HONOURS ALGEBRA 1 (MATH 245)
COURSE NOTES
FALL 2024
VERSION: December 5, 2024

EYAL Z. GOREN,
MCGILL UNIVERSITY
EYAL.GOREN@MCGILL.CA

These notes will be updated and corrected throughout the semester. I will mark with ✠ the point I got to in relation to revisions. If you find any mistakes kindly bring them to my attention, especially if they are in the part that was already revised. Thanks, E.G.

Contents

Part 1. Some Language and Notation of Mathematics 3
1. Sets 3
1.1. First definitions 3
1.2. Algebra of set operations 5
2. Proofs: ideas and techniques 6
2.1. Proving equality by two inequalities 6
2.2. Proof by contradiction and the contrapositive 7
2.3. Proof by Induction 7
2.4. Prove or disprove 9
2.5. The pigeonhole principle 9
3. Functions 10
3.1. Injective, surjective, bijective and inverse image 12
3.2. Composition of functions 12
3.3. The inverse function 13
4. Cardinality of a set 13
4.1. Cantor's diagonal argument 16
5. Relations 16
6. Number systems 19
6.1. The polar representation 21
6.2. The complex exponential function 22
6.3. The Fundamental Theorem of Algebra 23
7. Fields and rings - definitions and first examples 25
7.1. Some formal consequences of the axioms 27
8. Exercises 29

Part 2. Arithmetic in $\mathbb{Z}$ 33
9. Division, GCD and the Euclidean Algorithm 33
9.1. Division with residue 33
9.2. Division 34
9.3. GCD 34
9.4. The Euclidean algorithm 35
10. Primes and unique factorization 36
10.1. Primes 36
10.2. Applications of the Fundamental Theorem of Arithmetic 39
11. Exercises 42

Part 3. Congruences and modular arithmetic 44
12. Congruence modulo n 44
12.1. Fermat's little theorem 46
12.2. Solving equations in $\mathbb{Z}/n\mathbb{Z}$. 47
12.2.1. Linear equations. 47
12.2.2. Quadratic equations. 48
12.3. Public key cryptography; the RSA method 48
13. Exercises 50

Part 4. Polynomials and their arithmetic 52
14. The ring of polynomials 52
15. Division with residue 53
16. Arithmetic in $\mathbb{F}[x]$ 54
16.1. Some remarks about divisibility in a commutative ring T 54
16.2. GCD of polynomials 54
16.3. The Euclidean algorithm for polynomials 55
16.4. Irreducible polynomials and unique factorization 56
16.5. Roots 59
16.6. Eisenstein's criterion 60
16.7. Roots of polynomials in $\mathbb{Z}/p\mathbb{Z}$ 61

17. Exercises 62

Part 5. Rings 63
18. Some basic definitions and examples 63
19. Ideals 65
20. Homomorphisms 68
20.1. Units 70
21. Quotient rings 71
21.1. The quotient ring $\mathbb{F}[x]/(f(x))$ 73
21.2. Every polynomial has a root in a bigger field 75
21.3. Roots of polynomials over $\mathbb{Z}/p\mathbb{Z}$ 75
22. The First Isomorphism Theorem 76
22.1. Isomorphism of rings 76
22.2. The First Isomorphism Theorem 76
22.3. The Chinese Remainder Theorem 77
22.3.1. Inverting $\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ 78
23. Prime and maximal ideals 80
24. Exercises 82

Part 6. Groups 85
25. First definitions and examples 85
25.1. Definitions and some formal consequences 85
25.2. Examples 85
25.3. Subgroups 86
26. Permutation groups and dihedral groups 87
26.1. Permutation groups 87
26.2. Cycles 88
26.3. The Dihedral group 90
27. The theorem of Lagrange 91
27.1. Cosets 91
27.2. Lagrange's theorem 92
27.3. Orders of elements of $\mathbb{Z}/n\mathbb{Z}$ 93
27.3.1. Euler's totient function 93
28. Homomorphisms and isomorphisms 94
28.1. homomorphisms of groups 94
28.2. Isomorphism 95
29. Group actions on sets 96
29.1. Basic definitions 96
29.2. Basic properties 96
29.3. Some examples 97
30. The Cauchy-Frobenius Formula 98
30.1. Some applications to Combinatorics 99
31. Cauchy's theorem: a wonderful proof 102
32. Wallpaper groups 102
33. The first isomorphism theorem for groups 103
33.1. Normal subgroups 103
33.2. Quotient groups 104
33.3. The first isomorphism theorem 104
33.4. Groups of low order 105
33.4.1. Groups of order 1 105
33.4.2. Groups of order 2, 3, 5, 7 105
33.4.3. Groups of order 4 105
33.4.4. Groups of order 6 106
33.5. Odds and evens 106
33.6. Odds and Ends 107
34. Exercises 108

Part 7. Appendices 110

Appendix A: The Cantor-Bernstein Theorem 110
Appendix B: The irrationality of e 111
Appendix C: Euclidean rings 111
Appendix D: The complex exponential function 112
Index 115

Introduction.

It is important to realize that Algebra started in antiquity as an *applied* science. As every science, it was born of necessity; in this case, the need to solve everyday problems some of which are mentioned below. Since the 19-th century, if not even earlier, there is a growing side to Algebra which is purely theoretic. However, it is important to realize that some of the most abstract algebraic structures of the past, such as Galois fields (nowadays simply called finite fields), have become in modern days part of the foundation of certain branches of applied algebra, for example in applications to coding theory and cryptography. The lesson is (and please repeat that to any politician you happen to meet) is that what is at present considered "pure" and what is considered "applied" is our temporary perspective and in time many of the pure, seemingly useless, branches of mathematics turn out to have critical relevance to real world applications.

The word "algebra" is derived from the title of a book - *Hisab al-jabr w'al-muqabala* - written by the Farsi scholar Abu Ja'far Muhammad ibn Musa Al-Khwarizmi (790 - 840 AD). The word *al-jabr* itself comes from the root of *reunite* (in the sense of completing or putting together) and refers to one of the methods of solving quadratic equations (by completing the square) described in that book. The book can be considered as the first treatise on algebra. The word *algorithm* is in fact derived from the name *Al-Khwarizmi*. The book was very much concerned with methods for solving known practical problems; Al-Khwarizmi intended to teach (in his own words) *"... what is easiest and most useful in arithmetic, such as men constantly require in cases of inheritance, legacies, partition, lawsuits, and trade, and in all their dealings with one another, or where the measuring of lands, the digging of canals, geometrical computations, and other objects of various sorts and kinds are concerned."* [1]

Algebra I is a first course in Algebra. Little is assumed in the way of background. Though the course is self-contained, it puts some of the responsibility of digesting and exploring the material on the student, as is normal in university studies. You'll soon realize that we are also learning a new language in this course and a new attitude towards mathematics. The language is the language of modern mathematics; it is very formal, precise and concise. One of the challenges of the course is digesting and memorizing the new concepts and definitions. The new attitude is an attitude where any assumptions one is making while making an argument have to be justified, or at least clearly stated as a postulate, and from there on one proceeds in a logical and clear manner towards the conclusion. This is called *proof* and one of the main challenges in the course is to understand what constitutes a good proof and to be able to write proofs yourself.[2] A further challenge for most students is that the key ideas we learn in this course are very abstract, bordering on philosophy and art, yet they are truly scientific in their precision. You should *expect* to not understand everything right away; you should *expect* to need time to reflect on the meaning of the new ideas and concepts that we introduce.

Here are some pointers as to how to cope with the challenges of this course:

- Read the class notes and the textbook over and over again. Try and give yourself examples of the theorems and propositions and try and provide counterexamples when some of the hypotheses are dropped.
- Do lots and lots of exercises. The more, the better.
- This textbook attempts to be lean and so everything in it is important. Examples and exercises may contain important observations and referred to later.
- Explain to your friends, and possibly to your family, the material of the course.[3] Work together with your class mates on assignments, but write your own solutions in the end; try to understand different solutions to assignments and try and find flaws in your friends' solutions.
- Use the instructor's and the TA's office hours, as well as the math help center, to quickly close any gap and clarify any point you're not sure about.

[1] Cited from http://www-groups.dcs.st-and.ac.uk/ history/Biographies/Al-Khwarizmi.html.

[2] To quote former Prime Minister Jean Chrétien, *"The proof is the proof and when you have a good proof it is proven"*.

[3] Unless you start noticing that you are becoming less and less popular.

So what is this course really about?

We are going to start by learning some of the notation and language of mathematics. We are going to discuss sets and functions and various properties and operations one can perform on those. We are going to talk about proofs and some techniques of proof, such as “induction" and “proving the counter-positive". We are going “to split infinities"; some infinities are more infinite than others.

We are going to discuss different structures in which one can do arithmetic, such as rational, real and complex numbers, polynomial rings and yet more abstract systems called rings and fields. We are going to see unifying patterns of such systems and some of their applications. For example, a finite field, perhaps initially a strange beast, is a key notion in modern days computer science; it is a concept absolutely essential and fundamental for cryptographic schemes as well as data management.

We are then going to do some really abstract algebra for a while, learning about rings and homomorphisms and ideals. Our motivation is mostly to know how to construct a finite field and work with it in practice. We also lay the basis for further study in the following algebra courses.

The final section of the course deals with groups and group actions on sets. After the previous section on rings and fields we'll feel more comfortable with the abstract notions of group theory. Furthermore, there are going to be plenty of concrete examples in this section and applications to problems in combinatorics and to the study of symmetries. Here is one concrete example: imagine that a jewelry company wants, as a publicity stunt, to display all necklaces one can make using 10 diamonds and 10 rubies, perhaps under the (banal) slogan “to each their own". In each necklace 10 diamonds and 10 rubies are to be used. The necklace itself is just round, with no hanging or protruding parts. Thus, we can provide an example of such a necklace as

D D R R D R D R R R D D D D R R D R D R

(where the last R is adjacent to the first D). Now, when we consider a particular design such as above, we do want to identify it with the following design

D R R D R D R R R D D D D R R D R D R D

(we've put the first D at the last spot), because this is just the same pattern; if the necklace is put on a table, it is just rotating it a bit, or, alternately, looking at it from a slightly different angle. Also, note that the pattern

D D R R D R D R R R D D D D R R D R D R

is identified with

R D R D R R D D D D R R R D R D R R D D

which corresponds to flipping over the necklace as the second sequence is the first sequence read from right to left. Now the question is how many rubies and diamonds we need to purchase in order to make all the different designs? It turns out that this can be approached using the theory of group actions on sets and a general formula we'll develop can be applied here. It turns out that there are 4752 such different designs; that will require 47520 diamonds and 47520 rubies. Perhaps the idea should be reconsidered ;-)

Part 1. Some Language and Notation of Mathematics

1. Sets

1.1. **First definitions.** A **set** is a collection of elements. The notion of a set is logically not quite defined (what's a "collection"? an "element"?) but, hopefully, it makes sense to us. What we have is the ability to say whether an element is a member of a set or not. Thus, in a sense, a set is a *property*, and its elements are the objects having that property (the property **is** to be in the set).[4]

There are various ways to define sets:

(1) By writing it down:

$$S = \{1, 3, 5\}.$$

The set is named S and its elements are $1, 3$ and 5. The use of curly brackets is mandatory! Another example is

$$T = \{2, 3, \text{Jim's football}\}.$$

This is a set whose elements are the numbers $2, 3$ and Jim's football. It is assumed here that "Jim" refers to one particular individual.

A set can also be given as all objects with a certain property:

$$S_1 = \{\text{all beluga whales}\}.$$

Another example is

$$T_5 = \{n : n \text{ is an odd integer}, n^3 = n\}.$$

The colon means that the part that follows is the list of properties n must satisfy, i.e. the colon is shorthand for "such as". Note that this set is equal to the set

$$U^+ = \{n : n^2 = 1\}.$$

Our eccentric notation T_5, S_1, U^+ is just to make a point that a set can be denoted in many ways.

(2) Sometimes we write a set where the description of its elements is implicit, to be understood by the reader. For example:

$$\mathbb{N} = \{0, 1, 2, 3, \dots\}, \qquad \mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\},$$

and

$$\mathbb{Q} = \left\{\frac{a}{b} : a, b \in \mathbb{Z}, b \neq 0\right\}.$$

Thus $\mathbb{N}$ is the set of **natural numbers**, $\mathbb{Z}$ is the set of **integers** and $\mathbb{Q}$ the set of **rational numbers**.[5] The use of the letters $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ is standard. Other standard notation is

$$\mathbb{R} = \text{the set of } \textbf{real} \text{ numbers (= points on the line)},$$

and the **complex numbers**

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}.$$

Here i is the imaginary number satisfying $i^2 = -1$ (we'll come back to that in §6). Note that we sneaked in new notation. If A is a set, the notation $x \in A$ means x is an element (a member) of A, while $x \notin A$ means that x is not an element of A. Thus, the expression $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$ is saying that $\mathbb{C}$ is the set whose elements are $a + bi$, where a and b are real numbers; these are formal expressions and it is not assumed that we know how to add a to bi. For example, $1 + i, \sqrt{3} + \pi i$ are complex numbers. Every real number r is a complex number that we still write as r instead of the more formal expression $r + 0i$. For example, in the notation above, $3 \in S, 2 \notin S$, Jim's football $\in T$ but $\notin U^+$, $i \in \mathbb{C}$ but $i \notin \mathbb{R}$.

We haven't really defined any of these sets rigorously. We have assumed that the reader understands what we mean. This suffices for the level of this course. A rigorous treatment is usually

[4] Like some upscale gentlemen clubs, they are defined as much by those that aren't members as by those that are.

[5] For us 0 is a natural number, that is, we include it in $\mathbb{N}$, but some authors do not. The letter $\mathbb{Z}$ comes from "zahlen" meaning "numbers" in german and $\mathbb{Q}$ comes from "quotient".

given in a logic course for $\mathbb{N}$ (constructed via the Peano axioms) or in an analysis course for $\mathbb{R}$ (via Dedekind cuts). The set of real numbers can also be thought of as the set of all numbers written in a possibly infinite decimal expansion. Thus, $1, 2, 1/3 = 0.33333\ldots$ and indeed any rational number is an element of $\mathbb{R}$ as are $\pi = 3.1415926\ldots, e = 2.718281\ldots, \sqrt{2} = 1.414\ldots$ and so on. In fact π, e and $\sqrt{2}$ are not rational numbers; we will prove that $\sqrt{2}$ is not rational in Proposition 10.2.4, and that e is irrational in Appendix B. The proof that π is irrational is much harder.

Two sets A, B, are **equal**, $A = B$, if they have the same elements, that is, if every element of A is an element of B and vice-versa. Thus, for example, $A = \{1, 2\}$ is equal to $B = \{2, 1\}$ (the order doesn't matter), and $A = \{-1, 1\}$ is equal to $B = \{x \in \mathbb{R} : x^2 - 1 = 0\}$ (the description doesn't matter). Also $A = \{1, -1\}$ is equal to $B = \{1, 1, -1, 1\}$ (repetitions do not matter, either). We say that

$$A \subseteq B,$$

(A is **contained in** B), or simply $A \subset B$, if every element of A is an element of B. For example $\mathbb{N} \subset \mathbb{Z}$. Some authors use $A \subset B$ to mean A is contained in B but not equal to it. We do **not** follow this convention and for us $A \subset B$ allows $A = B$. If we want to say that A is contained in B and not equal to it, we shall use $A \subsetneqq B$. Note that $A = B$ holds precisely when both $A \subset B$ and $B \subset A$.

The notation

$$\emptyset$$

stands for the **empty set**. It *is* a set but it has no elements. Admittedly, that sounds funny... the logic behind is that we want the intersection of sets to always be a set. We let

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

be the **intersection** of A and B, the set of common elements, and we let

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

be the **union** of A and B. For example, $\{1, 3\} \cap \{n : n^2 = n\} = \{1\}$, $\mathbb{N} \cap \{x : -x \in \mathbb{N}\} = \{0\}$, $S_1 \cap T_5 = \emptyset$.[6]

We shall also need arbitrary unions and intersections. Let I be a non-empty set (thought of as an **index set**) and suppose that for each $i \in I$ we are given a set A_i. Then

$$\cap_{i \in I} A_i = \{x : x \in A_i, \forall i\},$$

($\forall$ means "for all") is the set of elements belonging to each A_i, and

$$\cup_{i \in I} A_i = \{x : x \in A_i, \text{ for some } i\},$$

is the set of elements appearing in at least one A_i. For example, define for $i \in \mathbb{Z}$,

$$A_i = \{x \in \mathbb{Z} : x \geq i\}$$

(so $A_{-1} = \{-1, 0, 1, 2, 3, 4, \ldots\}, A_0 = \{0, 1, 2, 3, \ldots\}, A_1 = \{1, 2, 3, 4, \ldots\}, A_2 = \{2, 3, 4, 5, \ldots\}$ and so on). Then $\cup_{i \in \mathbb{Z}} A_i = \mathbb{Z}$, while $\cap_{i \in \mathbb{Z}} A_i = \emptyset$.

Here's a another example: for every real number x, $0 \leq x \leq 1$ define

$$S_x = \{(x, y) : 0 \leq y \leq x, y \in \mathbb{R}\}.$$

Then $\cup_{0 \leq x \leq 1} S_x$ is the triangular area in the plane whose vertices are $(0, 0), (1, 0), (1, 1)$.

