= 0 \\ a_0 b_2 + a_1 b_1 + a_2 b_0 &= 0 \\ a_0 b_3 + a_1 b_2 + a_2 b_1 + a_3 b_0 &= 0 \\ &\ldots \end{aligned}$$

We can rewrite this to

$$\begin{aligned} b_0 &= a_0^{-1} \\ b_1 &= -a_0^{-1}(a_1 b_0) \\ b_2 &= -a_0^{-1}(a_1 b_1 + a_2 b_0) \\ b_3 &= -a_0^{-1}(a_1 b_2 + a_2 b_1 + a_3 b_0) \\ &\vdots \end{aligned}$$

and see that $g = f^{-1}$ exists (and is uniquely determined). At the same time we got an algorithm to compute g.

To obtain the first n members of $b_0, \ldots, b_{n-1}$ we need to compute 1 inverse, $2 + 3 + 4 + \cdots + n$ multiplication operations and $1 + 2 + 3 + \cdots + (n-1)$ addition operations in the field $\mathbf{T}$, i.e.

$$O(n^2)$$

operations in total. But that's too much!

For a reasonable algorithm, we would like a runtime of $O(n \log n)$. For this, it would be enough to find a procedure that would computes the n-th coefficient in only $O(\log n)$ steps. Or, in other words, we would like a procedure that doubles in every step the number of coefficients that we obtain. We will show a procedure inspired by Newton's method for finding the roots of arbitrary functions (see any introduction to numerical mathematics).

The computation of the first n terms of the series f^{-1} can be interpreted in the following way: for the given series f we are looking for a polynomial g of degree $< n$ such that

$$f \cdot g = 1 + 0x + 0x^2 + \cdots + 0x^{n-1} + x^n \cdot h;$$

where h is an arbitrary series, or in other words

$$f \cdot g \equiv 1 \pmod{x^n}.$$

(In the end, only the first n terms of the f series matter.)

Let's rewrite this relation as $x^n \mid fg - 1$. Then $x^{2n} \mid (fg-1)^2 = f^2g^2 - 2fg + 1$ and we get

$$f \cdot g \cdot (2 - fg) \equiv 1 \pmod{x^{2n}}.$$

We see that $g \cdot (2 - fg) \bmod x^{2n}$ consists of the first $2n$ terms of the inverse power series f^{-1} and thus we get, as promised, that one step doubles the number of terms double.

Algorithm 4.
Input: n, $f = \sum a_i x^i$ with $a_0 \neq 0$
Output: first n coefficients of f^{-1}
0. $g_0 := a_0^{-1}$
1. **for** $i = 1, \ldots, \lceil \log_2 n \rceil$ **do**
   $g_i := g_{i-1} \cdot (2 - f g_{i-1}) \bmod x^{2^i}$
2. **return** $g_{\lceil \log_2 n \rceil} \bmod x^n$

The correctness of the algorithm follows from the above considerations: each g_i contains exactly 2^i members of the f^{-1} power series. Regarding the time complexity, in the i-th step of the loop, we use the polynomial $f \bmod x^{2^i}$ of degree $< 2^i$ and the polynomial g_{i-1} of degree $< 2^{i-1}$. The i-th step (computing one difference and two multiplications) has thus a running time of $O(2^i \log 2^i) = O(i2^i)$ and we can express the total time complexity

$$O\left(\sum_{i=1}^{\lceil \log_2 n \rceil} i2^i\right) = O\left(\log n \cdot \sum_{i=1}^{\lceil \log_2 n \rceil} 2^i\right) = O\left(\log n \cdot (2^{\lceil \log_2 n \rceil + 1} - 1)\right) = O(n \log n)$$

(we used the estimate $i \quad d\log_2 ne$ and the formula for geometric sums).

Example. We compute the first 4 entries of the power series of $f^{\ 1}$, where

$$f = 1 \quad 2x + 3x^2 + x^4 \quad x^5:$$

The classical method yields

$$b_0 = a_0^{\ 1} = 1$$
$$b_1 = \quad a_0^{\ 1}(a_1 b_0) = \quad (\ \ 2 \ \ 1) = 2$$
$$b_2 = \quad a_0^{\ 1}(a_1 b_1 + a_2 b_0) = \quad ((\ \ 2) \ \ 2 + 3 \ \ 1) = 1$$
$$b_3 = \quad a_0^{\ 1}(a_1 b_2 + a_2 b_1 + a_3 b_0) = \quad ((\ \ 2) \ \ 1 + 3 \ \ 2 + 0 \ \ 1) = \quad 4$$

whereas Newton's algorithm gives us

$$g_0 = 1$$
$$g_1 = 1 \ \ (2 \quad f \ \ 1) \bmod x^2 = 1 + 2x$$
$$g_2 = (1 + 2x)(2 \quad f \ \ (1 + 2x)) \bmod x^4 = 1 + 2x + x^2 \quad 4x^3$$

(Note that the entire computation only depends on $f \bmod x^4$.)

Exercises.

1. Find the inverse power series of the polynomial $x^2 + x + 1$ in $Z_3[x]$. Try both algorithms and think about which one is better (theoretically, computationally) for this type of task.

2. Compute the quotient and remainder of the polynomials $x^4 + 2x^3 + 2$ and $x^3 \quad x$ in $Z[x]$ using algorithm 3.

3. Estimate the time complexity of the algorithms 3 and 4 expressed in the number of polynomial multiplications of degree $< n$.

12. Factorization of polynomials over finite fields

In this section, we are going to discuss algorithms to decompose polynomials into irreducible factors. Most such factorization algorithms assume that the input polynomial f is square-free, i.e. it is not divisible by any square of a polynomial. This is not a problem, as first we will first discuss an algorithm that decomposes a given polynomial into a product of square-free (but not necessarily irreducible) polynomials, with a time complexity of $O(n^3)$. In the second part, we will show Berlekamp's algorithm for factorizing polynomials from $F_q[x]$.

12.1. Square-free factorization.

Definition. A polynomial f is called square-free if there exists no nonconstant polynomial g such that $g^2 \mid f$. By a square-free decomposition of a polynomial f we mean pairwise distinct square-free polynomials $h_1, \ldots, h_k$ such that

$$f = h_1 \ \ h_2^2 \ \ h_3^3 \ \ \ldots \ \ h_k^k$$

(that is, h_i contains precisely those irreducible factors that occur in f to the i-th power).

Example. The square-free decomposition of

$$f = x^8 + x^7 \quad x^6 \quad x^5 \quad x^4 \quad x^3 + x^2 + x = (x^3 + x) \ \ (x \quad 1)^2 \ \ (x + 1)^3$$

in $Z[x]$ is given by $h_1 = x^3 + x$, $h_2 = x \quad 1$, $h_3 = x + 1$.

Let us recall that the characteristic of an integral domain is the smallest natural k satisfying $k \ \ 1 = 0$, if such k exists, and 0 otherwise. The finite field F_q, where $q = p^n$, has characteristic p.

The basic version of the algorithm for square-free factorization works for polynomials over any unique factorization domain (UFD) of characteristic 0 (and thus also for polynomials in more variables). In the case of a non-zero characteristic,

a minor problem occurs; we will demonstrate the resolution of this problem for polynomials in one variable over finite fields.

Lemma 12.1. Let f be a polynomial in $F_q[x]$, $q = p^n$, such that $f' = 0$. Then $f = g^p$ for some polynomial $g \in F_q[x]$.

We then denote the polynomial g by $\sqrt[p]{f}$.

Proof. If $f' = 0$, then all non-zero terms in f must have an exponent divisible by p, i.e. we can write $f = \sum a_i x^{ip}$. Let us define $g = \sum b_i x^i$ for b_i such that $b_i^p = a_i$ (we can choose $b_i = a_i^{p^{n-1}}$ because $b_i^p = a_i^{p^n} = a_i$ according to Proposition 9.3). Using the Proposition 9.1 we get $g^p = \sum b_i^p x^{ip} = f$.

The principle of the algorithm for square-free factorization is then described by the following theorem.

Theorem 12.2. Let $\mathbf{R}$ be a UFD of characteristic 0 or $\mathbf{R} = F_q$, and let f be a primitive polynomial in $\mathbf{R}[x]$. Then

(1) f is square-free if and only if $\gcd(f; f') = 1$.

(2) Let $f = \prod_{i=1}^{k} h_i^i$ be a square-free decomposition. Then

(a) if $\mathrm{char}(\mathbf{R}) = 0$, then $\gcd(f; f') = \prod_{i=1}^{k} h_i^{i-1}$;

(b) if $\mathbf{R} = F_q$, $q = p^n$, then $\gcd(f; f') = \prod_{p|i} h_i^i \prod_{p\nmid i} h_i^{i-1}$.

Proof. (1) ($\Rightarrow$) Let's assume that f is not square-free. Then $f = g^2 h$ for $g; h \in \mathbf{R}[x]$, and g non-constant. Thus $f' = 2gg'h + g^2h'$, so g is a common divisor of f and f'.

($\Leftarrow$) Let $f = \prod_{i=1}^{m} g_i$ be the decomposition of f into irreducible factors (i.e. $g_1; \ldots; g_m$ are pairwise non-associated irreducible polynomials). Then

$$f' = g_1' g_2 \cdots g_m + g_1 g_2' g_3 \cdots g_m + \cdots + g_1 \cdots g_{m-1} g_m'.$$

Let us assume that f and f' have some non-trivial common divisor. Then there is also some irreducible common divisor, and it is equal to some g_i. Since g_i divides f' and occurs in all summands in the above expression, except the i-th one, g_i must also divide the i-th term, i.e. $g_i \mid g_1 \cdots g_{i-1} g_i' g_{i+1} \cdots g_m$. Since $g_i \nmid g_j$ for any $j \neq i$, g_i must divide g_i'. But $\deg g_i' < \deg g_i$, and therefore $g_i' = 0$. In the case of characteristic 0, this means that g_i is constant, which is a contradiction. In the case of finite fields, there is the possibility that $g_i = g^p$ for some (non-constant) polynomial g, but this contradicts the irreducibility of g_i.