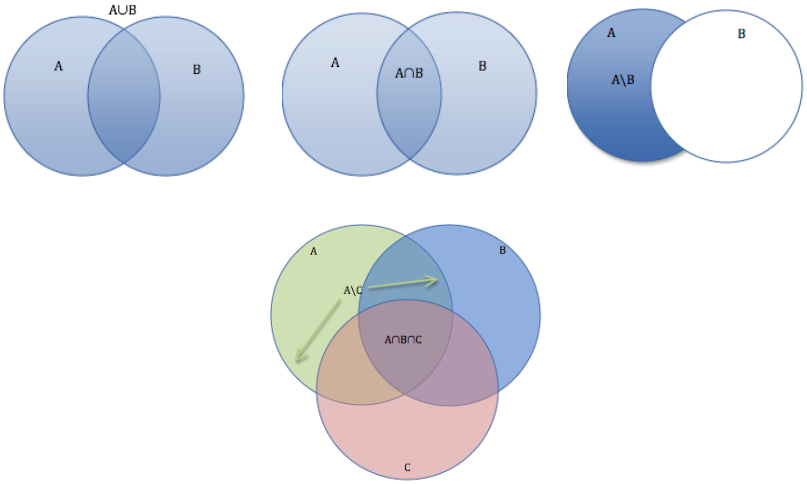
Yet another operation on sets is the **difference** of sets:

$$A \setminus B = \{x : x \in A, x \notin B\}.$$

We will also use the notation $A - B$ instead of $A \setminus B$. We remark that $A \cup B = B \cup A, A \cap B = B \cap A$ but $A \setminus B$ is not equal to $B \setminus A$, unless $A = B$.

[6] Now wrap your mind around this: We may form sets whose elements are sets themselves. For example, if $A_1 = \{1, 2\}, A_2 = \{3, 4\}$ then we can form the set $S := \{A_1, A_2\}$. The set S has precisely two elements, namely, A_1 and A_2. Note that 3, for example, is not an element of S. So far so good. But consider the set $T = \{\emptyset\}$. The set T is *not* empty - it has precisely one element. That element is the empty set. Likewise, the set $\{\emptyset, \{\emptyset\}\}$ has *two* elements. One element is the empty set, the other is the set T. As T is not the empty set, these two elements are indeed distinct. Thus, we have created matter out of vacuum.

A good way to decipher formulas involving two or three sets is by diagrams. For example:



The last diagram shows clearly that $A \cup B \cup C - (A \cap B \cap C) = (A - C) \cup (B - A) \cup (C - B)$, though that is not considered a proof!

Another definition we shall often use is that of the **cartesian product**. Let $A_1, A_2, \ldots, A_n$ be sets. Then

$$A_1 \times A_2 \times \cdots \times A_n = \{(x_1, x_2, \ldots, x_n) : x_i \in A_i, \text{ for } 1 \leq i \leq n\}.$$

In particular,

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

Example 1.1.1. Let $A = \{1, 2, 3\}, B = \{1, 2\}$. Then

$$A \times B = \{(1,1), (1,2), (2,1), (2,2), (3,1), (3,2)\}.$$

Note that $(3,1) \in A \times B$ but $(1,3) \notin A \times B$, because although $1 \in A, 3 \notin B$.

Example 1.1.2. Let $A = B = \mathbb{R}$. Then

$$A \times B = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{R}\}.$$

This is just the presentation of the plane in cartesian coordinates (which explains why we call such products “cartesian" products).

1.2. **Algebra of set operations.** Up till now we have only introduced notation and vocabulary. With the exception of the notion of the empty set, none of what we had said had any depth. We now wish to make general statements relating some of these operations. Once a statement is important enough to highlight it, it falls under the heading of a Lemma, a Proposition or a Theorem (or, more colloquially, a Claim, an Assertion and so on). Usually, “Lemma" is reserved for technical statements often to be used in the proof of a proposition or a theorem. “Proposition" and “Theorem" are more or less the same. They are used for claims that are more conceptual, or central, with “Theorem" implying even more importance. However, none of these rules is absolute. For example, consider the following proposition:

Proposition 1.2.1. *Let I be a set. Let A be a set and $B_i, i \in I$, be sets as well, then*

$$A \cap (\cup_{i \in I} B_i) = \cup_{i \in I} (A \cap B_i),$$

and

$$A \cup (\cap_{i \in I} B_i) = \cap_{i \in I} (A \cup B_i).$$

Furthermore,

$$A \setminus (\cup_{i \in I} B_i) = \cap_{i \in I} (A \setminus B_i),$$

and

$$A \setminus (\cap_{i\in I} B_i) = \cup_{i\in I}(A \setminus B_i).$$

We will prove parts of this Proposition below and the rest is left as an exercise.

2. Proofs: ideas and techniques

Proposition 1.2.1 is not obvious. It is not even clear at first sight whether it's correct. For that reason we insist on proofs in mathematics. Proofs give us confidence that we are making true statements and they reveal, to a lesser or higher extent, why the statements hold true. The proof should **demonstrate** that the statements made are true. In fact, in French the commonly used word for a proof is **démonstration**, in the past "demonstration" was often used instead of "proof" also in English (though today this usage is very rare), and the famous QED in the end of a proof stands for "quod erat demonstrandum", meaning, "which was to be demonstrated ". We shall prove some of the statements in the proposition now, leaving the rest as an exercise. Our method of proof for Proposition 1.2.1 is by a standard technique.

2.1. **Proving equality by two inequalities.** When one wants to show that two real numbers x, y are equal, it is often easier to show instead that $x \leq y$ and $y \leq x$ and to conclude that $x = y$.

In the same spirit, to show two sets A and B are equal, one may show that every element of A is an element of B and that every element of B is an element of A. That is, we prove two "inequalities", $A \subseteq B$ and $B \subseteq A$. Thus, our principle of proof is

$$A = B$$

if and only if

$$x \in A \Rightarrow x \in B \quad \text{and} \quad x \in B \Rightarrow x \in A.$$

(The notation $\Rightarrow$ means "implies that".)

Let us now prove the statement $A \cap (\cup_{i\in I} B_i) = \cup_{i\in I}(A \cap B_i)$. The way you should write it in an assignment, a test, or a research paper is as follows:

Proposition 2.1.1. *Let I be a set. Let A and $B_i, i \in I$, be sets then*

$$A \cap (\cup_{i\in I} B_i) = \cup_{i\in I}(A \cap B_i).$$

Proof. Let $x \in A \cap (\cup_{i\in I} B_i)$ then $x \in A$ and $x \in \cup_{i\in I} B_i$. That is, $x \in A$ and $x \in B_{i_0}$ for some $i_0 \in I$. Then $x \in A \cap B_{i_0}$ and so $x \in \cup_{i\in I}(A \cap B_i)$. We have shown so far that $A \cap (\cup_{i\in I} B_i) \subset \cup_{i\in I}(A \cap B_i)$.

Conversely, let $x \in \cup_{i\in I}(A \cap B_i)$. Then, there is some $i_0 \in I$ such that $x \in A \cap B_{i_0}$ and so for that i_0 we have $x \in A$ and $x \in B_{i_0}$. In particular, $x \in A$ and $x \in \cup_{i\in I} B_i$ and so $x \in A \cap (\cup_{i\in I} B_i)$. □

(The □ designates that the proof is complete. It is equivalent to writing QED.)

Let's also do

$$A \setminus (\cup_{i\in I} B_i) = \cap_{i\in I}(A \setminus B_i).$$

We use the same technique. Let $x \in A \setminus (\cup_{i\in I} B_i)$ thus $x \in A$ and $x \notin \cup_{i\in I} B_i$. That means that $x \in A$ and for all $i \in I$ we have $x \notin B_i$. That is, for all $i \in I$ we have $x \in A \setminus B_i$ and so $x \in \cap_{i\in I}(A \setminus B_i)$.

Conversely, let $x \in \cap_{i\in I}(A \setminus B_i)$. Then, for all $i \in I$ we have $x \in A \setminus B_i$. That is, $x \in A$ and $x \notin B_i$ for every i. Thus, $x \in A$ and $x \notin \cup_{i\in I} B_i$ and it follows that $x \in A \setminus (\cup_{i\in I} B_i)$.

2.2. **Proof by contradiction and the contrapositive. Proof by contradiction** is a very useful technique, even though using it too often shows lack of deeper understanding of the subject. Suppose that some statement is to be proven true. In this technique one assumes that the statement is false and then one proceeds to derive logical consequences of this assumption until an obvious contradiction arises. Here is an easy example that illustrates this:

Claim. *There is no solution to the equation $x^2 - y^2 = 1$ in positive integers.*

Proof. Assume not. Then there are positive integers x, y such that $x^2 - y^2 = 1$. Then, $(x-y)(x+y) = 1$. However, the only product of integers giving 1 is 1×1 or -1×-1 and, in any case, it follows that $x - y = x + y$. It follows that $2y = (x+y) - (x-y) = 0$ and so that $y = 0$. Contradiction (because we have assumed both x and y are positive). □

Claim. *If x and y are two integers whose sum is odd, then exactly one of them is odd.*

Proof. Suppose not. Then, either both x and y are odd, or both x and y are even. In the first case $x = 2a+1, y = 2b+1$, for some integers a, b, and $x + y = 2(a+b+1)$ is even; contradiction. In the second case, $x = 2a, y = 2b$, for some integers a, b, and $x + y = 2(a+b)$ is even; contradiction again. □

Somewhat related is the technique of **proving the contrapositive**. Let A and B be two assertions and let $\neg A, \neg B$ be their negation. Logically the implication

$$A \Rightarrow B$$

is equivalent to

$$\neg B \ \Rightarrow \ \neg A.$$

Here is an example. Let A be the statement "it rains" and B the statement "it's wet outside". Then $\neg A$ is the statement "it doesn't rain" and $\neg B$ is the statement "it's dry outside". The meaning of $A \Rightarrow B$ is "it rains therefore it's wet outside" and its contrapositive is "it's dry outside therefore it doesn't rain". Those statements are equivalent. Here is a mathematical example:

Claim. *If x and y are integers such that xy is even then either x or y are even.* [7]

Proof. The contrapositive is: *If both x and y are odd then xy is odd.* To prove that, write $x = 2a+1, y = 2b+1$ for some integers a, b. Then $xy = 4ab + 2a + 2b + 1 = 2(2ab + a + b) + 1$, is one more than an even integer and so is odd. □

2.3. **Proof by Induction.** Induction is perhaps the most fun technique. Its logical foundations also lie deeper than the previous methods. The principle of induction, to be explained below, rests on the following *axiom*:

Axiom: *Every non empty subset of $\mathbb{N}$ has a minimal element.*

We remark that the axiom is actually intuitively obviously true. The reason we state it as an axiom is that when one develops the theory of sets in a very formal way from fundamental axioms the assertion stated above doesn't follow from simpler axioms and, in one form or another, has to be included as an axiom.

Theorem 2.3.1. *(Principle of Induction) Let n_0 be a given natural number. Suppose that for every natural number $n \geq n_0$ we are given a statement P_n. Suppose that we know that:*

(1) P_{n_0} is true.

(2) If P_n is true then P_{n+1} is true.

Then P_n is true for every $n \geq n_0$.

[7] Note that in mathematics "or" always means "and/or".

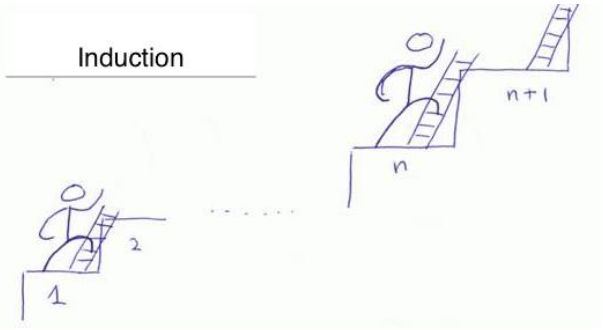
Proof. Suppose not. Then the set

$$S = \{n \in \mathbb{N} : n \geq n_0, P_n \text{ is false}\}$$

is a non-empty set and therefore has a minimal element a. Note that $a > n_0$, because P_{n_0} is true by (1), and so $a - 1 \geq n_0$. Now $a - 1 \notin S$ because of the minimality of a and so P_{a-1} is true. But then, by (2), also P_a is true. Contradiction. □

Remark 2.3.2. This proof is a gem of logic. Understand it and memorize it.

Remark 2.3.3. The mental picture I have of induction is a picture of a staircase. The first step is marked n_0 and each higher step by the integers $n_0+1, n_0+2, n_o+3, \ldots$. I know I can reach that first step and I know that if I can reach a certain step (say marked n) then I can reach the next one (marked $n+1$). I conclude that I can reach every step.



Example 2.3.4. Prove that for every positive integer n we have

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

(The statement P_n is $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ and it is made for $n \geq 1$, that is $n_0 = 1$.)

The first case is when $n = 1$ (this is called the **base case** of the induction). In this case we need to show that $1 = \frac{1\cdot(1+1)}{2}$, which is obvious.

Now, we assume that statement true for n, that is we assume that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2},$$

and we need to show it's true for $n+1$. That is, we need to prove that

$$1 + 2 + \cdots + n + (n+1) = \frac{(n+1)(n+2)}{2}.$$

(By achieving that we would have shown that if P_n is true then P_{n+1} is true.) We use the assumption that it's true for n (that's called the **induction hypothesis**) and write

$$\begin{aligned} 1 + 2 + \cdots + n + (n+1) &= \frac{n(n+1)}{2} + n + 1 \\ &= \frac{n(n+1) + 2(n+1)}{2} \\ &= \frac{n^2 + 3n + 2}{2} \\ &= \frac{(n+1)(n+2)}{2}. \end{aligned}$$

Thus, we have shown that P_{n+1} is true as well and the proof is complete.

Example 2.3.5. Here we prove the following statement: *Let $q \neq 1$ be a real number. Then for every $n \in \mathbb{N}$ we have*

$$1 + q + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q}.$$

The statement P_n is *"for every real number $q \neq 1$, $1+q+\cdots+q^n = \frac{1-q^{n+1}}{1-q}$"*. The base case, that is the first n for the statement is being claimed true, is $n=0$; the statement is then

$$1 = \frac{1-q}{1-q},$$

which is obviously true.

Now suppose that the statement is true for n. That is, suppose that

$$1+q+\cdots+q^n = \frac{1-q^{n+1}}{1-q}.$$

We need to show that

$$1+q+\cdots+q^{n+1} = \frac{1-q^{n+2}}{1-q}.$$

Indeed,

$$\begin{aligned} 1+q+\cdots+q^{n+1} &= (1+q+\cdots+q^n)+q^{n+1} \\ &= \frac{1-q^{n+1}}{1-q} + q^{n+1} \\ &= \frac{1-q^{n+1}}{1-q} + \frac{(1-q)q^{n+1}}{1-q} \\ &= \frac{1-q^{n+1}}{1-q} + \frac{q^{n+1}-q^{n+2}}{1-q} \\ &= \frac{1-q^{n+2}}{1-q}. \end{aligned}$$

Here are further examples of statements that are easy to prove by induction. (i) For $n \geq 0, n^2 \leq 4^n$. (ii) $1+3+\cdots+(2n-1) = n^2$, for $n \geq 1$. More thought is required for the following: (iii) Consider the following scenario: there are n people in a party and when the bell rings each is supposed to throw a pie at the person closest to him or her. (Well, things *were* getting wild...) Assume that all the distances between the people are mutually distinct. If n is odd then at least one person is not going to be hit by a pie.

2.4. **Prove or disprove.** A common exercise, and a situation one often faces in research, is to prove or disprove a particular statement. For example,

"Prove or disprove: for every natural number n, $4n+1$ is either a square or a sum of two squares."

At that point you are requested first to form a hunch, a guess, an opinion about whether the statement is true or false. To form that hunch you can try some examples ($4*0+1 = 1 = 1^2, 4*1+1 = 5 = 1^2+2^2, 4*2+1 = 9 = 3^2, 4*3+1 = 13 = 2^2+3^2$, ...) to see if the statement holds for these examples, or you can think whether the statement is similar to other statements you know to hold true/false, or, when at loss, throw a coin. After deciding on your initial position, if you believe the statement is true you should proceed to find a proof. If you don't, then you have two options. You can try and show that if the statement is true it will imply a contradiction to a known fact, or you can provide *one* counterexample. The statement being false doesn't mean it's false for *every* n; it means it's false for *at least one* n. In the case at hand, if we take $n=5$ we find that $4*5+1 = 21$, which is neither a square nor a sum of squares (just try all possibilities) and so the statement is false.

2.5. **The pigeonhole principle.** The pigeonhole principle is simple to state, yet it is a powerful tool. It states the following:[8]

If there are more pigeons than pigeonholes then in one pigeonhole there must be at least two pigeons.

[8] To quote Clint Eastwood: "I know things about pigeons, Lily".

Following are examples of applications of the pigeonhole principle. To appreciate its power, I recommend reading the statement first and trying to come up with an independent proof, before reading the provided proof.

Example 2.5.1. *Let a,b,c,d,e,f be 6 integers. Then there are among them two integers whose difference is divisible by 5.*

To prove this consider the remainders of a,b,c,d,e,f upon division by 5. The remainder is either $0,1,2,3$ or 4 (these are the 5 pigeonholes) but we get from our numbers 6 remainders (those are the pigeons). Therefore, there are two numbers among a,b,c,d,e,f with the same remainder. Say, a and b have the same remainder, say r. Then $a-b$ is divisible by 5 (because $a=5a'+r, b=5b'+r$ and so $a-b=5(a'-b')$).

A similar example is the following:

Example 2.5.2. *Let $n\geq 2$ be an integer. In any group of n people, there are two that have the same number of friends within the group.*

We prove that by induction on n. The case $n=2$ is trivial. Let $N\geq 2$; suppose the claim for all $n\leq N$, and consider a group of $N+1$ people. If there is a person with 0 friends, we can look at the rest of the people. This is a group of N people and the number of friends each has in this smaller group is the same as in the original group. We can apply induction to conclude that two have the same number of friends (initially in the smaller group, but in fact also in the larger group).

The other case is when each person in the group of $N+1$ people has at least 1 friend. By considering the number of friends each person in the group has, we get $N+1$ integer values between 1 and N and so, by the pigeonhole principle, two must be equal.

Here is another way to formulate the statement we have just proven. A **graph** is a collection of vertices and edges connecting them. We say that a graph is **simple** if it has no loops and no multiple edges, namely, an edge always goes from a vertex to a different vertex and between any two vertices there is at most one edge. A graph is called **finite** if it has finitely many vertices. The degree of a vertex v is the number of edges one of whose terminal points is v. What we proved is that in a finite simple graph two of the vertices must have the same degree. (Given a party, create a vertex for every person and connect two vertices if the corresponding persons are friends.)

Finally, notice that we have used a variant of Induction called "complete induction". Namely, we used the following principle. *Suppose that a statement P_n is given for every natural number greater or equal to a natural number n_0. Suppose that the statement P_{n_0} is true and that for every $n\geq n_0$, if all the statments $P_{n_0}, P_{n_0+1},\dots,P_n$ are true then so is P_{n+1}. Then P_n is true for every $n\geq n_0$.*

The proof is more or less the same as the proof of Theorem 2.3.1, so we leave it as an exercise.[9]

3. Functions

We are familiar with functions already from high-school, where functions were usually, probably always, functions of a real variable. For example, the function $y=\sin(x)$ or $y=3x+1$. Here $x\in\mathbb{R}$ and the value y is in $\mathbb{R}$ too. We are also used to plotting this function using cartesian coordinates as all the pairs (x,y) where $y=\sin(x)$ or $3x+1$, as the case may be. We want to abstract this notion and for a start we may say that we have a function $f\colon \mathbb{R}\to\mathbb{R}$, where for the first example $f(x)=\sin(x)$ and for the second $f(x)=3x+1$. That is, instead of writing y we write $f(x)$. In this language, the graph are all the points in the plane of the form $(x,f(x))$, or, what is the same, all the points (x,y) such that $y=f(x)$.

There are more formal and less formal ways to define a function. Here we take the most pedestrian approach. Let A and B be sets. A **function** f from A to B,

$$f\colon A\longrightarrow B,$$

[9] You should not take that as meaning that you can safely ignore the matter of finding a proof for complete induction. On the contrary, it means that you should understand the proof of Theorem 2.3.1 so well that it is clear to you how to prove the variant given here.

is a rule assigning to *each* element of A a *single* element of B. The set A is called the **source**, or the **domain**, of the function, and B the **target**, or **codomain**, of the function. For $a \in A$, $f(a)$ is called the **image** of a (under f) and $f(A) = \{f(a) : a \in A\}$ is the **image** of f.

Example 3.0.1. The simplest example is the **identity function**. Let A be any set and define

$$1_A \colon A \to A$$

to be the function sending each element to itself. Namely,

$$1_A(x) = x,$$

for any $x \in A$.

Example 3.0.2. Let $A = \{1,2,3\}, B = \{1,2\}$ and consider the following rules for $f \colon A \longrightarrow B$.

(1) $f(1) = 2, \quad f(2) = 1, \quad f(3) = 1.$
(2) $f(1) = 1$ or $2, \quad f(2) = 2, \quad f(3) = 1.$
(3) $f(1) = 1, \quad f(2) = 1.$

The first recipe defines a function from A to B. The second recipe does not, because 1 is assigned two possible values. The third also doesn't define a function because no information is given about $f(3)$.

Example 3.0.3. Let $\mathbb{R}_{\geq 0}$ denote the non-negative real numbers. Consider the following attempts to define functions.

(1) $f \colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \sqrt{x}.$
(2) $f \colon \mathbb{R}_{\geq 0} \to \mathbb{R}, \quad f(x) = y,$ where y is a real number such that $y^2 = x$.
(3) $f \colon \mathbb{R}_{\geq 0} \to \mathbb{R}, \quad f(x) =$ the non negative root of $x = +\sqrt{x}$.
(4) $f \colon \mathbb{R} \to \mathbb{R}$, $f(x) = 1/x$.

The first definition fails because -1 doesn't have a root in $\mathbb{R}$. The second definition fails because every positive number has 2 roots (differing by a sign) and it isn't clear which root one is supposed to take. This problem also exists in the first definition. The third definition does define a function. The fourth definition doesn't define a function, because the value $f(0)$ is not well-defined.

There are various ways to define a function. It can be done by writing down $f(a)$ for every $a \in A$ explicitly, it can be done by providing a formula, and it can be done by giving some other description. For example, take A to be the set of all people who ever lived, $B = A$ and $f : A \to A$ is given by

$$f(a) = a\text{'s mother.}$$

This definitely looks like a good definition at first sight. However, the astute reader will note the problem here. If this function was truly well-defined then the set A must be infinite, because if it were finite we would have a person who's a descendant of itself (consider a, $f(a)$, $f(f(a))$, $f(f(f(a)))$, ...). Since a mother is older than any of her children by at least a day, say (just not to worry about an exact number), it follows that if A and f are indeed well defined, that people have existed forever. The various ways to resolve the paradox, namely to provide an explanation as to **why** f is not well-defined, are rather amazing and I leave it to you as an amusing exercise. (And, no, I don't view that as a proof that God exists.)

Here is some more notation: the symbol $\forall$ means "for all". The symbol $\exists$ means "exists". The symbol $\exists!$ means "exists unique". A function can also be defined by using a set

$$\Gamma \subset A \times B,$$

with the following property: $\forall a \in A, \exists! b \in B$ such that $(a,b) \in \Gamma$. (Read: for all a in A there exists a unique b in B such that (a,b) is in Γ.) We then define $f(a)$ to be the unique b such that $(a,b) \in \Gamma$. Conversely, given a function f we let

$$\Gamma = \Gamma_f = \{(a, f(a)) : a \in A\}.$$

The set Γ_f is called the **graph** of f.

Example 3.0.4. Let A be a set and $\Gamma \subset A \times A$ the "diagonal",

$$\Gamma = \{(x,x) : x \in A\}.$$

The function defined by Γ is 1_A.

Let $f_1, f_2 : A \to B$ be functions. We say that $f_1 = f_2$ if for every $a \in A$ we have $f_1(a) = f_2(a)$. Equivalently (exercise!) if $\Gamma_{f_1} = \Gamma_{f_2}$. Note here that the description of f_1 and f_2 doesn't matter, only whether $f_1(a) = f_2(a)$ for all $a \in A$. For example, let $f_1 : \mathbb{R} \to \mathbb{R}$ be the constant function $f_1(x) = 1$ for all x. Let $f_2 : \mathbb{R} \to \mathbb{R}$ be the function $f_2(x) = \sin(x)^2 + \cos(x)^2$. Then $f_1 = f_2$, although they were defined differently.

3.1. **Injective, surjective, bijective and inverse image.** We introduce some properties of functions. Let

$$f\colon A \to B$$

be a function. Then:

(1) f is called **injective** if $f(a) = f(a') \Rightarrow a = a'$. I.e., different elements of A go to different elements of B. Such a function is also called **one-one**.
(2) f is called **surjective**, or **onto**, if $\forall b \in B, \exists a \in A$ such that $f(a) = b$. I.e., every element in the target is the image of some element in the source under the function f.
(3) f is called **bijective** if it is both injective and surjective. In that case, every element of B is the image of a unique element of A.

Let $f\colon A \to B$ be a function. Let $U \subset B$. We define the **pre-image** of U to be the set

$$f^{-1}(U) = \{a : a \in A, f(a) \in U\}.$$

If U consists of a single element, $U = \{u\}$, we usually write $f^{-1}(u)$ (instead of $f^{-1}(\{u\})$) and call it the **fibre of f over u.**

Example 3.1.1. (1) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Then f is neither surjective (a square is always non-negative) nor injective as $f(x) = f(-x)$. We have $f^{-1}([1,4]) = [1,2] \cup [-2,-1]$ and $f^{-1}(0) = \{0\}, f^{-1}(-1) = \emptyset$.
(2) $f\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, $f(x) = x^2$. Then f is surjective but not injective.
(3) $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $f(x) = x^2$. Then f is bijective.

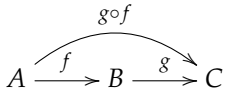
3.2. **Composition of functions.** Let

$$f\colon A \to B, \qquad g\colon B \to C,$$

be functions. We define their **composition**, $g \circ f$, to be the function:

$$g \circ f\colon A \to C, \qquad (g \circ f)(x) = g(f(x)).$$

The picture is



Lemma 3.2.1. *We have the following properties:*

(1) If $g \circ f$ is injective then f is injective.
(2) If $g \circ f$ is surjective then g is surjective.

Proof. Suppose that $g \circ f$ is injective. Let $a, a' \in A$ be elements such that $f(a) = f(a')$. We need to show that $a = a'$. We have $g(f(a)) = g(f(a'))$ or otherwise said, $(g \circ f)(a) = (g \circ f)(a')$. Since $g \circ f$ is injective, $a = a'$.

Suppose now that $g \circ f$ is surjective. Let $c \in C$. We need to show that there is an element $b \in B$ with $g(b) = c$. Since $g \circ f$ is surjective, there is $a \in A$ such that $(g \circ f)(a) = c$. Let $b = f(a)$ then $g(b) = g(f(a)) = (g \circ f)(a) = c$. □

A simple, but important property of composition is that it is associative: if $f\colon A \to B, g\colon B \to C, h\colon C \to D$ are functions then

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

3.3. **The inverse function.** Let $f\colon A \to B$ be a bijective function. In this case we can define the **inverse function**

$$f^{-1}\colon B \to A,$$

by the property

$$f^{-1}(b) = a \text{ if } f(a) = b.$$

This is well defined: since f is surjective such an a exists for every b and is unique (because f is injective). Thus f^{-1} is a function. It is easy to verify that

$$f^{-1} \circ f = 1_A, \qquad f \circ f^{-1} = 1_B.$$

To tie it up with previous definitions, note that if we let

$$\varphi : A \times B \to B \times A, \qquad \varphi(a,b) = (b,a),$$

then

$$\varphi(\Gamma_f) = \Gamma_{f^{-1}}.$$

In fact, this gives another way of defining f^{-1}.

4. Cardinality of a set

Imagine a group of students about to enter a lecture hall. The instructor wants to know if there are sufficient chairs for all the students. There are two ways to do that. One is to count both students and chairs separately and determine which is larger. The other is to ask each student to take a seat. If there are students left standing, the number of chairs is too small. If there are chairs left unoccupied, there are more chairs than students. In the remaining case there is a perfect match between students and chairs and so their number (cardinality) is equal.

This idea proves very powerful in discussing the cardinality ("size", "magnitude", "number of elements") of sets, finite or infinite.

George Cantor revolutionized mathematics,[10] and human thought, by defining two sets A, B **possibly infinite** to be of **equal cardinality**, denoted $|A| = |B|$, if there is a bijective function

$$f : A \to B.$$

[10] Cantor's own words are as fresh as ever; the following are the opening paragraphs of Cantor's paper whose title is translated into English as "Contributions to the founding of the theory of transfinite numbers". The translation is based on the Dover edition, but I have replaced "aggregate", "uniting" etc. by the modern terminology "set", "union" and so on, and the use of quotation marks by italics.

"By a *set* we are to understand any collection into a whole M of definite and separate objects m of our intuition or our thought. These objects are called the *elements* of M. In signs we express this thus:

(1) $$M = \{m\}.$$

We denote the union of many sets M, N, P, ..., which have no common elements, into a single set by

(2) $$(M, N, P, \dots).$$

The elements of this set are, therefore, the elements of M, of N, of P, ..., taken together. [....]
Every set M has a definite *power*, which we will also call its *cardinal number*. We will call by the name *power* or *cardinal number* of M the general concept which, by means of our active faculty of thought, arises from the set M when we make abstraction of the nature of its various elements m and of the order in which they are given."

What is striking to me is that Cantor's investigation is intimately tied with introspective meditation on the nature of thought and abstraction. The notion of cardinal number is not claimed to be absolute; it arises as a result of our thought process. Note the effort in trying to explain what cardinality *is*. In current times such efforts are rare. One simply says when two sets have the same cardinality, as we have done, dodging the issue of what that cardinality *is*.
Cantor suffered vicious personal attacks in reaction to his theory. Leopold Kronecker's public opposition and personal attacks included describing Cantor as a "scientific charlatan", a "renegade" and a "corrupter of youth." On the other hand, David Hilbert defended Cantor by declaring "No one shall expel us from the Paradise that Cantor has created." However, this came too late for Cantor, he was already dead by then.

(Note that then there is an inverse function $f^{-1}: B \to A$ which is also a bijection, so it doesn't matter if we require a bijection from A to B or from B to A.) He defined the cardinality of A to be no larger than B's if there is an injective function

$$f: A \to B,$$

and this is denoted $|A| \leq |B|$. We also say that the cardinality of A is less than B's, $|A| < |B|$, if $|A| \leq |B|$ and $|A| \neq |B|$. As a sanity check we'd like to know at least the following.

Proposition 4.0.1. *Let A, B, C be sets. If $|A| = |B|$ and $|B| = |C|$ then $|A| = |C|$.*

Proof. Let $f: A \to B, g: B \to C$ be bijections. Then

$$g \circ f: A \to C$$

is also a bijection. Indeed: if for some $x, y \in A$ we have $(g \circ f)(x) = (g \circ f)(y)$ then $g(f(x)) = g(f(y))$. Since g is injective, $f(x) = f(y)$ and, since f is injective, $x = y$.

To show $g \circ f$ is surjective, let $c \in C$ and choose $b \in B$ such that $g(b) = c$; such b exists since g is surjective. Since f is surjective, there is an $a \in A$ such that $f(a) = b$. Then $(g \circ f)(a) = g(f(a)) = g(b) = c$. □

To show the definitions and notations make sense at all, we surely need to know the following theorem.

Theorem 4.0.2 (Cantor-Bernstein)**.** *If $|A| \leq |B|$ and $|B| \leq |A|$ then $|A| = |B|$.*

The proof is by an ingenious and intricate argument, but requires no more background than we already have. We give it in Appendix A. We would like to explain though why the theorem is not obvious.

What we are given that there is *some* injective function f from A to B and *some* injective function g from B to A. Those functions need not be related to each other in any way. We should conclude from that there is a bijective function h from A to B. It is not necessarily true that $h = f$. One should somehow construct h from f and g. Here is an example: let A be the set of points in the plane in distance at most 1 from the origin (the closed unit disk) and B the square $[-1,1] \times [-1,1]$. The function

$$f: A \to B, \qquad f(x) = x,$$

is a well-defined injective function but not a bijection. The function

$$g: B \to A, \qquad g(b) = b/\sqrt{2},$$

is also a well-defined injective function, but not bijection. One can find a bijection from A to B, but it is not immediately clear how to find it based on the knowledge of f and g (and in fact in this particular example, it is better to "rethink the situation" rather than to deduce it from f and g).

Remark 4.0.3. By definition, $|A| \leq |B|$ if there is an injective function $f: A \to B$. We show here that if A is not the empty set, this is equivalent to the existence of a surjective function $g: B \to A$. Indeed, given f injective, choose some element $a_0 \in A$ and define $g: B \to A$ as follows: if $b = f(a)$ for some $a \in A$ then $g(b) = a$; otherwise, let $g(b) = a_0$. Then g is well defined as in the case where $b = f(a)$ the a is unique because f is injective. The function g is surjective, because given some $a \in A$ we find that $g(f(a)) = a$ and so every a is the image of some $b \in B$.

Conversely, given a surjective function $g: B \to A$, define $f: A \to B$ by choosing for every $a \in A$ some element $b \in B$ such that $g(b) = a$. There could be more than one b such that $g(b) = a$, but we just choose one of them and let $f(a) = b$. We get this way a function $f: A \to B$. We claim that it is injective. Indeed, if $f(a_1) = f(a_2)$ then $g(f(a_1)) = g(f(a_2))$. But by the construction of f, $g(f(a_1)) = a_1$ and $g(f(a_2)) = a_2$. Therefore, $a_1 = a_2$ and so f is injective.

A set A is called **countable** (or **enumerable**) if it is either finite or has the same cardinality as $\mathbb{N}$. If it is finite there is a bijective function $f: \{0, 1, \dots, n-1\} \to A$, where n is the number of elements of A and so $A = \{f(0), f(1), \dots, f(n-1)\}$. If A is infinite, there is a bijective function $f: \mathbb{N} \to A$. The elements of A are thus $\{f(0), f(1), f(2), f(3), \dots\}$. If we introduce the notation $a_i = f(i)$ then we can also enumerate the elements of A as $\{a_0, a_1, a_2, a_3, \dots\}$ and this explains the terminology.

Example 4.0.4. Let A be the set $\{0,2,4,6,\dots\}$ and B the set $\{0,1,4,9,16,25,\dots\}$. Then

$$|\mathbb{N}| = |A| = |B|.$$

Indeed, one verifies that the functions

$$f : \mathbb{N} \to A, \qquad f(x) = 2x,$$

and

$$g : \mathbb{N} \to B, \qquad g(x) = x^2,$$

are bijections. We can then conclude that $|A| = |B|$ and, if we want to, we can find the bijection. It is $h = g \circ f^{-1}$; that is, $h(x) = (x/2)^2$.

Example 4.0.5. The cardinality of $\mathbb{N}$ is the cardinality of $A = \{x \in \mathbb{N} : x \text{ is not a square}\}$.

Instead of trying to write a bijection explicitly, which is not that straight-forward, we use the Cantor-Bernstein theorem. It is important to digest the following argument!

The function

$$f : A \to \mathbb{N}, \qquad f(a) = a,$$

is injective. Thus, $|A| \leq |\mathbb{N}|$. Consider the function

$$f : \mathbb{N} \to A, \qquad f(x) = 4x + 3.$$

First, this is well defined. Namely, $4x+3$ is really in A. This is because if n is a square, n leaves residue 1 or 0 when divided by 4 (if $n = (2m)^2 = 4m^2$ the residue is zero; if $n = (2m+1)^2 = 4(m^2+m)+1$ the residue is one). But $4x+3$ leaves residue 3. Clearly f is injective and so $|\mathbb{N}| \leq |A|$.

Proposition 4.0.6. $|\mathbb{N}| = |\mathbb{Z}|$.