(2a) The derivative of the square-free expansion of f is

$$f' = \sum_{j=1}^{k} (h_j^j)' \prod_{i\neq j} h_i^i = \sum_{j=1}^{k} j h_j^{j-1} h_j' \prod_{i\neq j} h_i^i.$$

Note that $\prod_{i=1}^{k} h_i^{i-1}$ is a common divisor of f and f'. We are going to prove that this is the greatest common divisor. For this, assume that there exists a nonconstant polynomial g that divides both polynomials

$$\frac{f}{\prod_{i=1}^{k} h_i^{i-1}} = \prod_{i=1}^{k} h_i; \qquad \frac{f'}{\prod_{i=1}^{k} h_i^{i-1}} = \sum_{j=1}^{k} j \frac{h_j'}{h_j} \prod_{i\neq j} h_i.$$

Again, we can assume that the polynomial g is irreducible and thus that $g \mid h_m$ for some m. Similar to (1), since h_m occurs in all terms of the sum $\sum_{i=1}^{k} j \frac{h_j'}{h_j} \prod_{i\neq j} h_i$ except the m-th one, g must divide mh_m'. But the polynomial h_m is square-free, i.e. according to (1) it is coprime with h_m', a contradiction.

(2b) The derivation can be expressed as in (2a); because we are in characteristic p, all terms where $p \mid j$ are dropped, so we get

$$f' = \sum_{j=1}^{k} j\, h_j^{j-1} h_j' \prod_{i \neq j} h_i^i = \prod_{p \nmid i} h_i^i \left(\sum_{p \mid j} j\, h_j^{j-1} h_j' \prod_{p \mid i \neq j} h_i^i \right)$$

and we see that

$$\prod_{p \nmid i} h_i^i \prod_{p \mid i} h_i^{i-1}$$

is a common divisor of the polynomials f, f'. Similar to the previous case, it turns out to be the greatest common divisor.

12.1.1. Square-free factorization with characteristic 0.

For a given polynomial f, let

$$f_1 = \gcd(f, f'), \qquad g_1 = f/f_1$$

and further let us define inductively

$$g_{j+1} = \gcd(f_j, g_j), \qquad f_{j+1} = f_j/g_{j+1}.$$

In characteristic 0, then the following holds for $f = \prod_{i=1}^{k} h_i^i$:

j	f_j	g_j
1	$\prod_{i \geq 2} h_i^{i-1}$	$\prod_{i \geq 1} h_i$
2	$\prod_{i \geq 3} h_i^{i-2}$	$\prod_{i \geq 2} h_i$
3	$\prod_{i \geq 4} h_i^{i-3}$	$\prod_{i \geq 3} h_i$
	...	...
$k-1$	h_k	$h_{k-1} h_k$
k	1	h_k
$k+1$	1	1

Note that the length of the square-free decomposition, i.e. the value of k, is the step in which g_{k+1} turns into a constant value. In fact, much more can be seen from the table: the quotient g_j/g_{j+1} is equal to the sought factor h_j.

We remark that, since gcds are only defined up to associated elements, all polynomials in the table are defined up to $\sim$; so, by this procedure, we do not necessarily obtain the square-free decomposition of the polynomial f, but of some polynomial associated with f (this technical difficulty is of course easy to resolve).

Based on the above table, we can formulate an algorithm.

Algorithm 5 (Square-free factorization in Gaussian rings of characteristic 0).

Input: $f \in R[x]$ primitive, non-constant

Output: a square-free factorization $h_1, \ldots, h_k$ of a polynomial associated with f

1. $f_1 := \gcd(f, f')$, $g_1 := f/f_1$, $j := 1$
2. **while** $\deg g_j > 0$ **do**
   $g_{j+1} := \gcd(f_j, g_j)$, $f_{j+1} := f_j/g_{j+1}$, $h_j := g_j/g_{j+1}$
   $j := j + 1$
3. **return** $h_1, \ldots, h_{j-1}$

Proposition 12.3. The algorithm 5 is correct.

Proof. Let $f = \prod_{i=1}^{k} h_i^i$ be the input polynomial. It suffices to formally verify by induction that $f_j \sim \prod_{i \geq j+1} h_i^{i-j}$ and $g_j \sim \prod_{i \geq j} h_i$. It then follows that indeed $h_j \sim g_j/g_{j+1}$ and also that the algorithm stops for $j = k + 1$.

For $j = 1$, both assertions follow from Theorem 12.2. In an induction step we get

$$g_{j+1} \sim \gcd(f_j, g_j) = \gcd\Big(\prod_{i \geq j+1} h_i^{i-j}, \prod_{i \geq j} h_i \Big) = \prod_{i \geq j+1} h_i,$$

and hence $f_{j+1} = f_j / g_{j+1}\, k \prod_{i \quad j+1} h_i^{i \quad j} = \prod_{i \quad j+1} h_i = \prod_{i \quad j+2} h_i^{i \quad j \quad 1}$

Proposition 12.4. The time complexity of algorithm 5 is $O(nN(n))$, where $n = \deg f$ and $N(n)$ denotes the complexity of computing the gcd of two polynomials of degree n in $\mathbf{R}[x]$.

Proof. The algorithm computes $k+1$ values of $f_j; g_j$ and k values of h_j. Computing them requires computing a gcd and two divisions, with the divisions being less computationally demanding. So we can estimate the time complexity to be $(3k + 2)O(N(n)) = O(nN(n))$. (The estimate $k \quad n$ is the best possible: equality occurs if $h_1 = \quad = h_{n \quad 1} = 1$ and $\deg h_n = 1$.)

Example. Let us consider the polynomial

$$f = x^7 + x^6 \quad x^5 \quad x^4 \quad x^3 \quad x^2 + x + 1.$$

The algorithm 5 over the ring $Z[x]$ leads to

j	f_j	g_j	$h_{j \quad 1}$
1	$x^3 + x^2 \quad x \quad 1$	$x^4 \quad 1$	
2	$x + 1$	$x^2 \quad 1$	$x^2 + 1$
3	1	$x + 1$	$x \quad 1$
4	1	1	$x + 1$

Thus, the answer is $h_1 = x^2 + 1$, $h_2 = x \quad 1$, $h_3 = x + 1$, so

$$f = (x^2 + 1)(x \quad 1)^2(x + 1)^3.$$

Example. Let us consider the polynomial

$$f = (x + 1)^6.$$

The algorithm 5 over the ring $Z[x]$ then results in:

j	f_j	g_j	$h_{j \quad 1}$
1	$(x + 1)^5$	$x + 1$	
2	$(x + 1)^4$	$x + 1$	1
3	$(x + 1)^3$	$x + 1$	1
4	$(x + 1)^2$	$x + 1$	1
5	$x + 1$	$x + 1$	1
6	1	$x + 1$	1
7	1	1	$x + 1$

So $h_1 = \ldots = h_5 = 1$ and $h_6 = x + 1$.

12.1.2. Square-free factorization over finite fields. Let us now consider the same procedure over the field F_q. According to Theorem 12.2, we get the following values for the polynomial $f = \prod_{i=1}^{k} h_i^i$:

j	f_j	g_j
1	$\prod_{pji} h_i^i \quad \prod_{p\text{-}i \quad 2} h_i^{i \quad 1}$	$\prod_{p\text{-}i \quad 1} h_i$
2	$\prod_{pji} h_i^i \quad \prod_{p\text{-}i \quad 3} h_i^{i \quad 2}$	$\prod_{p\text{-}i \quad 2} h_i$
3	$\prod_{pji} h_i^i \quad \prod_{p\text{-}i \quad 4} h_i^{i \quad 3}$	$\prod_{p\text{-}i \quad 3} h_i$
	...	...
k	$\prod_{pji} h_i^i$	$\prod_{p\text{-}i \quad k} h_i$
$k+1$	$\prod_{pji} h_i^i$	1

The length of the square-free decomposition, i.e., the value of k, can again be determined by the fact that g_{k+1} turns out to be constant. There are two differences: in f_k we are left with the product of those square-free factors that are divisible by the power of p. The quotient g_j / g_{j+1} is equal to h_j if $p \nmid j$, otherwise it will be 1. In the process, we will find all square-free factors except the p-th, $2p$-th, etc.

At the end, it is enough to take the p-th root of the polynomial f_k and repeat the procedure.

Algorithm 6 (square-free factorization in finite fields).
Input: $f \in F_q[x]$ non-constant
Output: square-free factorization $h_1, \ldots, h_k$ of some polynomial associated with f
0. **if** $f' = 0$ **then goto 3.**
1. $f_1 := \gcd(f, f')$, $g_1 := f/f_1$, $j := 1$
2. **while** $\deg g_j > 0$ **do**
   $g_{j+1} := \gcd(f_j, g_j)$, $f_{j+1} := f_j/g_{j+1}$, $h_j := g_j/g_{j+1}$
   $j := j + 1$
   $f := f_j$
3. **if** $\deg f = 0$ **then return** $h_1, \ldots, h_{j-1}$
   else compute the square-free factorization $h_p, h_{2p}, \ldots, h_{lp}$ of the polynomial $\sqrt[p]{f}$,
   return $h_1, h_2, \ldots, h_{\max(j-1, lp)}$

The proof of correctness is analogous to that for the algorithm 5.