Proof. We define

$$f : \mathbb{Z} \to \mathbb{N}, \qquad g(x) = \begin{cases} 2x & x \geq 0 \\ -2x - 1 & x < 0. \end{cases}$$

Then f is a bijective function, as is easy to check. □

Proposition 4.0.7. $|\mathbb{N}| = |\mathbb{N} \times \mathbb{N}|$.

Proof. Define $f\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ by $f(n) = (n, 0)$. This is an injective function. Define

$$g\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}, \qquad g(n,m) = 2^n 3^m.$$

This is also an injective function. If $2^n 3^m = 2^a 3^b$ then $n = a, m = b$ by unique factorization (to be discussed in § 10). We conclude that $|\mathbb{N}| = |\mathbb{N} \times \mathbb{N}|$. □

Corollary 4.0.8. $|\mathbb{Z}| = |\mathbb{Z} \times \mathbb{Z}|$.

Proof. Let $h\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ be a bijection. A bijection $f\colon \mathbb{N} \to \mathbb{Z}$ induces a bijection

$$g = (f, f)\colon \mathbb{N} \times \mathbb{N} \to \mathbb{Z} \times \mathbb{Z},$$

and the composition

$$\mathbb{Z} \xrightarrow{f^{-1}} \mathbb{N} \xrightarrow{h} \mathbb{N} \times \mathbb{N} \xrightarrow{g} \mathbb{Z} \times \mathbb{Z}\ ,$$

is also a bijection. □

Exercise 4.0.9. Prove that $|\mathbb{N}| = |\mathbb{Q}|$. (Hint: there's an easy injection $\mathbb{Q} \to \mathbb{Z} \times \mathbb{Z}$).

When Cantor has laid down the foundations for the study of infinite cardinals he also dropped a bombshell:

4.1. **Cantor's diagonal argument.**

Theorem 4.1.1 (Cantor)**.** $|\mathbb{N}| \neq |\mathbb{R}|$.

The argument in the proof became known as **Cantor's diagonal argument** and is used in many proofs.

Proof. Suppose that $|\mathbb{N}| = |\mathbb{R}|$. We can then enumerate the real numbers as $a_0, a_1, a_2, \ldots$. Let us write the decimal expression of each number as

$$\begin{aligned}
a_0 &= \epsilon_0 b_1^0 \ldots b_{n(0)}^0 . c_0^0 c_1^0 c_2^0 c_3^0 \ldots \\
a_1 &= \epsilon_1 b_1^1 \ldots b_{n(1)}^1 . c_0^1 c_1^1 c_2^1 c_3^1 \ldots \\
a_2 &= \epsilon_2 b_1^2 \ldots b_{n(2)}^2 . c_0^2 c_1^2 c_2^2 c_3^2 \ldots \\
a_3 &= \epsilon_3 b_1^3 \ldots b_{n(3)}^3 . c_0^3 c_1^3 c_2^3 c_3^3 \ldots \\
&\vdots
\end{aligned}$$

where we agree to use $000000000\ldots$ instead of $999999999\ldots$ (so we write $1.00000000000\ldots$ and not $0.9999999999\ldots$, etc.). Here ϵ_i is a sign, $+$ or $-$, and each b_j^i, c_j^i is a digit, i.e. in $\{0, 1, \ldots, 9\}$.

Now consider the number

$$0.e_0 e_1 e_2 e_3 e_4 \ldots, \qquad e_i = \begin{cases} 3 & c_i^i \neq 3 \\ 4 & c_i^i = 3. \end{cases}$$

This is a real number that differs from each a_i at the $i+1$-th digit after the decimal dot and hence is not equal to any a_i. It follows that the list $a_0, a_1, a_2, \ldots$ cannot consist of **all** real numbers and so we arrive at a contradiction. □

Remark 4.1.2. The **Continuum hypothesis**, formulated by George Cantor, asserts that there is no set A such that $|\mathbb{N}| < |A| < |\mathbb{R}|$. Much later, Kurt Godel and Paul Cohen proved that neither the hypothesis, nor its negation, can be proven from the standard axioms of set theory. That is, in the axiom system we are using in mathematics, we cannot prove the hypothesis or provide a counter-example. These discoveries were nothing short of shocking. They came at a time where humanity was fascinated with its own power, especially the power of the intellect. These results proved the inherent limitations of thought.

One may ask at this point if there is a cardinality bigger than that of $\mathbb{R}$. That is, is there a set T such that $|\mathbb{R}| < |T|$. The answer to that is yes. One may take T to be the set of all subsets of $\mathbb{R}$. This is discussed in the exercises.

5. Relations

A **relation** on a set S is best described as a subset $\Gamma \subset S \times S$. For each $s \in S$, s is "related" to t if $(s,t) \in \Gamma$. Though the format reminds one of functions, the actual relevance of the notion of functions here is minimal. For example, usually for a given s there will be many elements t such that $(s,t) \in \Gamma$, which is the opposite of what we require for functions, where there is precisely one t for a given s. We shall usually denote that x **is related to** y, namely that $(x,y) \in \Gamma$, by $x \sim y$.

Note that so far the definition is wide enough to allow any Γ. Here are same basic examples:

(1) $\Gamma = S \times S$. In this case for any x, y, we have $x \sim y$. Any two elements are related.
(2) $\Gamma = \emptyset$. In this case for no x, y we have $x \sim y$. No elements are related (including an element with itself).
(3) $\Gamma = \{(s,s) : s \in S\}$ (the "diagonal"). In this case $x \sim x$ for all $x \in S$, but if $x \neq y$ then $x \not\sim y$.
(4) $\Gamma = \{(x,y) : x, y \in S, x \leq y\}$, where S is the interval of real numbers $[0,1]$. Then to say that $x \sim y$ means that $x \leq y$, in the sense of the usual inequality of real numbers.
(5) $\Gamma = \{(x,y) : x, y \in \mathbb{Z}, 5|(x-y)\}$ is the relation on $\mathbb{Z}$ where $x \sim y$ if $(x-y)$ is divisible by 5.

A relation on S is called **reflexive** if for all $x \in S$ we have $x \sim x$. In words: every element is related to itself. The relations in (1), (3), (4) and (5) are reflexive, but the relation in (2) is not (except in the trivial case where S is the empty set).

A relation is called **symmetric** if for all $x, y \in S$, if $x \sim y \Rightarrow y \sim x$. In words, whenever x is related to y also y is related to x. The relations (1), (2), (3), (5) are symmetric, but (4) is not.

A relation is called **transitive** for all $x, y, z \in S$, if $x \sim y$ and $y \sim z$ implies $x \sim z$. All the relations (1) - (5) are transitive.

A relation on a set S is called a **partial order** if it is reflexive and transitive and, in addition, if both $x \sim y$ and $y \sim x$ then $x = y$. We then use the notation $x \leq y$ for $x \sim y$ and, in this notation, we always have: $x \leq x$, the implication $x \leq y, y \leq z \Rightarrow x \leq z$ holds, and if both $x \leq y$ and $y \leq x$ then $x = y$. There may very well be x, y for which neither $x \leq y$ nor $y \leq x$ holds. A **linear order** (or a **simple order** or a **total order**) is a partial order such that for every x, y we have either $x \leq y$ or $y \leq x$.

For example, for real numbers we have the relation $\leq$ of "less or equal than". That is, for real numbers a, b we have the notion of $a \leq b$ and this is a linear ordering. It is reflexive ($a \leq a$), satisfies transitivity ($a \leq b$ and $b \leq c$ implies that $a \leq c$) and if both $a \leq b$ and $b \leq a$ then $a = b$. We can put a relation on the positive natural numbers $\mathbb{N}$, by saying that $a \sim b$ if $a|b$. Since we have $a|b$ and $b|c$ implies $a|c$ this relation is transitive. Also, if $a|b$ and $b|a$ then $a = b$ and, clearly, always $a|a$. Thus, this is a partial ordering of the natural numbers. Note that for $a = 2, b = 3$ neither $a|b$ nor $b|a$. That is, this is not a linear order.

Another important class of relations, even more important for this course than order relations, are the equivalence relations. They are very far from order relations. A relation is called an **equivalence relation** if it satisfies the following properties:

(1) (Reflexive) For every x we have $x \sim x$.
(2) (Symmetric) If $x \sim y$ then $y \sim x$.
(3) (Transitive) If $x \sim y$ and $y \sim z$ then $x \sim z$.

Equivalence relations arise when one wishes to *identify* elements in a given set R according to some principle. If we reflect on the meaning of "identify" we see that what we aim at is that x is identified with x (well, obviously!), that if x is identified with y then y is identified with x, and that if x is identified with y and y is identified with z then, by all accounts, x should be identified with z too. That is, we aim at an equivalence relation. And conversely, an equivalence relation is a way to identify elements in a set. We identify x with all the elements y such that $x \sim y$.

Example 5.0.1. (1) A **permutation** σ of a set A is a bijection $\sigma\colon A \to A$. A permutation of n elements is a bijection

$$\sigma\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}.$$

We denote S_n the set of all permutations σ of n elements. There are $n!$ permutations, that is, $|S_n| = n!$. Define a relation on S_n by saying for two permutations σ, τ that

$$\sigma \sim \tau \text{ if } \sigma(1) = \tau(1).$$

This is an equivalence relation.

(2) Here is a simple example, that will be generalized in §12. Define a relation on the integers $\mathbb{Z}$ by saying that $a \sim b$ if $a - b$ is even. This is an equivalence relation as $a - a = 0$ is always even, if $a - b$ is even then so is $b - a$ and if both $a - b$ and $b - c$ are even then so is $a - c = (a - b) + (b - c)$.

(3) Define a relation on sets[11] by saying that

$$A \sim B \text{ if } |A| = |B|.$$

This is an equivalence relation. The identity function $A \to A$ shows $A \sim A$. If $A \sim B$ there is a bijection $f\colon A \to B$ and then the inverse function $f^{-1}\colon B \to A$ is a bijection too, showing $B \sim A$. Finally, if $|A| = |B|$ and $|B| = |C|$ then $|A| = |C|$, by Proposition 4.0.1.

[11] We ignore here the fact that the collection of all sets is not a set and we only defined relations on a set. The example would be correct if we restricted ourselves to all sets that are subsets of some fixed given set.

Let S be a set and $S_i, i \in I$, be subsets of S. We say that S is the **disjoint union** of the sets S_i if $S = \cup_{i\in I} S_i$ and for any $i \neq j$ we have $S_i \cap S_j = \emptyset$. We denote this by

$$S = \coprod_{i\in I} S_i.$$

Lemma 5.0.2. *Let $\sim$ be an equivalence relation on a set S. Define the* **equivalence class** $[x]$ *of an element $x \in S$ as follows:*

$$[x] = \{y : y \in S, x \sim y\}.$$

This is a subset of S. The following holds:

(1) Two equivalence classes are either disjoint or equal.
(2) S is a disjoint union of equivalence classes.

Conversely, if S is a disjoint union $S = \coprod_{i\in I} U_i$ of non-empty sets U_i (this is called a **partition** *of S) then there is a unique equivalence relation on S for which the U_i are the equivalence classes.*

Proof. Let x, y be elements of S and suppose that $[x] \cap [y] \neq \emptyset$. Then, there is an element z such that $x \sim z, y \sim z$. Since $\sim$ is symmetric also $z \sim y$ and using transitivity $x \sim y$. Now, if $s \in [y]$ then $y \sim s$ and by transitivity $x \sim s$ and so $s \in [x]$ and we showed $[y] \subset [x]$. Since $x \sim y$ also $y \sim x$ and the same argument gives $[x] \subset [y]$. We conclude that $[x] = [y]$.

Every element of S lies in the equivalence class of itself. It follows that S is a disjoint union of equivalence classes.

To prove the second part of the lemma, we define that $x \sim y$ if both x and y lie in the same set U_i. It is clearly reflexive and symmetric. It is also transitive: $x \sim y$ means $x, y \in U_i$ for some i, $y \sim z$ means $y, z \in U_j$ for some j. But there is a unique U_i containing y because the union is a disjoint union. That is $U_i = U_j$ and so $x, z \in U_j$; that is, $x \sim z$. The equivalence classes are clearly the U_i. □

Example 5.0.3. We revisit Example 5.0.1 and find the equivalence classes. In the first case, there are n equivalence classes $[\sigma_1], \ldots, [\sigma_n]$ (each having $(n-1)!$ permutations in it), where we may choose σ_i to be the permutation:

$$\sigma_i(j) = \begin{cases} j & j \neq 1, i \\ 1 & j = i \\ i & j = 1 \end{cases}$$

That is, σ_i is the permutation sending 1 to i and i to 1 and leaving all other elements intact.

In the second case there are two equivalence classes, the even integers and the odd integers.

In the third case, the equivalence classes are the cardinalities, per definition (namely, one could say that a cardinality is an equivalence class of sets under the relation of existence of bijection). Some distinct equivalence classes are "all sets of cardinality 3" (that is, all sets that are in bijection with $\{1, 2, 3\}$, which are precisely the sets of 3 elements), "all sets of cardinality $\aleph_0$" (that is, all sets that are in bijection with $\mathbb{N}$ and this includes $\mathbb{Z}$, $\mathbb{N} \times \mathbb{N}$, $\mathbb{Q}$, ...), "all sets of cardinality $2^{\aleph_0}$", which is the cardinality of the real numbers $\mathbb{R}$ (see also Exercise 12).

We introduce the following terminology: Let S be a set with an equivalence relation. We say that a subset $T \subseteq S$ is a **complete set of representatives** if $S = \coprod_{t\in T} [t]$. That is S is the disjoint union of the equivalence classes of the elements in T: every equivalence class $[x]$ in S is equal to an equivalence class $[t]$ for a unique element $t \in T$.

Sometimes we want to index the elements of T and so as a notational variant we say that a subset $\{x_i : i \in I\} \subseteq S$, I some index set, is a complete set of representatives if the equivalence classes $[x_i]$ are disjoint and $S = \cup_{i\in I} [x_i]$. This means that every equivalence class is of the form $[x_i]$ for a unique $i \in I$. That is, if $i \neq j$ then $[x_i] \neq [x_j]$ (and in fact $[x_i] \cap [x_j] = \emptyset$).

6. Number systems

Again we start with an apology of a sort. The formal discussion of number systems is a rather involved piece of mathematics. Our approach is pragmatic. We assume that at some level we all know what are integers and real numbers and no confusion shall arise there. We use those to define more complicated notions.

As we have already said, we denote the **natural numbers** $\mathbb{N}$ by

$$\mathbb{N} = \{0, 1, 2, \dots\}, \qquad \mathbb{N}^+ = \{1, 2, 3, \dots\}.$$

We also denote the **integers** by

$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}.$$

The **rational numbers** are the set

$$\mathbb{Q} = \left\{ \frac{a}{b} : a, b \in \mathbb{Z}, b \neq 0 \right\},$$

where we identify $\frac{a}{b}$ with $\frac{c}{d}$ if $ad = bc$. The **real numbers** $\mathbb{R}$ are the "points on the line". Each real number has a decimal expansion such as $0.19874526348\dots$ that may or may not repeat itself from some point on. For example:

$$1/3 = 0.3333333\dots, 1/2 = 0.5000000\dots, 1/7 = 0.142857142857142857142857142857\dots,$$

$$\pi = 3.14159265358979323846264338 3\dots.$$

It is a fact that a number is rational if and only if from some point on its decimal expansion becomes periodic. The length of the period of rational number can be explained. For example, the period of a number of the form $1/n$, where $n \geq 1$ is an integer not divisible by 2 and 5 is the following.

Consider the sequence $1, 10, 100, 1000, \dots$ and their residues upon division by n, say $r_0 = 1, r_1, r_2, \dots$. The first $d > 0$ such that $r_d = 1$ is the period of the decimal expansion of $1/n$. For example, for $n = 3$ we have $1, 10, 100, \dots$ that give residues $1, 1, 1, \dots$ when divided by 3, and indeed $1 = 0.3333\dots$ has period 1. On the other hand, the residues of $1, 10, 100, \dots$ when divided by 7 are $r_0 = 1, r_1 = 3, r_2 = 2, r_3 = 6, r_4 = 4, r_5 = 5, r_6 = 1$ which predicts a period of 6 for the decimal expansion of $1/7$ and indeed $1/7 = 0.\mathbf{142857}142857142857142857142857\dots$.

The **complex numbers** are defined as the set

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}.$$

Here i is a formal symbol. We can equally describe the complex numbers as points $(a, b) \in \mathbb{R}^2$, the plane. The function $f : \mathbb{C} \to \mathbb{R}^2, f(a + bi) = (a, b)$ is bijective. The x-axis are now called the **real axis** and the y-axis the **imaginary axis**. If $z = a + bi$ is a complex number a is called the **real part** of z and is denoted $\mathrm{Re}(z)$ and b is called the **imaginary part** of z and is denoted $\mathrm{Im}(z)$. We therefore have

$$z = \mathrm{Re}(z) + \mathrm{Im}(z)i.$$

The point corresponding to z in the plane model is $(\mathrm{Re}(z), \mathrm{Im}(z))$.

One can perform arithmetic operations with the complex numbers using the following definitions:

$$-(a + bi) = -a - bi, \quad (a + bi) + (c + di) = (a + c) + (b + d)i.$$

Up to this point things look nice in the plane model as well:

$$-(a, b) = (-a, -b), \quad (a, b) + (c, d) = (a + c, b + d).$$

That is to say, the addition is then just the addition rule for vectors. The key point is that we can also define *multiplication*. The definition doesn't have any priori interpretation in the plane model; it's a new operation. We let

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i.$$

In particular,

$$i^2 = -1.$$

This shows that we have really gone beyond the realm of real numbers because there is no real number whose square is -1. The operations described above satisfy the usual rules of arithmetic, such as

$$(z + z') + z'' = z + (z' + z''), \quad z(z' + z'') = zz' + zz'', \dots$$

(We shall later say that the complex numbers form a **field**.) In fact, given these rules and $i^2 = -1$, there is no need to memorize the formula for multiplication as it just follows by expansion: $(a+bi)(c+di) = ac + adi + bci + bdi^2 = ac - bd + adi + bci = ac - bd + (ad+bc)i$.