Example. Consider the input polynomial

$$f = x^7 + x^6 - x^5 - x^4 - x^3 - x^2 + x + 1.$$

The algorithm 6 in $Z_3[x]$ runs as follows:

j	f_j	g_j	h_{j-1}
1	$x^4 - x^3 + x - 1$	$x^3 - x^2 + x - 1$	
2	$x^3 + 1$	$x - 1$	$x^2 + 1$
3	$x^3 + 1$	1	$x - 1$

We get $h_1 = x^2 + 1$, $h_2 = x - 1$ and we are left with the polynomial $f = x^3 + 1$. The third root is $x + 1$, so we will perform a new calculation with this polynomial as input. we store the result in h_3. We are getting

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

Example. Consider the polynomial

$$f = x^6 + x^4 + x^2 + 1.$$

The algorithm 6 in $Z_2[x]$ runs as follows: since $f' = 0$, we will immediately consider the square root, i.e. the polynomial $x^3 + x^2 + x + 1$.

j	f_j	g_j	h_{j-1}
1	$x^2 + 1$	$x + 1$	
2	$x + 1$	$x + 1$	1
3	1	$x + 1$	1
4	1	1	$x + 1$

Hence $\sqrt[p]{f} = (x + 1)^3$, so the result is $f = (x + 1)^6$.

Exercises.

1. Perform a square-free factorization of the polynomial $x^7 + x^6 + x^4 + x^3 + x + 1$ in $Z_3[x]$.

2. Perform a square-free factorization of the polynomial $x^{10} + x^6 + x^5 + x^3 + x^2 + 1$ in $Z_2[x]$.

3. Perform a square-free factorization of the polynomial $x^{10} + 2x^9 + 2x^8 + 2x^7 + x^6 + x^5 + 2x^4 + x^3 + x^2 + 2x + 1$ in the ring $Z[x]$.

4. Carefully, prove the correctness of Algorithm 6.

12.2. **Berlekamp's algorithm.** The easiest factorization algorithm in $\mathbb{F}_q[x]$ is Berlekamp's algorithm. Its complexity with respect to the degree of the given polynomial is cubic, the disadvantage of its basic version is the exponential complexity with respect to $l(q)$, where $l(q)$ denotes the number of digits of the number q (ie $l(q) = (\log q)$). The input is a monic square-free polynomial. The main principle behind Berlekamp's algorithm is the following statement.

Proposition 12.5. Let f be a monic square-free polynomial in $\mathbb{F}_q[x]$ and consider a nonconstant polynomial $h \in \mathbb{F}_q[x]$ satisfying

$$h^q \equiv h \pmod{f}.$$

Then

$$f = \prod_{a \in \mathbb{F}_q} \gcd(f, h - a).$$

Proof. The polynomials $h - a$ are pairwise coprime (because $\gcd(h - a_1, h - a_2) = \gcd(h - a_1, a_1 - a_2) = 1$), hence also all polynomials $\gcd(f, h - a)$ are pairwise coprime. Since each of them divides the polynomial f, also the product $\prod_{a \in \mathbb{F}_q} \gcd(f, h - a)$ divides f. It remains to prove that also f divides the product.

Let $f = g_1 g_2 \cdots g_m$ be the decomposition of f into irreducible factors. It follows from the assumption that $f \mid h^q - h$. If we apply Proposition 9.4 by substituting the polynomial h for the variable x we get $f \mid \prod_{a \in \mathbb{F}_q}(h - a)$. Since the g_i are irreducible and $h - a$ are pairwise coprime, for each i there exists exactly one element $a \in \mathbb{F}_q$ such that $g_i \mid h - a$, and thus also $g_i \mid \gcd(f, h - a)$. Due to the square-freeness, the polynomials g_i are pairwise distinct, so $f = g_1 \cdots g_m \mid \prod_{a \in \mathbb{F}_q} \gcd(f, h - a)$.

So, we proved that f and $\prod_{a \in \mathbb{F}_q} \gcd(f, h - a)$ are associated. Since they are both monic, they are equal.

Proposition 12.5 provides a non-trivial decomposition of the polynomial f whenever we have a (non-constant) polynomial h of degree less than $\deg f$ satisfying $h^q \equiv h \pmod{f}$. The question is where to get such h. Let's define

$$W = \{h \in \mathbb{F}_q[x] : \deg h < \deg f,\ h^q \equiv h \pmod{f}\}.$$

The next proposition tells us more about the structure of this set.

Proposition 12.6. Let f be a square-free polynomial from $\mathbb{F}_q[x]$ with irreducible decomposition $f = g_1 g_2 \cdots g_m$. Then

(1) for every polynomial $h \in W$ and every i:

$$h \bmod g_i \in \mathbb{F}_q;$$

(2) the set W forms a vector space over $\mathbb{F}_q$ of dimension m and the map

$$\varphi : W \to (\mathbb{F}_q)^m, \qquad h \mapsto (h \bmod g_1, \ldots, h \bmod g_m)$$

is a vector space isomorphism.

Proof. (1) In the proof of Proposition 12.5, we saw that for every nonconstant polynomial $h \in W$ and every i there exists an element $a \in \mathbb{F}_q$ such that $g_i \mid h - a$. That is, $h \bmod g_i = a \in \mathbb{F}_q$. For constant polynomials the statement is trivial.

(2) The given map φ clearly preserves addition and scalar multiplication. If we can verify that it is a bijection, it follows that W is a vector space over $\mathbb{F}_q$ and φ is an isomorphism. Let us choose an arbitrary vector $(a_1, \ldots, a_m) \in \mathbb{F}_q^m$ and consider the set of elements h satisfying the congruences $h \equiv a_i \pmod{g_i}$, $i = 1, \ldots, m$. Since g_i are pairwise coprime (due to square-freeness), the Chinese remainder theorem guarantees exactly one solution h modulo $g_1 \cdots g_m = f$ (i.e. exactly one of degree $< \deg f$). At the same time, according to Proposition 9.3

$$h^q - a_i^q = a_i - h \pmod{g_i};$$

and thus, thanks to the coprimeness of the g_i, $h^q \quad h$ modulo $g_1 \quad g_m = f$ holds. Thus, the map $'$ is a bijection, and the only preimage of the vector $(a_1;:::;a_m)$ is the polynomial h.

It remains the question of how to actually nd some non-trivial elements of W. Let us denote $n = \deg f$. Consider the polynomial

$$h = a_0 + a_1x + \quad + a_{n\ 1}x^{n\ 1} \ 2\ \mathrm{F}_q[x]:$$

We are interested in nding h with $h^q \bmod f = h$. From Proposition 9.1 and 9.3 we see that

$$h^q = \left(\sum_{i=0}^{n\ 1} a_i x^i\right)^q = \sum_{i=0}^{n\ 1} \left(a_i x^i\right)^q = \sum_{i=0}^{n\ 1} a_i^q x^{iq} = \sum_{i=0}^{n\ 1} a_i x^{iq}:$$

If we denote the coe cients of the polynomial $x^{jq} \bmod f$ by $q_{i;j}$, then

$$\begin{aligned}
1 = x^0 \bmod f &= q_{0;0} + q_{1;0}x + \quad + q_{n\ 1;0}x^{n\ 1} \\
x^q \bmod f &= q_{0;1} + q_{1;1}x + \quad + q_{n\ 1;1}x^{n\ 1} \\
&::: \qquad :::\\
x^{(n\ 1)q} \bmod f &= q_{0;n\ 1} + q_{1;n\ 1}x + \quad + q_{n\ 1;n\ 1}x^{n\ 1};
\end{aligned}$$

so

$$\begin{aligned}
h^q \bmod f &= \left(\sum_{j=0}^{n\ 1} a_j x^{jq}\right) \bmod f = \sum_{j=0}^{n\ 1} a_j (x^{jq} \bmod f) = \\
&= \sum_{j=0}^{n\ 1} \left(a_j \sum_{i=0}^{n\ 1} q_{i;j} x^i\right) = \sum_{i=0}^{n\ 1} \left(\sum_{j=0}^{n\ 1} q_{i;j} a_j\right) x^i:
\end{aligned}$$

If we name the coe cients of $h^q \bmod f = b_0 + b_1x + \quad + b_{n\ 1}x^{n\ 1}$, then we can write in matrix-form

$$\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n\ 1} \end{pmatrix} = Q \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n\ 1} \end{pmatrix}; \quad \text{where } Q = \begin{pmatrix} q_{0;0} & q_{0;1} & ::: & q_{0;n\ 1} \\ q_{1;0} & q_{1;1} & ::: & q_{1;n\ 1} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n\ 1;0} & q_{n\ 1;1} & ::: & q_{n\ 1;n\ 1} \end{pmatrix}$$

Thus, the equality $h^q \bmod f = h$ holds if and only if

$$Q \quad (a_0;:::a_{n\ 1})^T = (a_0;:::;a_{n\ 1})^T;$$

so if and only if

$$(Q \quad E) \quad (a_0;:::;a_{n\ 1})^T = (0;0;:::;0)^T$$

(here E denotes the identity matrix). We derive the following proposition

Proposition 12.7. Let Q be an $n \quad n$ matrix whose columns are the coe cients of polynomials

$$1;\ x^q \bmod f;\ x^{2q} \bmod f;\ :::;\ x^{(n\ 1)q} \bmod f:$$

Then the polynomial $h = a_0 + a_1x + ::: + a_{n\ 1}x^{n\ 1}$ lies in W if and only if $(a_0;:::;a_{n\ 1})$ is a solution of a homogeneous system of linear equations, given by the matrix $Q \quad E$.

According to Proposition 12.6, the dimension of the vector space W is equal to the number of irreducible factors of the polynomial f. On the other hand, it is equal to the dimension of the solution space of the given system, i.e. the value $n \quad \mathrm{rk}(Q \quad E)$, where rk denotes the rank of the matrix. Note that the rst column $Q \quad E$ matrix is zero, so the vectors $(a;0;0;:::;0)$ are always solutions; these vectors

correspond to constant polynomials in W, which we are not interested in because they do not provide a nontrivial decomposition.