Let $z = a + bi$ be a complex number. We define the **complex conjugate** of z, $\bar{z}$, as follows:

$$\bar{z} = a - bi.$$

Lemma 6.0.1. *The complex conjugate has the following properties:*

(1) $\bar{\bar{z}} = z$.

(2) $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2, \quad \overline{z_1 \cdot z_2} = \bar{z}_1 \cdot \bar{z}_2$.

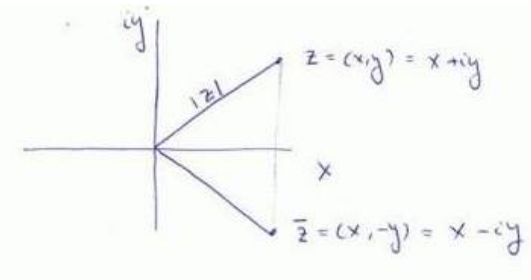
(3) $\mathrm{Re}(z) = \frac{z+\bar{z}}{2}, \quad \mathrm{Im}(z)i = \frac{z-\bar{z}}{2}$.

(4) Define for $z = a + bi$,

$$|z| = \sqrt{a^2 + b^2}.$$

(This is just the distance of the point (a,b) *from the origin.) Then* $|z|^2 = z \cdot \bar{z}$ *and the following holds:*

$$|z_1 + z_2| \le |z_1| + |z_2|, \qquad |z_1 \cdot z_2| = |z_1| \cdot |z_2|.$$



Proof. Denote $z = z_1 = a + bi, z_2 = c + di$.

We have $\bar{z} = a - bi$ and so $\bar{\bar{z}} = a - (-b)i = a + bi = z$. That is (1). For (2) we calculate

$$\begin{aligned}
\overline{z_1 + z_2} &= \overline{(a+c) + (b+d)i} \\
&= (a+c) - (b+d)i \\
&= a - bi + c - di \\
&= \overline{a+bi} + \overline{c+di} \\
&= \overline{z_1} + \overline{z_2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\overline{z_1 z_2} &= \overline{(ac - bd) + (ad + bc)i} \\
&= (ac - bd) - (ad + bc)i \\
&= (a - bi)(c - di) \\
&= \overline{a+bi} \cdot \overline{c+di} \\
&= \overline{z_1} \cdot \overline{z_2}.
\end{aligned}$$

We have $(z + \bar{z})/2 = ((a+bi) + (a-bi))/2 = a = \mathrm{Re}(z)$ and $(z - \bar{z})/2 = ((a+bi) - (a-bi))/2 = bi = \mathrm{Im}(z)i$, which is (3). Next, $|z|^2 = a^2 + b^2 = (a+bi)(a-bi) = z \cdot \bar{z}$. Now,

$$\begin{aligned}
|z_1 z_2|^2 &= z_1 z_2 \cdot \overline{z_1 z_2} \\
&= z_1 z_2 \cdot \overline{z_1} \cdot \overline{z_2} \\
&= z_1 \cdot \overline{z_1} \cdot z_2 \cdot \overline{z_2} \\
&= |z_1|^2 \cdot |z_2|^2.
\end{aligned}$$

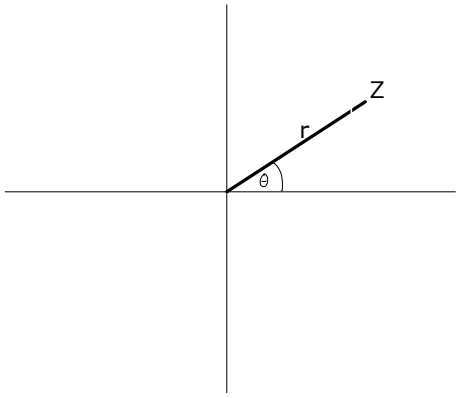
Thus, the assertion $|z_1 \cdot z_2| = |z_1| \cdot |z_2|$ follows by taking roots. The inequality $|z_1 + z_2| \leq |z_1| + |z_2|$ viewed in the plane model for complex numbers is precisely the assertion that the sum of the lengths of two sides of a triangle is greater or equal to the length of the third side. □

Example 6.0.2. If $z \neq 0$ then z has an inverse with respect to multiplication. Indeed, $z \cdot \frac{\bar{z}}{|z|^2} = \frac{z \cdot \bar{z}}{|z|^2} = 1$. We write

$$z^{-1} = \frac{\bar{z}}{|z|^2}.$$

Just to illustrate calculation with complex numbers, we calculate $1 + 2i + \frac{3-i}{1+5i}$. Using $z^{-1} = \bar{z}/|z|^2$, we have $\frac{1}{1+5i} = \frac{1-5i}{26}$ and so $\frac{3-i}{1+5i} = (3-i)(1-5i)/26 = -\frac{1}{13} - \frac{8}{13}i$ and thus $1 + 2i + \frac{3-i}{1+5i} = \frac{12}{13} + \frac{18}{13}i$.

6.1. **The polar representation.** Considering a complex number $z = a + bi$ in the plane model as the vector (a, b) we see that we can describe each complex number z by the length r of the corresponding vector (a, b) and the angle θ it forms with the real axis.



We have

$$r = |z|, \quad \sin\theta = \frac{\mathrm{Im}(z)}{|z|}, \quad \cos\theta = \frac{\mathrm{Re}(z)}{|z|}.$$

Lemma 6.1.1. *If z_1 has parameters r_1, θ_1 and z_2 has parameters r_2, θ_2 then $z_1 z_2$ has parameters $r_1 r_2, \theta_1 + \theta_2$ (up to multiples of 360° or 2π* **rad***).*

Proof. We have $r_1 r_2 = |z_1||z_2| = |z_1 z_2|$ and this shows that $r_1 r_2$ is the length of $z_1 z_2$. Let θ be the angle of $z_1 z_2$ then

$$\begin{aligned}
\sin\theta &= \frac{\mathrm{Im}(z_1 z_2)}{|z_1 z_2|} \\
&= \frac{\mathrm{Re}(z_1)\mathrm{Im}(z_2) + \mathrm{Re}(z_2)\mathrm{Im}(z_1)}{|z_1||z_2|} \\
&= \frac{\mathrm{Re}(z_1)}{|z_1|}\frac{\mathrm{Im}(z_2)}{|z_2|} + \frac{\mathrm{Re}(z_2)}{|z_2|}\frac{\mathrm{Im}(z_1)}{|z_1|} \\
&= \cos\theta_1 \sin\theta_2 + \cos\theta_2 \sin\theta_1 \\
&= \sin(\theta_1 + \theta_2).
\end{aligned}$$

Similarly, we get

$$\begin{aligned}\cos\theta &= \frac{\mathrm{Re}(z_1z_2)}{|z_1z_2|}\\ &= \frac{\mathrm{Re}(z_1)}{|z_1|}\frac{\mathrm{Re}(z_2)}{|z_2|} - \frac{\mathrm{Im}(z_1)}{|z_1|}\frac{\mathrm{Im}(z_2)}{|z_2|}\\ &= \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2)\\ &= \cos(\theta_1+\theta_2).\end{aligned}$$

It follows that $\theta = \theta_1 + \theta_2$ up to multiples of 360°. □

6.2. **The complex exponential function.** Let θ be any real number. Let $e^{i\theta}$ denote the unit vector whose angle is θ. Note that we *define* $e^{i\theta}$ this way. That is, $e^{i\theta}$ is *the* complex number of length 1 and angle θ. Clearly we have

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

If z is any complex number with length r and angle θ then we have the equality $z = |z|e^{i\theta}$. The formula we proved in Lemma 6.1.1 is

$$z_1z_2 = |z_1||z_2|e^{i(\theta_1+\theta_2)}$$

and in particular, if we take the complex numbers $e^{i\theta_1}, e^{i\theta_2}$ themselves, we find that

$$e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}.$$

Let $z = a + bi$ be a complex number then we define

$$e^z = e^a e^{ib},$$

where e^a is the usual exponential (a is a real number) and e^{ib} is as defined above. Combining the formulas for the real exponent with our definition for $e^{i\theta}$, we conclude that for any complex numbers z_1, z_2,

$$e^{z_1}e^{z_2} = e^{z_1+z_2}.$$

We have defined here e^z in a purely formal way. A more analytic approach, and in fact the "correct" approach, is the following:

We say that a sequence of complex numbers $z_1, z_2, \dots$ converges to a complex number A if

$$\lim_n |A - z_n| = 0.$$

In this sense, one can show that for every complex number z the series

$$1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots + \frac{z^n}{n!} + \dots$$

converges and is equal to e^z. This is well known to hold for z a real number, and so we see that our definition of e^z for z a complex number is a natural extension of the function e^z for z a real number. See Appendix D for more on the complex exponential function.

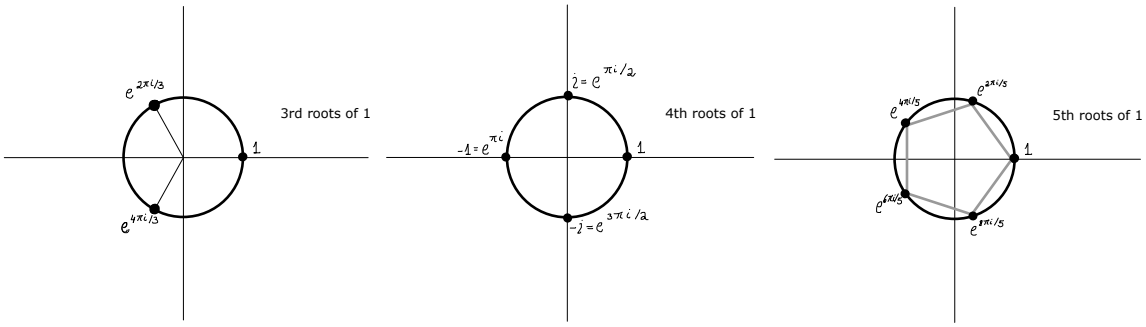
Example 6.2.1. Consider the polynomial $x^n - a = 0$, where a is a non-zero complex number. We claim that this equation has n distinct roots in $\mathbb{C}$. Write $a = re^{i\theta}$ (so $|a| = r$ and the line from 0 to a forms an angle θ with the real axis). A complex number $z = Re^{i\Theta}$ is a solution to the equation if and only if $z^n = R^n e^{in\Theta} = re^{i\theta}$. That is, if and only if

$$R^n = r, \qquad n\Theta \equiv \theta \pmod{2\pi}.$$

Thus, the solutions are exactly

$$z = r^{1/n}e^{i(\frac{\theta}{n} + j\cdot\frac{2\pi}{n})}, \quad j = 0, 1, \dots, n-1.$$

In particular, taking $a = 1$ the solutions are called the **roots of unity** of order n. There are precisely n of them: the points on the unit circle having angles $0, \frac{2\pi}{n}, 2\cdot\frac{2\pi}{n}, \dots, (n-1)\cdot\frac{2\pi}{n}$.



6.3. **The Fundamental Theorem of Algebra.** A **complex polynomial** $f(x)$ is an expression of the form

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

where n is a non-negative integer, x is a variable, and the coefficients a_i are complex numbers. If all the coefficients are real we may call it a **real polynomial**; if all the coefficients are rational numbers we may call it a **rational polynomial** and so on. But note that x^2+1 is both a rational, real and complex polynomial. The **zero polynomial**, denote 0, is the case when $n = 0$ and $a_0 = 0$.

A polynomial defines a function

$$f\colon \mathbb{C} \to \mathbb{C}, \quad z \mapsto f(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0.$$

The notation $\mapsto$ appearing in this formula means "maps to". If $a_n \neq 0$ then we say f has **degree** n. If $f(z) = 0$ for some particular complex number z, we say that z is a **root** (or a **solution**, or a **zero**) of the polynomial f.

Example 6.3.1. Consider the polynomial $f(x) = x^2+1$. It has degree 2 and $f(i) = i^2+1 = -1+1 = 0, f(-i) = (-i)^2+1 = -1+1 = 0$. So i and $-i$ are roots of f. This is a special case of Example 6.2.1, because $x^4-1 = (x^2+1)(x^2-1)$.

Theorem 6.3.2 (The Fundamental Theorem of Algebra)**.** *Let $f(x)$ be a complex polynomial of degree at least 1. Then $f(x)$ has a root in $\mathbb{C}$.*

Proofs of the theorem are beyond the scope of this course. There are many proofs. In the course *Honours Algebra 4* one sees an algebraic proof using Galois theory; in the course *Complex Variables and Transforms* one sees an analytic proof. The theorem is attributed to Gauss who proved it in 1799. There are some issues regarding the completeness of that proof; he later published several other proofs of the theorem. Gauss didn't discover the theorem, though; Many attempts and partial results were known before his work, and he was aware of that literature.

Proposition 6.3.3. *Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a complex non-zero polynomial of degree n. Then*

$$f(x) = a_n \prod_{i=1}^{n} (x - z_i),$$

for suitable complex numbers z_i, not necessarily distinct. The numbers z_i are all roots of f and any root of f is equal to some z_i. Moreover, this factorization is unique.

Proof. We prove the result by induction on n. For $n = 0$ we understand the product $\prod_{i=1}^{n}(x - z_i)$ as 1, by definition.[12] And so, the claim is just that a constant polynomial is equal to its leading coefficient. Clear.

Now, assume that f has degree at least one. By the Fundamental Theorem of Algebra there is a complex number z_n, say, such that $f(z_n) = 0$. We claim that for every complex number z we can write

$$f(x) = (x - z)g(x) + r,$$

[12] This is a convention: the empty product is equal to one, the empty sum is equal to zero.

where $g(x)$ is a polynomial of degree $n-1$ and leading coefficient a_n and r is a complex number. Indeed, write $g(x)=b_{n-1}x^{n-1}+\cdots+b_1x+b_0$ and equate coefficients in $(x-z)g(x)=b_{n-1}x^n+(b_{n-2}-zb_{n-1})x^{n-1}+\cdots+(b_0-zb_1)x$ and $f(x)$. We want complex numbers $b_0,\dots,b_{n-1}$ such that

$$b_{n-1}=a_n,\ (b_{n-2}-zb_{n-1})=a_{n-1},\dots,\ (b_0-zb_1)=a_1,$$

and there is no problem solving these equations. Thus, we can choose $g(x)$ with a leading coefficient a_n such that $f(x)-(x-z)g(x)=r$ is a constant. Note that $g(x)$ depends on z, but this is not reflected in our notation.

Now, apply that for the particular choice $z=z_n$. We have $f(x)-(x-z_n)g(x)=r$. We view r as a polynomial and substitute $x=z_n$. We get

$$f(z_n)-(z_n-z_n)g(z_n)=r.$$

Since $f(z_n)=0$ we conclude that $r=0$.

We showed that if $f(z_n)=0$ then

$$f(z_n)=(x-z_n)g(x),\quad g(x)=b_{n-1}x^{n-1}+\cdots+b_0.$$

In fact, $b_{n-1}=a_n$. Using the induction hypothesis, we have

$$g(x)=a_n\prod_{i=1}^{n-1}(x-z_i),$$

for some complex numbers z_i and so

$$f(x)=a_n\prod_{i=1}^{n}(x-z_i).$$

We note that $f(z_j)=a_n\prod_{i=1}^n(z_j-z_i)=0$, because the product contains the term (z_j-z_j). If $f(z)=0$ then $a_n\prod_{i=1}^n(z-z_i)=0$. But, if a product of complex numbers is zero one of the numbers is already zero (use $|z_1z_2\cdots z_a|=|z_1|\cdot|z_2|\cdot\ldots\cdot|z_a|$). Since $a_n\neq 0$, we must have $z=z_i$ for some i.

It remains to prove the uniqueness of the factorization. Suppose that

$$f(x)=a_n\prod_{i=1}^{n}(x-z_i)=a\prod_{i=1}^{n}(x-t_i).$$

Since the leading coefficient of f is a_n we must have $a=a_n$. We now argue by induction on the degree of f. The case of degree 0 is clear. Assume f has degree greater than zero. Then the t_i are roots of f and so t_1 is equal to some z_i, and we may re-index the z_i so that $t_1=z_1$. Dividing both sides by $x-z_1$ we then conclude that[13]

$$a_n\prod_{i=2}^{n}(x-z_i)=a_n\prod_{i=2}^{n}(x-t_i),$$

and, by induction, $z_i=t_i$ for all i. □

We remark that for $n=1,2$ the expression of f as a product is well-known from highschool:

$$ax+b=a\cdot\left(x-\frac{-b}{a}\right),$$

$$ax^2+bx+c=a\cdot\left(x-\frac{-b+\sqrt{b^2-4ac}}{2a}\right)\cdot\left(x-\frac{-b-\sqrt{b^2-4ac}}{2a}\right).$$

There are also formulas for the roots for polynomials of degree 3 and 4, but in degrees 5 and higher no such general formulas exist. Not that they are merely unknown; they cannot exist. This follows from Galois theory, taught in in the course Algebra 4.

The story of the solvability of polynomials is a fascinating story. One looks for formulas for solving polynomial equations that only involve the coefficients of the polynomials and elementary operations such as adding, subtracting, multiplying, dividing and taking n-th roots. This is called "solving by radicals". The formulas for solving linear and quadratic polynomials by radicals were known since antiquity: in some form or

[13] We say that $f(x)/g(x)=h(x)$ if $h(x)$ is a polynomial such that $f(x)=g(x)h(x)$. We shall see later that $h(x)$ is uniquely determined. In our case clearly $f(x)/(x-z_1)=a_n\prod_{i=2}^n(x-z_i)$.

another already to the Babylonians as early as 2000 BC and in a definite form to the Indian mathematician Brahmagupta in 628 AD. This knowledge propagated through Al-Khwarizmi and others to Europe, and one of the first European books containing such formulas was published in the 12^{th} century by the Spanish Jewish scholar Avraham bar Khiyya Ha-Nasi. The formulas for polynomials of degree 3 and 4 were published by the Italian mathematician Gerolamo Cardano in 1545 (the case of a quartic reduces to a cubic and was discovered by his student Ludovico Ferrari), and the case of cubic was also known to his contemporary and compatriot Niccolo Fontana, who went by the name Tartaglia. In fact, Tartaglia had told Cardano how to solve cubics but sworn him to complete secrecy. As it were, both were preceded, by a margin of three decades, by Scipione del Ferro that preferred taunting his colleagues with challenges for solving particular cubic equations than to publish his result. Cardano, after discovering del Ferro's work didn't consider himself bound by his promise to Tartaglia anymore and, indeed, in his book Ars Magna had attributed the solution of the cubic to de Ferro as well as acknowledging that Tartaglia had told him of the method, but not its proof. Cardano's publication didn't go well with Tartaglia and a bitter feud ensued. Indeed, the solutions to the cubic and quartic equations made Cardano and his book well-known through Europe and the scientific prestige easily translated into material benefits as well.

For a long time finding formula for equations of degree 5 was one the outstanding problems of mathematics and for a long time everyone was sure that such formulas must exist, although Gauss expressed some serious doubts. Niels Henrik Abel, a Norwegian mathematical genius who died in 1829 at the age of 27 (and after whom the Abel prize - the "Nobel prize for mathematics" - is named), proved that such formulas cannot exist. Independently, Évartise Galois who died at the age of 21 in 1832, proved not only the insolvability of the quintic, but in fact of all equations of any degree greater than 4, developing in the process the theory of groups - a theory that transformed algebra and number theory.

7. Fields and rings - definitions and first examples

In the examples of number systems we have already discussed there are implicit structures that we want to bring to light by defining them now in a formal way. At this point we just provide the definitions and reexamine previous examples. Later we shall enter a systematic development of the theory. The process we are about to carry out is very typical of how mathematics develops. Interesting structures are discovered (in this case arithmetic of integers, rational numbers, complex numbers, but later we shall also consider permutation groups). The typical (good) mathematician will ask herself what are the reasons that these structures are so rich, and proceed to distill the particular features responsible for that richness (in our case, we have "oprations" and certain identities between operations, or, for groups, we have "composition"). The next step, is to turn the table around and ask, now that we have these abstracted structures, what other examples may we find where these abstracted structures are manifest? In our context, these will be rings and fields, or, in the context of groups, these will abstract groups, no longer tied to permutations of objects, and we would want to know what properties are shared between all these examples.

An **operation** (more pedantically called a "binary operation") on a set R is a function

$$w : R \times R \to R.$$

That is, an operation is a rule taking two elements of R and returning a new one. For example:

$$w : \mathbb{C} \times \mathbb{C} \to \mathbb{C}, \quad w(z_1, z_2) = z_1 + z_2,$$

or

$$w : \mathbb{C} \times \mathbb{C} \to \mathbb{C}, \quad w(z_1, z_2) = z_1 z_2.$$

Often, for a general set R, we may denote $w(z_1, z_2)$ by $z_1 + z_2$, or $z_1 z_2$, if we want to stress the fact that the operation behaves like addition, or multiplication. We shall of course focus on mathematical examples, but one can certainly be more adventurous. For example, we can take R to be the set of sound waves and the operation w to be the juxtaposition of two sound waves.

Definition 7.0.1. A **ring** R is a non-empty set together with two operations, called "addition" and "multiplication" that are denoted, respectively, by

$$(x,y)\mapsto x+y,\qquad (x,y)\mapsto xy.$$

One requires the following axioms to hold:

(1) $x+y=y+x$, $\forall x,y\in R$. (Commutativity of addition)
(2) $(x+y)+z=x+(y+z)$, $\forall x,y,z\in R$. (Associativity of addition)
(3) There exists an element in R, denoted 0, such that $0+x=x$, $\forall x\in R$. (Neutral element for addition)
(4) $\forall x\in R$, $\exists y\in R$ such that $x+y=0$. (Inverse with respect to addition)
(5) $(xy)z=x(yz)$, $\forall x,y,z\in R$. (Associativity of multiplication)
(6) There exists an element $1\in R$ such that $1x=x1=x, \forall x\in R$. (Neutral element for multiplication)
(7) $z(x+y)=zx+zy, (x+y)z=xz+yz$, $\forall x,y,z\in R$. (Distributivity)

We remark that for us, **by definition**, a ring **always** has an identity element with respect to multiplication. Most, but not all, authors follow this convention.

Definition 7.0.2. Note that the multiplication operation is not assumed to be commutative in general. If $xy=yx$ for all $x,y\in R$, we say R is a **commutative ring**. If for every non-zero $x\in R$ there is an element $y\in R$ such that $xy=yx=1$, and also $0\neq 1$ in R, we call R a **division ring**. A commutative division ring is called a **field**.

Example 7.0.3. $\mathbb{Z}$ is a commutative ring. It is not a division ring and so it is not a field.

Example 7.0.4. The rational numbers $\mathbb{Q}$ form a field. The real numbers $\mathbb{R}$ form a field. In both cases we assume the properties of addition and multiplication as "well known". The complex numbers also form a field, in fact we have at some level already used all the axioms implicitly in our calculations, but now we prove it formally using that $\mathbb{R}$ is a field.

Proposition 7.0.5. *$\mathbb{C}$ is a field.*

Proof. Let $z_1=a_1+b_1i, z_2=a_2+b_2i, z_3=a_3+b_3i$. We verify the axioms:
1. $z_1+z_2=(a_1+a_2)+(b_1+b_2)i=(a_2+a_1)+(b_2+b_1)i=z_2+z_1$.
2. $(z_1+z_2)+z_3=[(a_1+a_2)+(b_1+b_2)i]+a_3+b_3i=[(a_1+a_2)+a_3]+[(b_1+b_2)+b_3]i=[a_1+(a_2+a_3)]+[b_1+(b_2+b_3)]i=z_1+[(a_2+a_3)+(b_2+b_3)i]=z_1+(z_2+z_3)$.
3. Clearly $0+z_1=z_1$.
4. We have $(-a_1-b_1i)+(a_1+b_1i)=(-a_1+a_1)+(-b_1+b_1)i=0+0i=0$.
5. $(z_1z_2)z_3=[(a_1+b_1i)(a_2+b_2i)](a_3+b_3i)=((a_1a_2-b_1b_2)+(a_1b_2+b_1a_2)i)(a_3+b_3i)=(a_1a_2-b_1b_2)a_3-(a_1b_2+b_1a_2)b_3+((a_1a_2-b_1b_2)b_3+(a_1b_2+b_1a_2)a_3)i=a_1a_2a_3-b_1b_2a_3-a_1b_2b_3-b_1a_2b_3+(a_1a_2b_3-b_1b_2b_3+a_1b_2a_3+b_1a_2a_3)i$. One now develops the product $z_1(z_2z_3)$ in the same way and checks that the answers match. We don't do that here.
6. Clearly $1\cdot z_1=z_1\cdot 1=z_1$.
7. $z_1(z_2+z_3)=(a_1+b_1i)((a_2+a_3)+(b_2+b_3)i)=a_1(a_2+a_3)-b_1(b_2+b_3)+(b_1(a_2+a_3)+a_1(b_2+b_3)i)=(a_1a_2-b_1b_2)+(b_1a_2+a_1b_2)i+(a_1a_3-b_1b_3)+(b_1a_3+a_1b_3)i=(a_1+b_1i)(a_2+b_2i)+(a_1+b_1i)(a_3+b_3i)=z_1z_2+z_1z_3$.

Before proving the next property, we check that $z_1z_2=z_2z_1$. We have $z_1z_2=(a_1+b_1i)\ (a_2+b_2i)=a_1a_2-b_1b_2+(a_1b_2+b_1a_2)i=a_2a_1-b_2b_1+(a_2b_1+b_2a_1)i=(a_2+b_2i)(a_1+b_1i)=z_2z_1$. In particular, $(z_1+z_2)z_3=z_3(z_1+z_2)=z_3z_1+z_3z_2=z_1z_3+z_2z_3$.

Finally, as we have already seen, if $z_1\neq 0$ then $z_1\cdot\frac{\bar{z}_1}{|z_1|^2}=1$. We proved that $\mathbb{C}$ is a field. □

Example 7.0.6. Here is an example of a non-commutative ring. The elements of this ring are 2-by-2 matrices with entries in $\mathbb{R}$. We define

$$\begin{pmatrix}a&b\\c&d\end{pmatrix}+\begin{pmatrix}\alpha&\beta\\\gamma&\delta\end{pmatrix}=\begin{pmatrix}a+\alpha&b+\beta\\c+\gamma&d+\delta\end{pmatrix},\qquad \begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}\alpha&\beta\\\gamma&\delta\end{pmatrix}=\begin{pmatrix}a\alpha+b\gamma&a\beta+b\delta\\c\alpha+d\gamma&c\beta+d\delta\end{pmatrix}.$$

The verification of the axioms is a straightforward, but tiresome, business. We shall not do it here. The zero element is $\begin{pmatrix}0&0\\0&0\end{pmatrix}$ and the identity element is $\begin{pmatrix}1&0\\0&1\end{pmatrix}$. To see that the ring is not commutative we give the following example.

$$\begin{pmatrix}0&1\\0&0\end{pmatrix}\begin{pmatrix}0&0\\1&0\end{pmatrix}=\begin{pmatrix}1&0\\0&0\end{pmatrix},\qquad \begin{pmatrix}0&0\\1&0\end{pmatrix}\begin{pmatrix}0&1\\0&0\end{pmatrix}=\begin{pmatrix}0&0\\0&1\end{pmatrix}.$$

7.1. **Some formal consequences of the axioms.** We note some useful formal consequences of the axioms defining a ring:

(1) **The element 0 appearing in axiom (3) is unique.** Indeed, if q is another element with the same property then $q+x=x$ for any x and in particular $q+0=0$. But also, using the property of 0 and commutativity, we have $q+0=0+q=q$. So $q=0$.

(2) **The element y appearing in axiom (4) is unique.** Indeed, if for a given x we have $x+y=x+y'=0$ then $y=y+(x+y')=(y+x)+y'=(x+y)+y'=0+y'=y'$. We shall denote this element y by $-x$.

(3) **We have $-(-x)=x$ and $-(x+y)=-x-y$**, where, technically $-x-y$ means $(-x)+(-y)$. To prove that, it is enough, after what we have just proven, to show that $-x+x=0$ and that $(x+y)+(-x-y)=0$ (after all, $-(-x)$ is that unique element that when added to $-x$ gives 0, etc.). The first is clear, and $(x+y)+(-x-y)=x+(-x)+y+(-y)=0+0=0$.

Although this proof is simple and short, it is based on an important idea, worth internalizing. We prove that an object is equal to another object by proving they both have a property known to uniquely determine the object. Namely, the proof that $-(-x)=x$ consists of showing that the objects x and $-(-x)$ both have the property that adding them to $-x$ gives zero. Since this is known to uniquely characterize the additive inverse of $-x$, denoted $-(-x)$, we conclude that $x=-(-x)$.

(4) **The element 1 in axiom (6) is unique.** (Use the same argument as in (1)).

(5) **We have $x\cdot 0=0, 0\cdot x=0$.** Indeed, $x\cdot 0=x\cdot(0+0)=x\cdot 0+x\cdot 0$. Let $y=x\cdot 0$ then $y=y+y$ and so $0=-y+y=-y+(y+y)=(-y+y)+y=0+y=y$.

We shall see many examples of rings and fields in the course. For now, we just give one more definition and some examples.

Definition 7.1.1. Let R be a ring. A subset $S\subset R$ is called a **subring** if $0,1\in S$ and if $a,b\in S$ implies that $a+b,-a$ and $ab\in S$.

Note that the definition says that the operations of addition and multiplication in R give operations of addition and multiplication in S (namely, the outcome is in S and so we get functions $S\times S\to S$), satisfying all the axioms of a ring. It follows that S is a ring whose zero element is that of R and whose identity is, likewise, that of R.

For example, $\mathbb{Q}$ is a subring of $\mathbb{R}$ and $\mathbb{Z}$ is a subring of $\mathbb{Q}$, as well of $\mathbb{R}$.

Example 7.1.2. Consider the set $\{0,1\}$ with the following addition and multiplication tables.

$+$	0	1
0	0	1
1	1	0

$\times$	0	1
0	0	0
1	0	1

One can verify that this is a ring by directly checking the axioms. (We shall later see that this is the ring of integers modulo 2).

Example 7.1.3. Consider all expressions of the form $\{a+b\sqrt{2}: a,b\in\mathbb{Z}\}$. We use the notation $\mathbb{Z}[\sqrt{2}]$ for this set. This set is actually a ring. Since $\mathbb{Z}[\sqrt{2}]\subset\mathbb{R}$ and $\mathbb{R}$ is a ring (even a field!), it is enough to check it's a subring. Indeed, $0,1\in\mathbb{Z}[\sqrt{2}]$. Suppose $a+b\sqrt{2}, c+d\sqrt{2}\in\mathbb{Z}[\sqrt{2}]$. Then: (1) $(a+b\sqrt{2})+(c+d\sqrt{2})=(a+c)+(b+d)\sqrt{2}\in\mathbb{Z}[\sqrt{2}]$; (2) $-(a+b\sqrt{2})=-a-b\sqrt{2}\in\mathbb{Z}[\sqrt{2}]$; (3) $(a+b\sqrt{2})(c+d\sqrt{2})=(ac+2bd)+(ad+bc)\sqrt{2}\in\mathbb{Z}[\sqrt{2}]$.

Let now

$$\mathbb{Q}[\sqrt{2}]=\{a+b\sqrt{2}: a,b\in\mathbb{Q}\}.$$

This is a field. The verification that this is a subring of $\mathbb{C}$ is the same as above. It is thus a commutative ring in which $0 \neq 1$. We need to show inverse for multiplication. If $a+b\sqrt{2}$ is not zero then either a or b are not zero. If

$$c = a^2 - 2b^2$$

is zero then either $b = 0$ (but then $a \neq 0$ and so $c \neq 0$, so this case doesn't happen), or $\sqrt{2} = a/b$ is a rational number. We shall prove in Proposition 10.2.4 that this is not the case. Thus, $c \neq 0$. Now, $\frac{a}{c} - \frac{b}{c}\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$ and it is easy to check that

$$(a+b\sqrt{2})\left(\frac{a}{c} - \frac{b}{c}\sqrt{2}\right) = 1.$$

8. Exercises

(1) Let B be a given set. What does the equality $A \cup B = B \cap C$ imply on A? on C?

(2) Calculate the following intersection and union of sets (provide short explanations, if not complete proofs).

Notation: If a, b are real numbers we use the following notation:
$[a,b] = \{x \in \mathbb{R} | a \leq x \leq b\}$.
$[a,b) = \{x \in \mathbb{R} | a \leq x < b\}$.
$(a,b] = \{x \in \mathbb{R} | a < x \leq b\}$.
$(a,b) = \{x \in \mathbb{R} | a < x < b\}$.
We also use
$[a,\infty) = \{x \in \mathbb{R} | a \leq x\}$.
$(-\infty,b] = \{x \in \mathbb{R} | x \leq b\}$.
$(-\infty,\infty) = \mathbb{R}$.
If $A_1, A_2, A_3, \ldots$ are sets, we may write $\cup_{i=1}^N A_i$ for $A_1 \cup A_2 \cup \cdots \cup A_N$ and $\cup_{i=1}^\infty A_i$ for $\cup_{i \in \{1,2,3,\ldots\}} A_i$.

(a) Let $N \geq 1$ be a natural number. What is $\cup_{n=1}^N [-n,n]$? What is $\cap_{n=1}^N [-n,n]$?

(b) What is $\cup_{n=1}^\infty [n,n+1]$? What is $\cup_{n=1}^\infty (n,n+2)$?

(c) What is $\cup_{n=1}^\infty (n,n+1)$? What is $\cup_{n=1}^\infty (1/n,1]$?

(d) Let $A_n = \{x^n : x \in \mathbb{N}\}$. What is $\cap_{n=1}^\infty A_n$?

(e) Let $B = (-1,1) \times (-1,1)$, the open square in the plane. Write B as an infinite union of closed discs of positive radius (where a closed disc with center (v_0, v_1) and positive radius is a set of the form $\{(v_0,v_1) + (x,y) : x^2 + y^2 \leq r\}$ for some fixed positive real number r).

(f) Prove that one cannot write this B as a finite union of discs.

(3) Using the intervals $[0,2]$ and $(1,3]$ and the operations of union, intersection and difference, create 8 different sets. For example, the set $[0,1] \cup (2,3]$ is $([0,2] \setminus (1,3]) \cup ((1,3] \setminus [0,2])$.

(4) Let A, B and C be sets. Prove or disprove:

(a) $(A \setminus B) \setminus C = A \setminus (B \setminus C)$;

(b) $(B \setminus A) \cup (A \setminus B) = (A \cup B) \setminus (A \cap B)$.

(5) Prove that
$$A \setminus (\cap_{i \in I} B_i) = \cup_{i \in I} (A \setminus B_i).$$

(6) Prove that the Principle of Induction (Theorem 2.3.1) implies the statement: "every non-empty subset of $\mathbb{N}$ has a minimal element".

(7) Prove by induction that for $n \geq 1$,
$$1^3 + \cdots + n^3 = (1 + \cdots + n)^2.$$
You may use the formula for the right hand side previously given.

(8) Prove by induction that $2^n > n^2$ for $n \geq 5$. It is false if we take $n \geq 0$ (because it fails for $n = 2$), but it happens to hold for $n = 0$. Where would your proof break down if you try to argue by induction starting at $n = 0$?