The core idea of Berlekamp's algorithm is now as follows:

(1) We solve the system of equations with the matrix $Q - E$ by Gaussian elimination. If the dimension of the solution space is 1, the polynomial f is irreducible. Otherwise, we take any solution of the given system di erent from $(a; 0; 0; \ldots; 0)$ and denote the corresponding polynomial by h.
(2) From the formula in Proposition 12.5 we obtain a non-trivial decomposition $f = g_1 g_2 \ldots g_l$.
(3) If l is equal to the dimension of the solution space, we are done. Otherwise, we continue recursively for each polynomial $g_1; g_2; \ldots; g_l$.

We remark that this procedure can be noticeably optimized: instead of randomly choosing h in step 1, we can compute a basis of the solution space $h_1 = 1; h_2; \ldots; h_m$ (the polynomial $h_1 = 1$ and its multiples correspond to the irrelevant constant polynomials). We then can use h_2 to nd a non-trivial decomposition of f, and nd further decompositions using polynomial h_3, etc. (Here we apply Proposition 12.5 to the individual factors, the assumption $h^q \equiv h \pmod{f'}$ holds for any $f' \mid f$.) We proceed in this way until we nd m non-trivial factors. It is, of course, necessary to show that the stated procedure leads to the goal.

In the description of the algorithm, we identify polynomials of degree $< n$ with the elements of $(\mathbb{F}_q)^n$. The variable F will, in each step, contain a decomposition of the polynomial f, which is re ned in step 4.

Algorithm 7 (Berlekamp's algorithm).
Input: $f \in \mathbb{F}_q[x]$ square-free, monic, of degree n
Output: decomposition of f into irreducible polynomials $g_1; \ldots; g_m$ in $\mathbb{F}_q[x]$
1. $Q :=$ matrix with columns $x^0 \bmod f$, $x^q \bmod f$, ..., $x^{(n-1)q} \bmod f$
2. compute a base $h_1 = 1; h_2; \ldots; h_m$ of the solution space of $(Q - E)h = 0$
3. $i := 2$, $F := \{f\}$
4. **while** $|F| < m$ **do**
   replace each $g \in F$ by non-trivial factors from the decomposition
   $$g = \prod_{a \in \mathbb{F}_q} \gcd(g; h_i - a)$$
   $i := i + 1$
5. **return** F

Proposition 12.8. The algorithm 7 is correct.

Proof. To prove the correctness, it remains to show that every two pairwise distinct factors of the polynomial f can be separated by some polynomial h_k, i.e. that for every $i; j$ there exists an index k and di erent elements $a; b \in \mathbb{F}_q$ such that $g_i \mid \gcd(f; h_k - a)$ and $g_j \mid \gcd(f; h_k - b)$. Since $g_i; g_j \mid f$, it su ces to search for $k; a; b$ such that $g_i \mid h_k - a$ and $g_j \mid h_k - b$, i.e. such that $h_k \bmod g_i = a$ and $h_k \bmod g_j = b$. In other words, for each $i; j$ we look for k such that $h_k \bmod g_i \neq h_k \bmod g_j$. The existence of such k follows easily from Proposition 12.6: the vectors $\varphi(h_1); \ldots; \varphi(h_m)$ form the basis of the space $(\mathbb{F}_q)^m$, so it is not possible for them to have all the same i-th and j-th component { recall that $\varphi(h) = (\ldots; h \bmod g_i; \ldots; h \bmod g_j; \ldots)$.

Example. Consider the polynomial

$$f = x^4 + 1 \in \mathbb{Z}_3[x]:$$

Since $\gcd(f; f^0) = 1$, it is square-free, so we can use Berlekamp's algorithm. First, we compute the Q matrix. It holds that

$$\begin{aligned} x^0 \bmod f &= 1; \\ x^3 \bmod f &= x^3; \\ x^6 \bmod f &= 2x^2; \\ x^9 \bmod f &= x; \end{aligned}$$

thus

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}:$$

By Gaussian elimination we get the row echelon form

$$Q \quad E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}:$$

One possible basis of the solution space is $(1; 0; 0; 0); (0; 1; 0; 1)$, which corresponds to the polynomials $h_1 = 1$, $h_2 = x + x^3$. Thus, the polynomial f decomposes into 2 irreducible factors. We compute

$$\begin{aligned} \gcd(f; h_2 \quad 0) &= \gcd(x^4 + 1; x^3 + x) = 1; \\ \gcd(f; h_2 \quad 1) &= \gcd(x^4 + 1; x^3 + x + 2) = x^2 + 2x + 2; \\ \gcd(f; h_2 \quad 2) &= \gcd(x^4 + 1; x^3 + x + 1) = x^2 + x + 2: \end{aligned}$$

Thus, we obtain the decomposition

$$x^4 + 1 = (x^2 + 2x + 2)(x^2 + x + 2):$$

Proposition 12.9. The time complexity of Algorithm 7 in the field F_q is $O(n^3 q l(q)^2)$, where $l(q)$ denotes the number of digits of the number q.

Proof. Computing $x^q \bmod f$ by a binary algorithm has complexity $O(n^2 l(q)^3)$. In step 1, when calculating the Q matrix, we use the relation

$$x^{iq} \bmod f = (x^{(i \ 1)q} \bmod f) \ \ x^q \bmod f;$$

so we need n divisions with remainder in $\mathrm{F}_q[x]$, dividing the x^q-multiple of the previous polynomial (degrees $< 2n$) by the polynomial f of degree n. Therefore, Step 1 has running time $O(n^3 l(q)^3)$. The time complexity of the Gaussian elimination of the matrix $Q \quad E$ of size $n \quad n$ over the field F_q is $O(n^3 l(q)^2)$. We go through the loop in step 4 at most m times, with $m \quad n$. In a given loop, for each element $g \ 2 \ F$, we compute $\gcd(g; h_i \quad a)$ for each $a \ 2 \ \mathrm{F}_q$, which is a total of q gcd computations with each g. The total complexity of step 4. is therefore $m \ \ q \ \ \sum_{g2F} O(n \deg g l(q)^2) = O(n^2 q l(q)^2 (\sum_{g2F} \deg g)) = O(n^3 q l(q)^2)$.

Berlekamp's algorithm has exponential complexity with respect to the length of the number q, because in step 4 we go through all elements a of the field F_q. At the same time, the other steps have a complexity of only $O(n^3 l(q)^3)$, so this is the bottleneck of this procedure.

No deterministic algorithm is know that is of polynomial complexity with respect to both n and $l(q)$. However, there are a number of probabilistic algorithms (e.g. Cantor-Zassenhaus or Kaltofen-Shoup) that, on average, compute a factorization in polynomial time. The first such algorithm was discovered by Berlekamp himself.

Exercises.

1. Decompose the polynomial $x^7 + 2x^5 + 2x^4 + x^3 + 2x + 2$ into a product of irreducibles in $Z_3[x]$.

2. Decompose the polynomial $x^7 + 4x^6 + 2x^5 + 4x^3 + 3x^2 + 4x + 2$ into a product of irreducibles in $Z_5[x]$. (Don't forget to do the square-free factorization first.)

3. Based on Berlekamp's algorithm, design an algorithm for testing the irreducibility of a polynomial $f \in F_q[x]$ and calculate its time complexity.

13. General algebraic structures

13.1. **Algebraic structures.** So far, the reader has become familiar with three classical algebraic theories: linear algebra, commutative algebra, and group theory. Each of these disciplines studies a speci c type of algebraic structures, namely vector spaces, commutative rings, and groups. Every such algebraic structure can be described as a set on which some operations are de ned (e.g. a ring $\mathbf{R}$, is given by the carrier set R, and the operations $+$, , , 0). Each theory then imposes some conditions on these operations, so-called axioms, which are based on properties that are shared by key examples (e.g., in commutative algebra, the ring axioms are satis ed by the integers Z, or polynomial rings $\mathbf{R}[x]$). We generalize these observations in the following de nition of algebraic structures.

Definition. By a language (or signature) we mean the set together with a mapping ar : ! N [$f0g$. The meaning of this de nition is as follows: is the set of operational symbols that we will use in a given theory, and the function ar assigns an arity to each symbol. We say that the symbol 2 is $ar(\ )$-ary. Instead of 1-ary we say unary, instead of 2-ary we say binary. As a rule, in x characters $+; \ ; \ ;$, etc. are only used for binary symbols, post x characters $^0; \ ^1$ etc. are sometimes used for unary symbols;

Let A be a set. By n-ary operation on A we mean a map from the Cartesian power $A^n = A \ ::: \ A$ to A. Speci cally, a 0-ary operation is a mapping from a one-element set to A, so it can be interpreted as a constant.

An algebraic structure in the language is the pair $\mathbf{A} = (A; \ )$, where A is a non-empty set, called the carrier set (also universe, or domain of $\mathbf{A}$), and is a map from to the set of operations on A, that assigns a $ar(\ )$-ary operation $^{\mathbf{A}}$ to every symbol .

Example. Groups are algebraic structures $\mathbf{G} = (G; \ )$ in the language $= f \ ;^0; eg$, where

$$ar(\ ) = 2; \ ar(^0) = 1; \ ar(e) = 0;$$

satisfying the following identities for all $a; b; c \ 2 \ G$:

$$a \ ^{\mathbf{G}} (b \ ^{\mathbf{G}} c) = (a \ ^{\mathbf{G}} b) \ ^{\mathbf{G}} c;$$

$$a \ ^{\mathbf{G}} e^{\mathbf{G}} = e^{\mathbf{G}} \ ^{\mathbf{G}} a = a;$$

$$a \ ^{\mathbf{G}} a^{0\mathbf{G}} = a^{0\mathbf{G}} \ ^{\mathbf{G}} a = e^{\mathbf{G}};$$

Example. Rings with unity are algebraic structures $\mathbf{R} = (R; \ )$ in the language $= f+; \ ; \ ;0;1g$, where $ar(+) = ar(\ ) = 2$, $ar(\ ) = 1$, $ar(0) = ar(1) = 0$, such that $(R; +^{\mathbf{R}}; \ ^{\mathbf{R}}; 0^{\mathbf{R}})$ is an Abelian group, $^{\mathbf{R}}$ is asociative, the left and right associativity laws hold for $+^{\mathbf{R}}; \ ^{\mathbf{R}}$, and $a \ ^{\mathbf{R}} 1^{\mathbf{R}} = 1^{\mathbf{R}} \ ^{\mathbf{R}} a = a$ for every $a \ 2 \ R$.

If it is clear from the context, whether we are talking about an operation symbol (e.g. $+$) or an operation of an algebra (e.g. $+^{\mathbf{A}}: A^2 \ ! \ A$), we will omit the superscript.

Example. A Latin square $(a_{i;j})_{i;j \ 2 \ X}$ over a set X can be considered as an algebraic structure $(X; \ )$ with one binary operation, where $u \ v = a_{u;v}$. Algebraic structures formed from Latin squares are called quasigroups.

The following two examples show some pitfalls in the structure de nition.

Example. In a eld $\mathbf{T}$ on a set T, forming the inverse $a \mapsto a^{-1}$ is not an operation on T, as it is not de ned on 0. Thus, elds cannot be regarded as algebraic structure in the language $f+; \ ; \ ^{-1}; 0; 1g$. We can, however, regard them as algebraic structures in the language of rings with unity $f+; \ ; \ ; 0; 1g$, by de ning them as those commutative rings with unity in which, for every $0 \neq a \in T$ there is exactly one $b \in T$ such that $a \ b = 1$.

Example. Vector spaces over the eld $\mathbf{T}$ can be considered as algebraic structures in the language $f+; \ ; 0g \cup ff \ : \ \in Tg$, where $ar(+) = 2$, $ar(\ ) = 1$, $ar(0) = 0$ and $ar(f \ ) = 1$ for all $\in T$, and $(V; +; \ ; 0)$ is an Abelian group and for all $a; b \in V$, $\ ; \ \in T$,

$$f_{(\ +_T \ )}(a) = f \ (a) + f \ (a); \quad f \ _T \ (a) = f \ (f \ (a));$$
$$f \ (a + b) = f \ (a) + f \ (b); \quad f_1(a) = a:$$

We interpret the symbols f as scalar multiplication by the element , i.e. $f \ (v) = v$. The scalar product in the traditional sense (so, as a map $T \ V \to V$) is then not an operation of this algebraic structure.

With lattices and Boolean algebras will see some more examples of interesting classes of algebraic structures in Section 14.

In the algebraic theories mentioned above, some basic constructions recurrently appear: substructures, direct products, homomorphisms, and quotients. We will next give a common framework for these concepts.

13.2. **Substructures.**

Definition. Let f be an n-ary operation on the set A and $B \ A$. We say that a subset B is closed under the operation f if for all $b_1; : : : ; b_n \in B$

$$f(b_1; : : : ; b_n) \in B:$$

Let $\mathbf{A} = (A; \ )$ be a structure in the language . By a substructure (or subalgebra) of $\mathbf{A}$ we mean the algebraic structure $\mathbf{B} = (B; \ )$ in the same language, where $B \ A$ is closed under all operations from and contains restrictions on operations from to the set B, i.e., $^{\mathbf{B}} = {}^{\mathbf{A}}|_B$ for all $\in$. We denote by $\mathbf{B} \ \mathbf{A}$.

This de nition is compatible with de nitions of substructures you already know: a substructure of a vector space is the same as a subspace, a substructure of a ring is a subring, a substructure of a group is a subgroup.

Example. The operations of the given structure matter: the natural numbers N form a substructure of the structure $(Z; +)$, but not of $(Z; \ )$.

Proposition 13.1 (intersection of substructures)**.** Let $\mathbf{A}$ be an algebraic structure and let $\mathbf{B}_i$, $i \in I$, be a family of substructures. Then $\bigcap_{i \in I} B_i$ is either the empty set or also a substructure of $\mathbf{A}$.

If the intersection is non-empty we denote the corresponding substructure by $\bigcap_{i \in I} \mathbf{B}_i$.

Proof. Let $B = \bigcap_{i \in I} B_i$ and let us assume that $B \neq ;$. Let be a symbol of arity n and $b_1; : : : ; b_n \in B$. Then $b_1; : : : ; b_n \in B_i$ for all $i \in I$, so $^{\mathbf{A}}(b_1; : : : ; b_n) \in B_i$ for all $i \in I$, since every set B_i is closed under this operation. Thus $^{\mathbf{A}}(b_1; : : : ; b_n) \in \bigcap_{i \in I} B_i = B$.

Next, consider a subset $; \subseteq X \quad A$ of the carrier set of $\mathbf{A}$. We de ne the substructure generated by the set X to be the smallest substructure (with respect to inclusion) of $\mathbf{A}$ containing X. For short, we write $hX\,i_{\mathbf{A}}$. Such a substructure always exists: just take the intersection of all substructures containing X, i.e.,

$$hX\,i_{\mathbf{A}} = \bigcap_{X \quad B;\ \mathbf{B} \quad \mathbf{A}} \mathbf{B}:$$

By Proposition 13.1, this intersection is also a substructure; clearly it contains the set X, and is the smallest among all such substructures.

We can nd the elements of the substructure $hX\,i_{\mathbf{A}}$ by starting with the elements of the set X and by applying the operations of the structure $\mathbf{A}$ we obtain other elements. If no more new elements arise, that is, when the resulting subset is closed to all operations of the structure $\mathbf{A}$, we have found the substructure.

Example. For a given $n \in \mathbb{Z}$:

$$hni_{(\mathbb{Z};+)} = \{kn : k \in \mathbb{N}\},$$
$$hni_{(\mathbb{Z};\ )} = \{n^k : k \in \mathbb{N}\}.$$

Exercises.

1. List all substructures of the structure $\mathbf{A} = (\{1;2;3;a;b;c\};f)$ with one unary operation de ned $f(a) = f(1) = 2$, $f(b) = f(2) = 3$, $f(c) = f(3) = 1$. Draw an ordered set of substructures, with the inclusion order of support sets.

2. Find all substructures of the structure $(\mathbb{Z};f)$ with one unary operation de ned by $f(k) = k + 1$.

3. Let $\mathbf{A} = (A;\ )$ be a structure with one binary operation. Prove that the set

$$\{a \in A : (x\ \ a)\ \ y = x\ \ (a\ \ y) \text{ for all } x;y \in A\}$$

is either empty or contains a substructure of the structure $\mathbf{A}$. Give an example of a structure in which this set is empty.

4. Describe the algebraic structures $h2i_{(\mathbb{Z};\ )}$ and $h2i_{(\mathbb{Q}\setminus\{0\};:)}$.

Let $\mathbf{T}_3 = (T_3;\ )$ be a structure consisting of all $\{1;2;3\} \to \{1;2;3\}$, together with the composition operation . Verify that $\mathbf{T}_3$ is generated by the permutations (1 2 3), (1 2) and the mapping $1 \mapsto 1; 2 \mapsto 2; 3 \mapsto 1$. Prove that $\mathbf{T}_3$ is not generated by any two-element set.

5. Prove that every substructure of $(\mathbb{N};+)$ is generated by nitely many elements.

13.3. Homomorphisms and isomorphisms.

Definition. Let $\mathbf{A}, \mathbf{B}$ be algebraic structures in the same language . A mapping $' : A \to B$ is called a homomorphism between $\mathbf{A}$ and $\mathbf{B}$ if

$$'(\ ^{\mathbf{A}}(a_1;\dots;a_n)) = \ ^{\mathbf{B}}('(a_1);\dots;'(a_n))$$

for each n-ary symbol $\in$ and all $a_1;\dots;a_n \in A$. We say that $'$ preserves the operations of these algebras. We write $' : \mathbf{A} \to \mathbf{B}$.

The following terminology is used for special types of homomorphisms:

An embedding is an injective homomorphism (sometimes the notation $\mathbf{A} \hookrightarrow \mathbf{B}$ is used),

An isomorphism is a bijective homomorphism (notation $\mathbf{A} \simeq \mathbf{B}$),

and further

an endomorphism of the structure $\mathbf{A}$ is a homomorphism $\mathbf{A} \to \mathbf{A}$,

an automorphism of the structure $\mathbf{A}$ is an isomorphism $\mathbf{A} \to \mathbf{A}$.

Note that the identity map $id : \mathbf{A} \to \mathbf{A}$, $x \mapsto x$, is always an automorphism.

The image of a given homomorphism $' : \mathbf{A} \to \mathbf{B}$ is de ned by $\mathrm{Im}(') = f'(a) : a \in A g$. The image always forms a substructure of the structure $\mathbf{B}$: if is an n-ary symbol and $b_1;:::;b_n \in \mathrm{Im}(')$, then $b_1 = '(a_1);:::;b_n = '(a_n)$ for some $a_1;:::;a_n \in A$ and holds

$$^{\mathbf{B}}(b_1;:::;b_n) = {}^{\mathbf{B}}('(a_1);:::;'(a_n)) = '({}^{\mathbf{A}}(a_1;:::;a_n)) \in \mathrm{Im}(').$$

Homomorphisms are uniquely determined by their values on generators. (But it is not true that a given map on generators can always be extended to a homomorphism: this is a speci c feature of e.g. vector spaces, or, in general, so-called free algebras.) This principle can be used to nd all homomorphisms between two structures.