(9) Let C_n denote the number of ways to cover the squares of a $2 \times n$ checkers board using plain dominos. Therefore, $C_1 = 1, C_2 = 2, C_3 = 3$. Compute C_4 and C_5 and find a recursive formula for C_n. Prove by induction that $C_n = \frac{1}{\sqrt{5}} \cdot \left((\frac{1+\sqrt{5}}{2})^{n+1} - (\frac{1-\sqrt{5}}{2})^{n+1} \right)$.

(10) Prove by induction that for $n \geq 1$
$$1 + 3 + 5 + \cdots + (2n-1) = n^2.$$

(11) Let $A = \{1,2,3,4\}$ and $B = \{a,b,c\}$.

(a) Write 4 different surjective functions from A to B.

(b) Write 4 different injective functions from B to A.

(c) How many functions are there from A to B?

(d) How many surjective functions are there from A to B?

(e) How many injective functions are there from A to B?

(f) How many functions are there from B to A?

(g) How many surjective functions are there from B to A?

(h) How many injective functions are there from B to A?

(12) Let A be a set with a elements, B a set with b elements, where a, b are finite. Show that the number of functions $f\colon B \to A$ is a^b. One sometimes write the set of functions from B to A using the notation A^B. In this notation, prove that $|A^B| = |A|^{|B|}$.

In particular, using this interpretation what should 0^b be equal to if $b > 0$? what should a^0 be equal to if $a > 0$? what should 0^0 be equal to?

Note that this allows us to define the power of any cardinality to any cardinality. For example, if $\aleph_0 = |\mathbb{N}|$ then $\aleph_0^{\aleph_0} := |\mathbb{N}^{\mathbb{N}}| = |\{f\colon \mathbb{N} \to \mathbb{N}\}|$ and $2^{\aleph_0} = |\{f\colon \mathbb{N} \to \{0,1\}\}|$. One can show that $2^{\aleph_0} = |\mathbb{R}|$ by first showing $|\mathbb{R}| = |(0,1)|$ by providing explicitly a bijection and then using binary expansion of real number to show $2^{\aleph_0} = |(0,1)|$ (which is a bit tedious as the binary expansion is not quite unique: $1/2$ is both $0.1000\ldots$ and $0.011111\ldots$). In fact, also $\aleph_0^{\aleph_0} = 2^{\aleph_0}$. See also Exercise 19.

(13) Let $f : A \to B$ be a function.

a) Prove that f is bijective if and only if there exists a function $g : B \to A$ such that $f \circ g = 1_B$ and $g \circ f = 1_A$.

b) Prove or disprove: if there exists a function $g : B \to A$ such that $f \circ g = 1_B$ then f is bijective.

(14) Consider $\mathbb{N} \times \mathbb{N}$ as a rectangular array:

$$\begin{array}{cccccc} (0,0) & (0,1) & (0,2) & (0,3) & \ldots \\ (1,0) & (1,1) & (1,2) & (1,3) & \ldots \\ (2,0) & (2,1) & (2,2) & (2,3) & \ldots \\ (3,0) & (3,1) & (3,2) & (3,3) & \ldots \\ \vdots & & & & \end{array}$$

Count the elements of $\mathbb{N} \times \mathbb{N}$ using diagonals as follows:

$$\begin{array}{cccccc} 0 & 1 & 3 & 6 & 10 & \ldots \\ 2 & 4 & 7 & 11 & & \ldots \\ 5 & 8 & 12 & & & \ldots \\ 9 & 13 & & & & \ldots \\ 14 & & & & & \ldots \\ \vdots & & & & & \end{array}$$

This defines a function

$$f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$$

where $f(m,n)$ is the number appearing in the (m,n) place. (For example, $f(0,0) = 0, f(3,1) = 13, f(2,2) = 12$.) Provide an explicit formula for f (it is what one calls "a polynomial function in the variables m, n". It may be a good idea to first find a formula for $f(0,n)$). Note that this f is a bijection $\mathbb{N} \times \mathbb{N} \to \mathbb{N}$. And there is nothing wild about it. It is a rather simple polynomial.

(15) Prove that if $|A_1| = |A_2|$ and $|B_1| = |B_2|$ then $|A_1 \times B_1| = |A_2 \times B_2|$.

(16) Let A, B be sets and assume A is not the empty set. Prove that $|A| \leq |B|$ if and only if there is a surjective function $B \to A$.

(17) Prove that $|\mathbb{N}| = |\mathbb{Q}|$. (Hint: Show two inequalities; note that there is an easy injection $\mathbb{Q} \to \mathbb{Z} \times \mathbb{Z}$).

(18) Prove or disprove: if $|A_1| = |A_2|$, $A_1 \supseteq B_1, A_2 \supseteq B_2$, and $|B_1| = |B_2|$ then $|A_1 \setminus B_1| = |A_2 \setminus B_2|$.

(19) Let A be a set. Then $|A| < |2^A|$, where 2^A is the set of all subsets of A. (Another common notation for the set of subsets of A is $\mathscr{P}(A)$.) Prove this as follows: First show $|A| \leq |2^A|$ by constructing an injection $A \to 2^A$. Suppose now that there is a bijection

$$A \to 2^A, \qquad a \mapsto U_a.$$

Define a subset U of A by

$$U = \{a : a \notin U_a\}.$$

Show that if $U = U_b$ we get a contradiction. (This is some sort of "diagonal argument"). Put all this together to conclude $|A| < |2^A|$.

(20) Prove that a number is rational if and only if from some point on its decimal expansion becomes periodic.

(21) A relation can be either reflexive or not, symmetric of not, transitive or not. This gives a priori 8 possibilities (e.g., reflexive, non-symmetric, transitive). For each possibility either give an example of such a relation, or indicate why this possibility doesn't occur. Whenever possible, give "natural examples".

(22) Let A be a set.
   (a) Let $\Gamma = A \times A$. What relation does Γ define on A? Is it an equivalence relation? For every $a \in A$ write the set of elements $b \in A$ such that $a \sim b$.
   (b) Let $\Gamma = \{(a, a) : a \in A\}$. What relation does Γ define on A? Is it an equivalence relation? For every $a \in A$ write the set of elements $b \in A$ such that $a \sim b$.
   (c) Define a relation on non-zero real numbers by saying that $a \sim b$ if a/b is a rational number. Show that this is an equivalence relation.
   (d) Define a relation on complex numbers by saying that $z_1 \sim z_2$ if $|z_1| = |z_2|$. Show that this is an equivalence relation. How do the equivalence classes look like graphically?
   (e) Let $f(x)$ be a polynomial of the form $x^n + a_{n-1}x^{n-1} + \cdots + a_0$, where $n \geq 1$ and $a_i \in \mathbb{R}$. Describe a relation on $\mathbb{R}$ by saying that $a \sim b$ if $f(a) = f(b)$. Prove that this is an equivalence relation. Prove that if n is odd, there is an element $a \in \mathbb{R}$ whose equivalence class consists only of itself. (This question requires some use of calculus.)

(23) Prove that if a product $z_1 \cdot z_2$ of complex numbers is equal to zero then at least one of z_1, z_2 is zero.

(24) Let f be the complex polynomial $f(x) = (3+i)x^2 + (-2-6i)x + 12$. Find a complex number z such that the equation $f(x) = z$ has a unique solution (use the formula for solving a quadratic equation).

(25) Find the general form of a complex number z such that: (i) z^2 is a real number (i.e., $\mathrm{Im}(z^2) = 0$). (ii) z^2 is a purely imaginary number (i.e., $\mathrm{Re}(z^2) = 0$). (iii) $z^2 = \bar{z}$. (iv) $\mathrm{Im}(z^2 + \bar{z}) = 0$. (v) $\mathrm{Re}(z^2 + \bar{z}) = 0$. Also, in each of these cases, plot the answer on the complex plane.

(26) **The ring of 2×2 matrices over a field $\mathbb{F}$.**
Let $\mathbb{F}$ be a field. We consider the set

$$M_2(\mathbb{F}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{F} \right\}.$$

It is called the two-by-two matrices over $\mathbb{F}$. We define the addition of two matrices as

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} + \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{pmatrix}.$$

We define multiplication by

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{pmatrix}.$$

Prove that this is a ring. For each of the following subsets of $M_2(\mathbb{F})$ determine if they are subrings or not.

(a) The set $\left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in M_2(\mathbb{F}) \right\}$.

(b) The set $\left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \in M_2(\mathbb{F}) \right\}$.

(c) The set $\left\{ \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} \in M_2(\mathbb{F}) \right\}$.

(d) The set $\left\{ \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} \in M_2(\mathbb{F}) \right\}$.

(e) The set $\left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \in M_2(\mathbb{F}) \right\}$.

Remark: One defines in a very similar way the ring of $n \times n$ matrices with entries in a field $\mathbb{F}$.

(27) The zero ring. Let e be a formal symbol and let $R = \{e\}$. Define $e + e = e, e \cdot e = e$. Verify that R is a ring in which $1 = 0$. Prove that if S is any ring in which $1 = 0$ then S has a unique element and if we choose to denote by e then $e + e = e, e \cdot e = e$.

(28) Prove that a commutative ring with 2 elements is a field.

(29) Prove that a commutative ring with 3 elements is a field.

Part 2. Arithmetic in $\mathbb{Z}$

In this part of the course we are going to study arithmetic in the ring of integers $\mathbb{Z}$. We are going to focus on particular properties of this ring. Our choice of properties is motivated by an analogy to be drawn later between integers and polynomials. In fact, there is a general class of rings to which one can extend this analogy, called Euclidean rings; they are discussed briefly in Appendix C.

We are focusing here on rather simple properties of the integers. However, the integers have been the playground of mathematicians for millennia and the fascination with their properties never vaned. Starting from the Greek mathematicians and philosophers, for whom integers were manifestation of the divine, simple problems concerning integers were explored. Remarkably, for most of those problems we still do not know the answer. For example, a positive integer n is called perfect if the sum of its divisors, excluding itself, is equal to n. Thus, 6 is a perfect number as $6 = 1 + 2 + 3$, as is 28. We do not know if odd perfect numbers exist and we do not know if there are infinitely many even perfect numbers (although I would suspect that the answer to the first is no, and for the second is yes). Because of the special role prime numbers play in arithmetic, perhaps the most substantial open problems are problems concerning prime numbers, and some of these are mentioned below.

9. Division, GCD and the Euclidean Algorithm

9.1. Division with residue.

Theorem 9.1.1. *(Division with residue) Let a, b be integers with $b \neq 0$. There exist integers q, r such that*

$$a = qb + r, \quad 0 \leq r < |b|.$$

Moreover, q and r are uniquely determined. r is called the **residue** *and q the* **quotient**.

Proof. For simplicity, assume $b > 0$. Very similar arguments prove the case $b < 0$.

Consider the set

$$S = \{a - bx : x \in \mathbb{Z}, a - bx \geq 0\}.$$

It is the set of all non-negative "residues". We claim that S is a non-empty set. Indeed, if $a \geq 0$ take $x = 0$ and it follows that $a \in S$. If $a < 0$ take $x = a$ and $a - bx = a(1 - b) \geq 0$ (because $b > 0$ and so $b \geq 1$). That is, $a(1 - b) \in S$. Since S is a non-empty subset of $\mathbb{N}$, it follows that S has a minimal element $r = a - bq$ for some q. Then $r < b$; otherwise, $0 \leq r - b = a - b(q + 1)$ is an element of S as well and smaller that r, which is a contradiction. It follows that

$$a = bq + r, \quad 0 \leq r < b.$$

We now show that q and r are unique. Suppose

$$a = bq' + r', \quad 0 \leq r' < b.$$

If $q = q'$ then also $r = a - bq = a - bq' = r'$. Else, either $q > q'$ or $q' > q$. Note that

$$0 = bq + r - (bq' + r') = b(q - q') + (r - r').$$

If $q > q'$ then $r' = r + b(q - q') \geq r + b \geq b$. Contradiction. If $q < q'$ we get $r = r' + b(q' - q) \geq b$ and again a contradiction. □

Example 9.1.2. Let $a = 40$ and $b = 13$. We then have

$$40 = 3 \cdot 13 + 1,$$

and as $0 \leq 1 < 13$ we have divided 40 by 13 with residue 1. If we take $a = 40$ and $b = -13$ then

$$40 = -3 \times -13 + 1$$

and since $0 \leq 1 < |-13|$ we have indeed divided 40 by -13 with residue 1. On the other hand, take $a = -40$ and $b = 13$. We have $-40 = -3 \cdot 13 - 1$, but this is not proper division with residue as the residue, namely -1, is negative. The proper way to do that is

$$-40 = -4 \times 13 + 12.$$

Indeed $0 \leq 12 < 13$.

9.2. **Division.**

Definition 9.2.1. Let a, b be integers. We say that $a|b$ (read, a **divides** b) if there is an element $c \in \mathbb{Z}$ such that $b = ac$.

Here are some properties:

(1) $a|b \Rightarrow a|-b$.
(2) $a|b \Rightarrow a|bd$ for any $d \in \mathbb{Z}$.
(3) $a|b, a|d \Rightarrow a|(b \pm d)$.
(4) if $a|b$ and $b|a$ then $a = \pm b$.

Proof. Write $b = ac$. Then $-b = a \cdot (-c)$ and so $a|-b$. Also, $bd = a \cdot (cd)$ and so $a|bd$.

Write also $d = ae$. Then $b \pm d = a \cdot (c \pm e)$ and so $a|(b \pm d)$.

Now, for the last property. For some integers c, d we have $b = ac, a = bd$ and so we find that $a = bd = acd$ and so $a(1 - cd) = 0$. If $a = 0$ then also $b = 0$ and the conclusion $a = \pm b$ is definitely true. Otherwise, we must have $cd = 1$. But, as c and d are integers, we must have $c = d = \pm 1$ and the conclusion follows. □

Remark 9.2.2. Extreme cases are sometimes confusing. It follows from the definition that 0 divides only 0, that every number divides 0 and that ± 1 (and only them) divide any number.

Example 9.2.3. Which are the integers that can divide both n and $2n + 5$? If a divides n, then a divides $2n$ and so, if a divides $2n + 5$ then a divides 5, because $5 = (2n + 5) - 2n$. Therefore $a = \pm 1, \pm 5$ are the only possibilities. However, note that this does not imply that they actually occur in every case. If $n = 0$, they all do. If $n = 1$ only ± 1 are common divisors.

Corollary 9.2.4. *Let $a \neq 0$. $a|b$ if and only if in dividing b in a with residue, $b = aq + r$, the residue r is zero.*

Proof. If the residue $r = 0$ then $b = aq$ and so $a|b$. If $a|b$ and $b = aq + r$ then $a|(b - aq)$, i.e., $a|r$. But $r < |a|$ and so that's possible only if $r = 0$. □

9.3. **GCD.**

Definition 9.3.1. Let $a, b,$ be integers, not both zero. The **greatest common divisor** (gcd) of a and b, denoted $\gcd(a, b)$ or just (a, b) if the context is clear, is the largest (positive) integer dividing both a and b.

Theorem 9.3.2. *Let $a, b,$ be integers, not both zero, and $d = (a, b)$ their gcd. Then every common divisor of a and b divides d. There are integers u, v such that*

$$d = ua + vb.$$

Moreover, d is the minimal positive number that has the form $ua + vb$.

Proof. Let

$$S = \{ma + nb : m, n \in \mathbb{Z}, ma + nb > 0\}.$$

First note that $S \neq \emptyset$. Indeed, $aa + bb \in S$. Let D be the minimal element of S. Then, for some $u, v \in \mathbb{Z}$ we have $D = ua + vb$.

We claim that $D = d$. To show $D|a$, write $a = qD + r, 0 \leq r < D$. Then, $D > r = a - qD = a - q(ua + vb) = (1 - qu)a - qvb$. If $r \neq 0$ then $r = (1 - qu)a - qvb$ is an element of S smaller than D and that's a contradiction. It follows that $r = 0$, that is $D|a$. In the same way, $D|b$.

On the other hand, let e be any common divisor of a and b. Then e also divides $ua + vb = D$. It follows that D is the largest common divisor of a, b, so $D = d$, and also that any other common divisor of a and b divides it. □

Corollary 9.3.3. *If $a|bc$ and $gcd(a, b) = 1$ then $a|c$.*

Proof. We have $1 = ua + vb$ for some integers u, v. Since $a|uac$ and $a|vbc$ we have $a|uac + vbc = c$. □

9.4. **The Euclidean algorithm.** [14] The question arises: how do we compute in practice the gcd of two integers? This is a very practical issue, even in the simple task of simplifying fractions! As we shall see, there are two methods. One method uses the prime factorization of the two numbers; we shall discuss that later. The other method, **which is much more efficient**, is the Euclidean algorithm.

Theorem 9.4.1. *(The Euclidean Algorithm) Let a, b be positive integers with $a \geq b$. If $b|a$ then $\gcd(a,b) = b$. Else perform the following recursive division with residue:*

$$\begin{aligned} a &= bq_0 + r_0, && 0 < r_0 < b, \\ b &= r_0q_1 + r_1, && 0 \leq r_1 < r_0, \\ r_0 &= r_1q_2 + r_2, && 0 \leq r_2 < r_1, \\ &\vdots \end{aligned}$$

For some t we must first get that $r_{t+1} = 0$. That is,

$$\begin{aligned} r_{t-2} &= r_{t-1}q_t + r_t, && 0 < r_t < r_{t-1} \\ r_{t-1} &= r_tq_{t+1}. \end{aligned}$$

Then r_t is the gcd of a and b.

Before giving the proof we provide two examples.
1). Take $a = 113, b = 54$. Then

$$\begin{aligned} 113 &= \underline{54} \cdot 2 + \underline{5} \\ \underline{54} &= \underline{5} \cdot 10 + 4 \\ \underline{5} &= \underline{4} \cdot 1 + 1 \\ \underline{4} &= \underline{1} \cdot 4. \end{aligned}$$

Thus $\gcd(113, 54) = 1$.