Exercise. Find all homomorphisms $(\mathbb{Z};) \to (\mathbb{Z};)$ (for the binary operation).

Solution. Note that $(\mathbb{Z};) = \langle 1 \rangle$, so all other values can be computed from the value in 1. Consider the homomorphism $'$. If $'(1) = k$, we can prove by induction that $'(a) = ka$ for all $a \geq 0$. Let's rst observe $'(0) = '(1 \ 1) = '(1) \ '(1) = k \ k = 0$ and $'(\ 1) = '(0 \ 1) = '(0) \ '(1) = \ k$. In the induction step, note that $'(a) = '((a \ 1) \ (\ 1)) = '((a \ 1)) \ '(\ 1) = (a \ 1)k \ (\ k) = ak$ holds; we can proceed similarly for negative a. It remains to verify that we have indeed obtained a homomorphism: $'(a \ b) = k(a \ b) = ka \ kb = '(a) \ '(b)$ for all $a; b$.

If we are not able to e ectively make use of generating sets, we can also try to use elements with special properties that are preserved by each homomorphism (see also the discussion of invariants in Section 1.3):

Exercise. Find all homomorphisms $(\mathbb{Z};) \to (\mathbb{Z}; +)$.

Solution. Consider a homomorphism $' : (\mathbb{Z};) \to (\mathbb{Z}; +)$. From the equality $'(0) = '(0 \ 0) = '(0) + '(0)$ it follows that $'(0) = 0$ and we get $0 = '(0) = '(n \ 0) = '(n) + '(0) = '(n)$ for each $n \in \mathbb{Z}$. Thus, there exists only a single homomorphism, the map $n \mapsto 0$.

The following properties can be proven similar to the group and ring case:

Proposition 13.2. Let $\mathbf{A}; \mathbf{B}; \mathbf{C}$ be algebraic structures in the same language and let $' : \mathbf{A} \to \mathbf{B}$ and $: \mathbf{B} \to \mathbf{C}$ be homomorphisms. Then

(1) the composition $'$ is a homomorphism $\mathbf{A} \to \mathbf{C}$;
(2) if $'$ is an isomorphism, then the inverse map $'^{1}$ is an isomorphism of $\mathbf{B} \to \mathbf{A}$.

It follows from Proposition 13.2 that the automorphisms of a given structure $\mathbf{A}$ form a subgroup of the symmetric group $\mathbf{S}_A$, denoted by $\mathbf{Aut}(\mathbf{A})$.

We say that the structures $\mathbf{A}$ and $\mathbf{B}$ are isomorphic, we denote $\mathbf{A} \simeq \mathbf{B}$ if there is an isomorphism $\mathbf{A} \to \mathbf{B}$. As with groups and rings, isomorphism can be thought of as copying operations from one support set to another, i.e., two algebraic structures are isomorphic if they di er only by renaming elements. It follows from Propositon 13.2 that isomorphisms induce an equivalence relation on the class of all algebras in the given language.

Recall that an isomorphism invariant is a property V such that whenever $\mathbf{A}$ has the property V and $\mathbf{B} \simeq \mathbf{A}$, then $\mathbf{B}$ also has the property V. In Section 1.3, we discussed several such invariants for groups (minimum number of generators, equalities, certain types of special elements). In general, an invariant is any property that can be expressed using so-called rst-order formulas in a given language, i.e. expressions only using quanti ers, variables, logical conjunctions, equality, and

operations from the given language. The exact formulation and proof can be found in most textbooks on mathematical logic.

Exercises.

1. Let us denote $N_0 = N \setminus \{0\}$. Find all homomorphisms $(N_0; +) \to (N;)$ and $(N;) \to (N_0; +)$.

2. Let $\mathbf{T}_n = (T_n;)$ be the structure consisting of all mappings on the set $\{1; \ldots; n\}$, together with the composition . Decide whether there is a homomorphism $\mathbf{T}_n \to (N; +)$ or $\mathbf{T}_n \to (\{0; 1\};)$.

3. Find all homomorphisms $\mathbf{A} \to \mathbf{B}$ where $\mathbf{A} = (\{a; b; c; d\}; f)$, $f(a) = f(b) = c$, $f(c) = f(d) = d$ and $\mathbf{B} = (\{0; 1\}; g)$, $g(0) = g(1) = 1$.

4. Decide which of the following structures are isomorphic: $(N;)$, $(2N;)$, $(3N;)$, $(N \setminus \{2N;)$, $(Q^+;)$.

5. Prove Proposition 13.2.

13.4. **Congruences and quotient structures.**

In the beginning of Section 2, we investigated the idea of the constructing quotient object, by identifying closely related elements. This idea appears throughout mathematics. In the following, we discuss how we can apply it to algebraic structures.

Let $\mathbf{A} = (A;)$ be an algebraic structure in the language . Consider an equivalence on the set A, which will tell us which elements we want to identify. We would like to define the operations of the factor structure $\mathbf{A}/$ in such a way that the result of the n-ary operation $^{\mathbf{A}/}$ on the blocks $[a_1]; \ldots; [a_n]$ should be equal to the block $[\ ^{\mathbf{A}}(a_1; \ldots; a_n)]$. However, in order for such an operation to be well defined, the equivalence cannot be arbitrary. The equivalences for which the construction works are called congruences.

Definition. Let $\mathbf{A}$ be a structure in the language . An equivalence on the support A is called a congruence of $\mathbf{A}$ if for every n-ary symbol 2 and all $a_1; \ldots; a_n; b_1; \ldots; b_n \ 2 \ A$ holds

$$a_1 \quad b_1; \ \ldots; \ a_n \quad b_n \quad) \quad ^{\mathbf{A}}(a_1; \ldots; a_n) \quad ^{\mathbf{A}}(b_1; \ldots; b_n):$$

For a unary symbol ^{0}, this condition says that if $a \quad b$, then $a^0 \quad b^0$. For a binary symbol the condition says

$$a \quad b; \ c \quad d \quad) \quad a \quad c \quad b \quad d;$$

which, as the reader will easily deduce, is equivalent to the condition $a \quad b \ ) \ a \quad c \quad b \quad c$ and $c \quad a \quad c \quad b$ for all c. Constants (i.e. 0-ary symbols) play no role in the definition, since $c \quad c$ for every c.

Example. Consider the structure $(Z; +; \ ; \ )$. In Algebra 1, we proved that the relation $\pmod n$ is a congruence of it.

Caution: Unlike for groups or rings, where congruence are already completely determined by one equivalence class that additionally forms a subalgebra (the ideal $[0]$ in rings and the normal subgroup $[e]$ in groups), this does not hold in general!

Definition. Let $\mathbf{A}$ be an algebraic structure and a congruence. Consider the set $A/ = \{[a] : a \ 2 \ A\}$ and define operations on by

$$^{\mathbf{A}/} \ ([a_1] \ ; \ldots; [a_n] \ ) = [\ ^{\mathbf{A}}(a_1; \ldots; a_n)]$$

for each n-ary symbol 2 and all $a_1; \ldots; a_n \ 2 \ A$. From the definition of congruence, we see that the operations are well defined: if we label the blocks in a different way, i.e. if $[a_1] = [b_1]; \ldots; [a_n] = [b_n]$, then $[\ ^{\mathbf{A}}(a_1; \ldots; a_n)] = [\ ^{\mathbf{A}}(b_1; \ldots; b_n)]$, i.e.

the result of the operation is independent of the representant. The algebraic structure

$$\mathbf{A}/\theta = (A/\theta; (f^{\mathbf{A}/\theta} : f \in \Omega))$$

is called the quotient structure of $\mathbf{A}$ by the congruence θ.

Let $\varphi : \mathbf{A} \to \mathbf{B}$ be a homomorphism. By its kernel we mean the relation on A defined

$$a \ker(\varphi) b \iff \varphi(a) = \varphi(b).$$

The following statement says that the kernel is a congruence of the structure $\mathbf{A}$, and that every congruence is the kernel of some homomorphism.

Proposition 13.3 (kernels vs. congruences)**.** Let $\mathbf{A}$ be a structure and θ be a relation on its support set A. Then θ is a congruence of the structure $\mathbf{A}$ if and only if it is the kernel of some homomorphism from $\mathbf{A}$ to some structure $\mathbf{B}$.

Proof. ($\Leftarrow$) Consider the map

$$\varphi : A \to A/\theta; \quad a \mapsto [a].$$

It immediately follows from the definition of quotient structure that it is a homomorphism $\mathbf{A} \to \mathbf{A}/\theta$. Its kernel consists of those pairs (a, b) for which $[a] = [b]$, i.e. $a \theta b$.

($\Rightarrow$) Consider some homomorphism $\varphi : \mathbf{A} \to \mathbf{B}$. Its kernel is an equivalence relation, so we only have to prove that it is even a congruence. Consider an n-ary symbol f. Let $a_1, \ldots, a_n, b_1, \ldots, b_n \in A$ be such that $a_i \ker(\varphi) b_i$, i.e. $\varphi(a_i) = \varphi(b_i)$, for all i. Then

$$\begin{aligned}\varphi(f^{\mathbf{A}}(a_1, \ldots, a_n)) &= f^{\mathbf{B}}(\varphi(a_1), \ldots, \varphi(a_n)) \\ &= f^{\mathbf{B}}(\varphi(b_1), \ldots, \varphi(b_n)) = \varphi(f^{\mathbf{A}}(b_1, \ldots, b_n));\end{aligned}$$

and thus $f^{\mathbf{A}}(a_1, \ldots, a_n) \ker(\varphi) f^{\mathbf{A}}(b_1, \ldots, b_n)$.

Similar to groups and rings, the homomorphism theorem and the 1st isomorphism theorem apply.

Theorem 13.4 (homomorphism theorem)**.** Let $\varphi : \mathbf{A} \to \mathbf{B}$ be a homomorphism between algebras.

(1) If θ is a congruence of the structure $\mathbf{A}$ that is a subset of the kernel $\ker(\varphi)$, then the map

$$\psi : \mathbf{A}/\theta \to \mathbf{B}; \quad [a] \mapsto \varphi(a)$$

is well defined and a homomorphism.

(2) (1st isomorphism theorem) $\mathbf{A}/\ker(\varphi) \cong \mathbf{Im}(\varphi)$.

Proof. (1) First, it is necessary to verify that the mapping ψ is well defined: if $[a] = [b]$, i.e. if $a \theta b$, then $a \ker(\varphi) b$, and thus $\varphi(a) = \varphi(b)$. It is a homomorphism because for the n-ary symbol f and every $a_1, \ldots, a_n \in A$

$$\begin{aligned}f^{\mathbf{A}/\theta}([a_1], \ldots, [a_n]) &= [f^{\mathbf{A}}(a_1, \ldots, a_n)] = \varphi(f^{\mathbf{A}}(a_1, \ldots, a_n)) \\ &= f^{\mathbf{B}}(\varphi(a_1), \ldots, \varphi(a_n)) = f^{\mathbf{B}}(\psi([a_1]), \ldots, \psi([a_n])).\end{aligned}$$

(2) We apply (1) to the congruence $\ker(\varphi)$ itself: the resulting homomorphism ψ is injective, since $a \ker(\varphi) b \iff \varphi(a) = \varphi(b)$.

Exercises.

1. Let us define the relation $x \sim y \iff |x| = |y|$ on the set of complex numbers. Decide whether it is a congruence of the structure $(\mathbb{C}; +)$ or $(\mathbb{C}; \cdot)$.

2. Let us de ne the relation $x \quad y$, $x \quad [x] = y \quad [y]$ on the set of real numbers. Decide whether it is a congruence of the structure $(R; +)$ or $(R;)$.

3. Find all congruences of the structures (a) $\mathbf{A} = (fa; b; c; dg; f)$ where $f(a) = f(b) = c$ and $f(c) = f(d) = d$, (b) $\mathbf{A} = (f0; : : : ; n \quad 1g; f)$, where $f(k) = k + 1 \pmod{n}$. Based on these observations, try to describe all structures with one unary operation that have only trivial congruences (i.e., only equality $=$ and the full relation $A \quad A$).

14. Partial orders and lattices

14.1. Partially ordered sets.

In this section, we discuss some basic facts about partial orders. The reader should know most of it from previous courses.

Definition. A (binary) relation on a set A is called a partial order if it is

(1) re exive, i.e. $a \quad a$ for all $a \; 2 \; A$,
(2) transitive, i.e. $a \quad b$ and $b \quad c$ imply $a \quad c$,
(3) antisymmetric, i.e. $a \quad b$ and $b \quad a$ implies $a = b$.

We also say that $(A; \quad)$ is a partially ordered set (or poset for short). A partial order is called a linear order if in addition $a \quad b$ or $b \quad a$ holds for all $a; b \; 2 \; A$. We denote closed intervals by

$$[a; b] = fu \; 2 \; A : \; a \quad u \quad bg:$$

If $a \quad b$ and $a \; 6= \; b$, we write $a < b$.

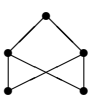
Example. We already encountered two important orders on the natural numbers:

(N;) given by the standard ordering $1 < 2 < 3 < : : :$ (this is a linear order);
(N; j) given by divisibility, i.e. a is \less than" b if $a \; j \; b$ (this partial order that is not linear).

Example. Another class of examples arises from ordering sets of sets by inclusion:

($P(X)$;): for the set $P(X)$ of all subsets of the given set X, we say that A is smaller than B if $A \quad B$;
($Eq(X)$;): on the set $Eq(X)$ of all equivalences on the set X, we say that is less than if , i.e. if $a \quad b$ implies $a \quad b$;
Similarly, the substructures of a given structure can be (partially) ordered by inclusion. Also the congruences of a given structure are (partially) ordered by inclusion.

Finite posets can be described using the so-called Hasse diagram. It is a description of by the corresponding directed graph, but we do not draw all the loops (given by re exivity), omit all edges whose existence is guaranteed by transitivity, and draw undirected edges instead of arrows, so that larger elements are on top. For example:

$\mathbf{A} =$  $\mathbf{B} =$

Definition. With respect to a partial order $(A; \quad)$ an element $a \; 2 \; A$ is

the greatest element if for every $b \; 2 \; A \; b \quad a$ holds;
the least element if $b \quad a$ holds for every $b \; 2 \; A$;
maximal if there is no $b \; 2 \; A$ such that $b > a$;
minimal if there is no $b \; 2 \; A$ such that $b < a$.

Note that, in general such elements do not need to exist. If there is a least/greatest element, then it is unique.

Examples.

The ordered set $\mathbf{A}$ has a greatest element, which is also the unique maximal element. It has no least element, but two minimal elements.
The ordered set $\mathbf{B}$ has a greatest (and at the same time maximal) and a least (and at the same time minimal) element. It is a linear order.
Both the ordered sets $(\mathbb{N};)$ and $(\mathbb{N}; j)$ have 1 as the least element but have no maximal element.
The minimal elements of the poset $(\mathbb{N} \setminus \{1\}; j)$ are exactly the prime numbers.

Definition. Let $(A;)$ be a partial order and $B \subseteq A$. We say that an element $a \in A$ is

an upper bound of B, if $a \geq b$ for all $b \in B$;
the supremum of B, if it is the least upper bound of B; we write $a = \sup B$;
a lower bound of B, if $a \leq b$ for all elements $b \in B$;
the in mum of B, if it is the greatest lower bound of B; we write $a = \inf B$.

Examples.

In the partial order $\mathbf{A}$, the subset consisting of the two minimal elements has neither supremum nor in mum. There is no in mum because it has no lower bounds. There is no supremum because it has three upper bounds, none of which is a least upper bound.
In every linearly ordered set, every nonempty nite subset has both a supremum and an in mum, with $\sup B = \max B$ and $\inf B = \min B$. It may not exist for in nite, e.g., $\sup \mathbb{N}$ in $(\mathbb{N};)$.
In the poset $(P(X);)$, every subset $B \subseteq P(X)$ has an in mum and a supremum, where $\inf B$ is equal to the intersection of all sets from B and $\sup B$ is equal to the union of all sets from B.
In the poset $(\mathbb{N}; j)$, every nite subset has an in mum and supremum, where $\inf B$ is equal to the gcd of all numbers from B and $\sup B$ is equal to the lcm of all numbers from B. On the other hand, for example, the supremum of the set of all prime numbers does not exist.

Definition. We call a poset a lattice if it contains the supremum and in mum of all two-element subsets (it is easy to prove by induction that then there also exist the suprema and in ma of all nonempty nite subsets). We call it a complete lattice if there are suprema and in ma of all subsets. In lattices, we usually denote by abbreviated

$$a \vee b = \sup\{a; b\} \quad \text{a} \quad a \wedge b = \inf\{a; b\};$$

and call the symbols $\vee; \wedge$ meet and join.

It follows from the de nition that in a complete lattice there is always a smallest and largest element (obtained by $\sup ;$ and $\inf ;$ respectively).

Examples.

Linearly ordered sets are always lattices by $a \vee b = \max(a; b)$, $a \wedge b = \min(a; b)$. They don't have to be complete lattices, an example is $(\mathbb{N};)$.
$(P(X);)$ is a complete lattice. For $U \subseteq P(X)$ $\sup U = \bigcup_{A \in U} A$, $\inf U = \bigcap_{A \in U} A$.
$(\mathbb{N}; j)$ is a lattice with $a \vee b = \operatorname{lcm}(a; b)$, $a \wedge b = \gcd(a; b)$. It is not a complete lattice.

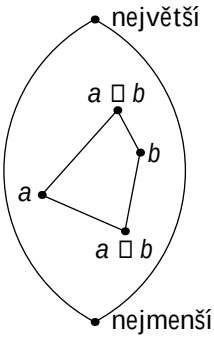


Figure 8. Meet and join in a lattice

To verify whether a given order is a complete lattice, it is sufficient to verify only the existence of infima (or only the existence of suprema).

Proposition 14.1. A partially ordered set in which infima of all subsets exist is a complete lattice. Symmetrically, also a poset in which the suprema of all subsets exist is a complete lattice.

Proof. Let us denote the given poset by $(A; \preceq)$. It follows from the definition of suprema that

$$\sup B = \inf\{a \in A : a \succeq b \text{ for every } b \in B\};$$

so suprema can be defined using infima. Dually, $\inf B = \sup\{a \in A : a \preceq b$ for every $b \in B\}$.

For some algebraic constructions we need Zorn's lemma. It is one of the forms of the so-called axiom of choice, one of the basic axioms of set theory. Therefore, Zorn's lemma is not proven but rather postulated. A chain in a partially ordered set means a subset that is linearly ordered.

Axiom 14.2 (Zorn's lemma)**.** Let $(X; \preceq)$ be a nonempty partially ordered set. Assume that every chain in $(X; \preceq)$ has an upper bound. Then $(X; \preceq)$ contains at least one maximal element.

Lastly, we will define isomorphism of ordered sets. Let $(A; \preceq)$ and $(B; \trianglelefteq)$ be two ordered sets. We call the mapping $\varphi : A \to B$ monotone if for every $a, b \in A$

$$a \preceq b \quad \Rightarrow \quad \varphi(a) \trianglelefteq \varphi(b).$$

We call a mapping φ an isomorphism between these ordered sets if it is bijective and both mappings φ, φ^{-1} are monotone. The composition of monotone mappings is monotone, but note that the inverse of a monotone bijection may not be monotone: an example is the identical mapping $x \mapsto x$ between the ordered sets $(\mathbb{N}; |)$ and $(\mathbb{N}; \leq)$ which is monotone in one direction only ($a \mid b$ implies $a \leq b$, but not vice versa).

Exercises.

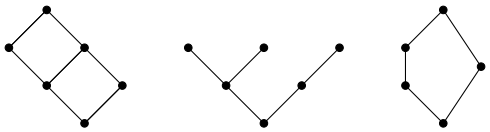
1. Draw the Hasse diagrams of the partial orders $(\{1, \ldots, 10\}; |)$, $(\mathcal{P}(\{1, 2, 3\}; \subseteq)$ and $(\mathrm{Eq}(\{1, 2, 3\}; \subseteq)$.

2. Determine whether there exists a partial order that

has at least two maximum and at least two minimum elements,
has at least two largest elements,
has at least one maximum but no minimum element,
has at least one maximum but no minimum element,
has exactly one minimum but no smallest element.

In the case it exists, nd an example that is as small as possible.

3. Which of the following posets are lattices?



4. Describe some linear order on the set N N and check if it is isomorphic to (N;), or not.

14.2. **Lattices and Boolean algebras.** Lattices can be viewed in two ways: as ordered sets, in which each pair $a; b$ of elements has both a supremum $a \_ b$ and in mum $a \wedge b$, but also as algebraic structures with two binary operations $\_$ and $\wedge$. The properties of $\_$ and $\wedge$ can be described abstractly, by the following axioms:

Definition. An algebraic structure $(A; \wedge; \_)$ with two binary operations is called a lattice, if for all $a; b; c \in A$:

$(a \wedge b) \wedge c = a \wedge (b \wedge c); \quad a \wedge b = b \wedge a; \quad a \wedge a = a;$

$(a \_ b) \_ c = a \_ (b \_ c); \quad a \_ b = b \_ a; \quad a \_ a = a;$

$a \wedge (a \_ b) = a; \quad a \_ (a \wedge b) = a$ (these last two conditions are also called the absorption laws).

Proposition 14.3 (lattices as orders vs. algebraic structures). (1) If $(A;)$ is a lattice (as a partially ordered set), then $(A; \inf; \sup)$ is a lattice (as an algebraic structure).

(2) If $(A; \wedge; \_)$ is a lattice (as an algebraic structure), and we de ne $a \quad b$, $a \wedge b = a$, then $(A;)$ is a lattice (as a partially ordered set).

Proof. (1) We verify the lattice axioms. It is easy (albeit somewhat technical) to verify that $(a \wedge b) \wedge c = \inf\{a; b; c\} = a \wedge (b \wedge c)$. Obviously, $a \wedge b = \inf\{a; b\} = b \wedge a$ and $a \wedge a = \inf\{a; a\} = a$. Analogous statements apply to $\_$ and sup. Furthermore, $a \wedge (a \_ b) = \inf\{a; \sup\{a; b\}\} = a$, since $a \quad \sup\{a; b\}$. We can verify the second absorption law in an analogous way.

(2) Re exivity of follows from $a \wedge a = a$. For transitivity: if $a \quad b \quad c$, then $a \wedge b = a$ and $b \wedge c = b$. So $a \wedge c = (a \wedge b) \wedge c = a \wedge (b \wedge c) = a \wedge b = a$, which implies $a \quad c$. For the antisymmetry, assume that $a \quad b$ and $b \quad a$. Then $a = a \wedge b = b \wedge a = b$.

We next claim that in ma of 2-element subsets exist and $\inf\{a; b\} = a \wedge b$. First, note that $a \wedge b$ is a lower bound of both elements, since $(a \wedge b) \wedge a = (a \wedge a) \wedge b = a \wedge b$, and analogously for b. Furthermore, if $c \quad a; b$ is another common lower bound, then $(a \wedge b) \wedge c = a \wedge (b \wedge c) = a \wedge c = c$, so $c \quad a \wedge b$.

Now note that the condition $a \quad b$ can be equivalently expressed as $a \_ b = b$: if $a \wedge b = a$, absorption implies $a \_ b = (a \wedge b) \_ b = b$; and conversely, if $a \_ b = b$, the absorption follows $a \wedge b = a \wedge (a \_ b) = a$.

Using this observation, it is easy to prove the existence of suprema by $\sup\{a; b\} = a \_ b$: we use the same argument as for in ma with the operation $\_$ instead of $\wedge$.

You can nd more about lattices in any textbook on universal algebra, or lattice theory.

Boolean algebras are one of the basic algebraic tools of mathematical logic. Consider the logical values $T; F$. If we interpret $\wedge$ and $\_$ as conjunction and disjunction, we get a lattice. If we add the negation $'$, we get an even richer structure.

Definition. An algebraic structure $(A; \wedge; \_; '; 0; 1)$ with two binary operations $\wedge; \_$, a unary operation $'$ and constants $0; 1$ is called a Boolean algebra, if the following conditions hold for all $a; b; c \in A$:

$(A; \wedge; \vee)$ is a lattice,
$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ (i.e. the lattice is distributive),
$a \wedge 0 = 0$, $a \vee 1 = 1$,
$a \wedge a^0 = 0$, $a \vee a^0 = 1$.

It is easy to derive other useful properties from the axioms:

$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ (dual distributivity),
$(a^0)^0 = a$, $0^0 = 1$, $1^0 = 0$,
$(a \wedge b)^0 = a^0 \vee b^0$ a $(a \vee b)^0 = a^0 \wedge b^0$ (de Morgan's laws).

We leave the proof to the reader.

Example. A basic example of a Boolean algebra is the set algebra

$$(P(X); [; \setminus; ;; X);$$

with the intersection, union, and complement operation ($A = X \setminus A$).

It is not difficult to prove that every finite Boolean algebra is isomorphic to some set algebra. There are, however, many other types of infinite Boolean algebras.

Example. As we already mentioned that the truth values $T; F$ together with conjunction, disjunction, and negation form the Boolean algebra $(\{T; F\}; \wedge; \vee; :; F; T)$. This algebra is tied to basic propositional logic.

Let us now consider a general theory T in a language L (for example, group theory or Peano arithmetic). Let F_L denote all first-formulas in the language L. We call two formulas $'; \$$ T-equivalent if the equivalence of $' \$$ can be proved in the theory of T. Let us consider the set F_T, in which we put one formula from each equivalence class. The so-called Lindenbaum algebra (or Lindenbaum-Tarski algebra) of the theory T is the Boolean algebra $(F_T; \wedge; \vee; :; F; T)$.

Lindenbaum algebras measure the incompleteness of a given theory: if every formula is provable or falsifiable from the axioms, then F_T will have only two elements, otherwise there will be more (e.g. for group theory, where neither commutativity nor its negation can be proved from the group axioms).

You can find more about Boolean algebras in most textbooks on logic and set theory. Furthermore, many non-classical logics are based on generalizations of Boolean algebras.

Appendix

15. Dictionary

English	Čeština	Deutsch	Italiano
group	grupa	Gruppe	gruppo
abelian group	Abelova grupa	abelscher Gruppe	gruppo abeliano
ring	okruh	Ring	anello
field	tělo	Körper	campo
commutative ring	komutativní okruh	kommutativer Ring	anello commutativo
integral domain	obor integrity	Integritätsring/Integritätsbereich	dominio d'integrità
unique factorization domain (UFD)	Gaussův obor integrity	faktorieller/gaußscher Ring	dominio a fattorizzazione unica
principal ideal domain (PID)	Obor hlavních ideálů	Hauptidealring	dominio ad ideali principali
Euclidean domain/ring	Eukleidovský obor/okruh	Euklidischer Ring	dominio euclideo
homomorphism	homomorfismus	Homomorphismus	omomorfismo
isomorphism	isomorfismus	Isomorphismus	isomorfismo
normal subgroup	normální podgrupa	Normalteiler	sottogruppo normale
ideal	ideál	Ideal	ideale
quotient/factor ring	faktorokruh	Faktorring/Quotientenring	anello quoziente
quotient/factor group	faktorgrupa	Faktorgruppe/Quotientengruppe	gruppo quoziente
kernel	jádro	Kern	nucleo
image	obraz	Bild	immagine
divisor	dělitelnost	Teiler	divisore
prime number	prvočíslo	Primzahl	numero primo
irreducible	ireducibilní	irreduzibel	irriducibile
coprime/relatively prime	nesoudělná	teilerfremd	coprimi
greatest common divisor (gcd)	největší společný dělitel (NSD)	größter gemeinsamer Teiler (ggT)	massimo comun divisore (MCD)
least common multiple (lcm)	Nejmenší společný násobek (NSN)	kleinstes gemeinsames Vielfaches (kgV)	minimo comune multiplo (mcm)
Chinese remainder theorem	Čínská věta o zbytcích	chinsesischer Restklassensatz	teorema cinese del resto
field extension	tělesové rozšíření	Körpererweiterung	estensione di campi
algebraic/transcendental element	algebraický/transcendentní prvek	algebraisches/transzendentes Element	numero algebraico/trancendente
rupture field	kořenové nadtěleso	Nullstellenkörper	
splitting field	rozkladové těleso	Zerfällungskörper	campo di spezzamento
set	množina	Menge	insieme
partial order	částečné uspořádání	partielle Ordnung/Halbordnung	ordine
linear/total order	lineární/úplné uspořádání	lineare Ordnung/Totalordnung	ordine semplice/lineare/totale
lattice	svaz	Verband	reticolo
Boolean algebra	Booleova algebra	boolesche Algebra	algebra di Boole