2). Now take $a = 442, b = 182$. Then

$$\begin{aligned} 442 &= \underline{182} \cdot 2 + \underline{78} \\ \underline{182} &= \underline{78} \cdot 2 + 26 \\ \underline{78} &= \underline{26} \cdot 3 \end{aligned}$$

and so $\gcd(442, 182) = 26$.

Proof. Let $d = \gcd(a,b)$. We claim that $d|r_n$ for every n. We prove that by induction: First, $d|a, d|b$ then $d|(a - bq_0) = r_0$. Suppose that $d|r_i, i = 0, 1, 2, \ldots, n$. Since $r_{n+1} = r_{n-1} - r_nq_{n+1}$ we get that $d|r_{n+1}$ as well. In particular, $d|r_t$.

We now show that $r_t|a, r_t|b$. It then follows that $r_t|d$ and therefore $r_t = d$. We again prove that by induction. We have $r_t|r_t$ and $r_t|r_tq_{t+1} = r_{t-1}$. Suppose we have already shown that r_t divides $r_t, r_{t-1}, \ldots, r_n$. Then, since $r_{n-1} = r_nq_{n+1} + r_{n+1}$ we also get $r_t|r_{n-1}$. Therefore, r_t divides $r_0, r_1, \ldots, r_t$. Again, $b = r_0q_1 + r_1$ and so $r_t|b$ and then $a = bq_0 + r_0$ and so $r_t|a$. □

[14] Euclid (about 325 - 265 BC), after whom Euclidean geometry and the Euclidean algorithm are named was a Greek mathematician, most famous for his magnum opus "Elements" consisting of 13 books devoted to plane and spacial geometry, number theory and algebra. Besides landmark results, such as the infinitude of primes and the Fundamental Theory of Arithmetic, the Elements are important for having set mathematics on the right course. Euclid's postulates laying the logical foundation to plane geometry, served as model for the development of mathematics at large. The Elements is the most successful textbook ever written and have served since its inception until the 19th and even early 20th century as a textbook in high school and university. Abraham Lincoln kept a copy of Euclid in his saddlebag, and studied it late at night by lamplight; he related that he said to himself, "You never can make a lawyer if you do not understand what demonstrate means; and I left my situation in Springfield, went home to my father's house, and stayed there till I could give any proposition in the six books of Euclid at sight".

A further bonus supplied by the Euclidean algorithm is that it allows us to find u, v such that $\gcd(a,b) = ua + vb$. We just illustrate it in two examples:
1). Take $a = 113, b = 54$. Then, as we saw,

$$\begin{aligned} 113 &= \underline{54} \cdot 2 + \underline{5} \\ \underline{54} &= \underline{5} \cdot 10 + 4 \\ 5 &= 4 \cdot 1 + 1 \\ 4 &= 1 \cdot 4. \end{aligned}$$

and so $\gcd(113, 54) = 1$. We have $1 = 5 - 4 \cdot 1$, and we keep substituting for the residues we now have expressions involving previous residues (the important numbers to modify are the **residues** not the quotients q_i). $4 = 54 - 5 \cdot 10$ and we get $1 = 5 - (54 - 5 \cdot 10) = -54 + 5 \cdot 11$. Next, $5 = 113 - 54 \cdot 2$ and we get $1 = -54 + 5 \cdot 11 = -54 + (113 - 54 \cdot 2) \cdot 11 = 54 \cdot (-23) + 113 \cdot 11$. Thus,

$$1 = \gcd(54, 113) = -23 \cdot 54 + 11 \cdot 113.$$

2). Now take $a = 442, b = 182$. Then

$$\begin{aligned} 442 &= \underline{182} \cdot 2 + \underline{78} \\ \underline{182} &= \underline{78} \cdot 2 + 26 \\ 78 &= 26 \cdot 3 \end{aligned}$$

and so $\gcd(442, 182) = 26$. Here the process is easier: $26 = 182 - 78 \cdot 2 = 182 - (442 - 182 \cdot 2) \cdot 2 = 5 \cdot 182 - 2 \cdot 442$.

$$26 = \gcd(182, 442) = 5 \cdot 182 - 2 \cdot 442.$$

10. Primes and unique factorization

10.1. **Primes.**

Definition 10.1.1. An integer $p \neq 0, \pm 1$ is called **prime** if its only divisors are $\pm 1, \pm p$.

The phrase "prime number" is usually used to denote a prime positive integer. An integer $p > 1$ is prime iff its only positive divisors are 1 and p.

The sieve of Eratosthenes: [15] This is a method that allows one to construct rapidly a list of all primes less than a given number N. We illustrate that with $N = 50$. One writes all the numbers from 2 to 50:

2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40,41, 42, 43, 44, 45,
46, 47, 48, 49, 50

The first number on the list is prime. This is 2. We write it in bold-face and cross all its multiples (we denote crossing out by an underline):

2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44,
45, 46, 47, 48, 49, 50

The first number on the list not in bold-face and not crossed out is prime. This is 3. We write it in bold-face and cross all its multiples (we denote crossing out by another underline):

[15] Eratosthenes of Cyrene, 276BC - 194BC, was a Greek mathematician and is famous for his work on prime numbers and for measuring the diameter of the earth. For more see http://www-groups.dcs.st-and.ac.uk/%7Ehistory/Biographies/Eratosthenes.html

2, **3**, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

The first number on the list not in bold-face and not crossed out is prime. This is 5. We write it in bold-face and cross all its multiples (we denote crossing out by an underline):

2, **3**, 4, **5**, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

The first number on the list not in bold-face and not crossed out is prime. This is 7. We write it in bold-face and cross all its multiples (we denote crossing out by yet another underline):

2, **3**, 4, **5**, 6, **7**, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

The next number 11 is already greater than $\sqrt{N} = \sqrt{50} \sim 7.071\ldots$. So we stop, because any number is a product of prime numbers (see Lemma 10.1.3 below) and so any number less or equal to N, which is not prime, has at least one prime divisor smaller or equal to $\sqrt{N}$. Thus, any number left on our list is prime.

2, **3**, 4, **5**, 6, **7**, 8, 9, 10, **11**, 12, **13**, 14, 15, 16, **17**, 18, **19**, 20, 21, 22, **23**, 24, 25, 26, 27, 28, **29**, 30, **31**, 32, 33, 34, 35, 36, **37**, 38, 39, 40, **41**, 42, **43**, 44, 45, 46, **47**, 48, 49, 50

Theorem 10.1.2 (The Fundamental Theorem of Arithmetic)**.** *Every non-zero integer n is a product of primes. (We allow the empty product, equal by definition to 1). That is, one can write every non-zero integer n as*

$$n = \epsilon p_1 p_2 \cdots p_m,$$

where $\epsilon = \pm 1$ and $0 \leq p_1 \leq p_2 \leq \cdots \leq p_m$ are primes ($m \geq 0$). Moreover, this way of writing n is unique.

Proof. We first show n can be written this way. We may assume n is positive (if n is negative, apply the statement to $-n$, $-n = p_1 p_2 \cdots p_m$ and thus $n = -1 \cdot p_1 p_2 \cdots p_m$).

Lemma 10.1.3. *Every positive integer is a product of primes numbers. (We allow the empty product, equal by definition to 1).*

Proof. Suppose not. Then the set of integers S that are not a product of prime numbers has a minimal element, say n_0. n_0 is not one, or a prime, because in those cases it is a product of primes (1, as said, is the empty product). Thus, there are integers $1 < s < n_0, 1 < t < n_0$ such that $n_0 = st$. Note that s and t are not in S because they are smaller than n_0. Thus, $s = q_1 q_2 \cdots q_a$ is a product of primes, $t = r_1 r_2 \cdots r_b$ is a product of primes and therefore $n = q_1 q_2 \cdots q_a r_1 r_2 \cdots r_b$ is also a product of primes. This is a contradiction - a contradiction to our initial assumption that there are positive integers that are not a product of prime numbers. Thus, every positive integer is a product of prime numbers. □

Choosing the sign ϵ appropriately and ordering the primes in increasing order we conclude that any non-zero integer $n = \epsilon p_1 p_2 \cdots p_m$, where $\epsilon = \pm 1$ and $0 \leq p_1 \leq p_2 \leq \cdots \leq p_m$ are primes. We now show uniqueness. For this we need the following important fact.

Proposition 10.1.4. *Let $p > 1$ be an integer. The following are equivalent:*

(1) p is a prime number;
(2) if $p|ab$ then $p|a$ or $p|b$.

Proof. Suppose p is prime and $p|ab$. If $p \not| a$ then $\gcd(p, a) = 1$ and so, as we have already seen (Corollary 9.3.3), $p|b$.

Now suppose that p satisfies (ii). Let $p = st$. By replacing s by $-s$ and t by $-t$, we may assume that s, t are positive. As $p = st$, $p|st$ and so $p|s$, say. So $s = ps'$ and $p = ps't$. But we must have then

that $s' = t = 1$, because s', t are positive integers. Therefore $p = s$. So p has no proper divisors and hence is prime. □

Remark 10.1.5. Suppose that p is a prime dividing a product of integers $q_1 q_2 \cdots q_t$. Arguing by induction on t, one concludes that p divides some q_i.

We now finish the proof of the theorem. Suppose that

$$n = \epsilon p_1 p_2 \cdots p_m,$$

and also

$$n = \mu q_1 q_2 \cdots q_t,$$

are two expressions of n as in the statement of the theorem. First, ϵ is negative if and only if n is, and the same holds for μ. So $\epsilon = \mu$. We may then assume n is positive and $\epsilon = \mu = 1$ and we argue by induction on n. The case $n = 1$ is clear: a product of one or more primes will be greater than 1 so the only way to express n is as the empty product. Assume the statement holds for $1, 2, \ldots, n-1$ and consider two factorizations of n:

$$n = p_1 p_2 \cdots p_m,$$

and

$$n = q_1 q_2 \cdots q_t.$$

First, note that $m \geq 1$ and $t \geq 1$ because $n > 1$. Assume that $p_1 \leq q_1$ (the argument in the other case goes the same way). We have $p_1 | n$ and so $p_1 | q_1 q_2 \cdots q_t$. It follows that p_1 divides some q_i but then, q_i being prime, $p_1 = q_i$. Furthermore, $p_1 \leq q_1 \leq q_i = p_1$, so $p_1 = q_1$. We then have the factorizations

$$\frac{n}{p_1} = p_2 \cdots p_m = q_2 \cdots q_t.$$

Since $n/p_1 < n$ we may apply the induction hypothesis and conclude that $m = t$ and $p_i = q_i$ for all i. □

The Fundamental Theorem exhibits the prime numbers as the building blocks of the integers. In itself, it doesn't tell us if there are finitely many or infinitely many of them.

Theorem 10.1.6 (Euclid)**.** *There are infinitely many prime numbers.*

Proof. Let $p_1, p_2, \ldots, p_n$ be distinct prime numbers. We show then that there is a prime not in this list. It follows that there couldn't be only finitely many prime numbers.

Consider the integer $n = p_1 p_2 \cdots p_n + 1$ and its prime factorization. Let q be a prime dividing n. If $q \in \{p_1, p_2, \ldots, p_n\}$ then $q | p_1 p_2 \cdots p_n$ and so $q | (n - p_1 p_2 \cdots p_n)$, that is $q | 1$, which is a contradiction. Thus, q is a prime not in the list $\{p_1, p_2, \ldots, p_n\}$. □

So! We know that every integer is a product of prime numbers, we know there are infinitely many prime numbers. That teaches us about the integers, and invites some more questions:

– **How frequent are prime numbers?** The Prime Number Theorem asserts that the number of primes in the interval $[1, n]$ is roughly $n/\log n$, in the sense that the ratio between the true number and the estimate $n/\log n$ approaches 1 as n goes to infinity. The result was conjectured by Gauss[16] at the age of 15 or 16 and proven by J. Hadamard and Ch. de la Vallée Poussin, independently, in 1896. The famous Riemann hypothesis, currently (Fall 2023) one of the Clay Institute million dollar millennium prize problems, is initially a conjecture about the zeros of some complex valued function – the so-called Riemann zeta function. An equivalent formulation is the following: let $\pi(x)$ be the number of prime numbers in the interval $[1, x]$, for $x > 1$, then for some constant C, we have

$$|\pi(x) - \frac{x}{\log x}| < C\sqrt{x} \cdot \log x.$$

[16] Johann Carl Friedrich Gauss, 1777 - 1855, worked in a wide variety of fields in both mathematics and physics including number theory, analysis, differential geometry, geodesy, magnetism, astronomy and optics. His work has had an immense influence in many areas of Science. He is mentioned in the same breath with Euclid, Archimedes and Newton as one of the giants of mathematics.

– **How small can the gaps between consecutive primes be?** For example, we have $(3,5)$, $(5,7)$, $(11,13)$, $(17,19)$, ... But, are there infinitely many such pairs?? The answer is believed to be yes but no one has proven it yet (Fall 2023). This is called the **Twin Primes Conjecture.** Incidentally, it is not hard to prove that the gap between primes can be arbitrarily large. Given an integer N, consider the numbers

$$N!+2, N!+3, N!+4, \ldots, N!+N.$$

This is a set of $N-1$ consecutive integers none of which is prime. In recent breakthroughs, originating with Yitang Zhang in 2013, it was established that there is a computable constant C such that there are infinitely many pairs of primes at most C apart. Zhang provided $C = 70,000,000$ and J. Maynard improved that not much after to $C = 600$; "Polymath project 8" - a large group of mathematicians working on the problem using an online platform - improved the bound further to $C = 256$. It is believed that their methods can be improved to yield $C = 6$, but not (oh, painful irony!) to yield $C = 2$, leaving the twin primes conjecture just beyond reach.

– **How far does one need to go until the next prime shows up?** For example, it is known that there is always a prime between n and $2n$. It follows rather easily from the prime number theorem for $n \gg 0$, but it in fact holds for every n. There is a rather elementary, yet ingenious, proof of this fact, the technical background for which is not much more than Stirling's formula. We sometimes give it in the Number Theory course.

– **What about adding primes? Goldbach's conjecture** asserts that every even integer greater than 2 is the sum of two prime numbers. For example, 4 = 2 + 2, 6 = 3 + 3, 8 = 3+5, 10 = 3+7, 12 = 5 + 7, 14 = 3 + 11, 16 = 5+ 11, ... It has been verified (Winter 2016) up to $n \leq 4 \times 10^{17}$, but no proof is currently known (Fall 2023).

The "odd Goldbach conjecture" states that every odd integer greater than 5 is the sum of 3 primes. Remarkably this is known. Series of works over a century proved better and better results towards the odd conjecture. In particular, it was known in 2002 to be true for every odd integer greater than $C = 2 \cdot 10^{1346}$, which is computationally out of reach. Harald Helfgott proved the full conjecture in 2013, by improving the constant C greatly so that the remaining cases could be verified by direct computation (and, in fact, were already verified prior to his work).

10.2. **Applications of the Fundamental Theorem of Arithmetic.**

Proposition 10.2.1. *Let a, b be non-zero integers. Then $a|b$ if and only if one can write $a = \epsilon p_1^{a_1} \cdots p_m^{a_m}$ and $b = \mu p_1^{a_1'} \cdots p_m^{a_m'} q_1^{b_1} \cdots q_t^{b_t}$ (products of distinct primes) with $a_i' \geq a_i > 0$ for all $i = 1, \ldots, m$, $b_i > 0$ or all $i = 1, \ldots, t$.*

Proof. Clearly for such factorizations it follows that $a|b$, in fact

$$b/a = (\mu/\epsilon) p_1^{a_1'-a_1} \cdots p_m^{a_m'-a_m} q_1^{b_1} \cdots q_t^{b_t}.$$

Conversely, if $a|b$, write $a = \epsilon p_1^{a_1} \cdots p_m^{a_m}$ and $b/a = \nu p_1^{a_1''} \cdots p_m^{a_m''} q_1^{b_1} \cdots q_t^{b_t}$, with $\nu = \pm 1$ and $a_i'' \geq 0$. Then $b = (\nu\epsilon) p_1^{a_1+a_1''} \cdots p_m^{a_m+a_m''} q_1^{b_1} \cdots q_t^{b_t}$ and let $\mu = \nu\epsilon, a_i' = a_i + a_i''$. □

Corollary 10.2.2. *Let $a = p_1^{a_1} \cdots p_m^{a_m}, b = p_1^{b_1} \cdots p_m^{b_m}$ with p_i distinct prime numbers and a_i, b_i non-negative integers. (Any two positive integers can be written this way). Then*

$$\gcd(a,b) = p_1^{\min(a_1,b_1)} \cdots p_m^{\min(a_m,b_m)}.$$

The next proposition establishes the existence of real numbers that are not rational. It can be generalized considerably. In fact, it is known that in randomly choosing a number in the interval $[0,1]$ the probability of picking a rational number is zero. So, although there are infinitely many rational numbers, even in the interval $[0,1]$, they are still a rather meagre set inside the real numbers. In fact, recall that by Cantor's diagonal argument, we know that the cardinality of $\mathbb{R}$ is greater than that of $\mathbb{Q}$, so there must be some real number

that is not rational. However, the advantage of Proposition 10.2.4 is that it shows that a **specific** number is irrational. We will need the following result.

Proposition 10.2.3. *Any non-zero rational number q can be written as*

$$q = \epsilon p_1^{a_1} \dots p_m^{a_m},$$

where $\epsilon = \pm 1$, the p_i are distinct prime numbers and $a_1, \dots, a_m$ are non-zero integers (possibly negative). Moreover, this expression is unique (up to reordering the primes).

Proof. By definition, for some integers a, b we have $q = a/b$. Let us write:

$$a = \epsilon_a r_1^{s_1} \cdots r_n^{s_n}, \quad b = \epsilon_b r_1^{t_1} \cdots r_n^{t_n},$$

where $\epsilon_a, \epsilon_b \in \{\pm 1\}$, the r_i are distinct primes and s_i, t_i are non-negative integers, possibly zero. This is always possible to do because we allow zero exponents here. Then, clearly,

$$q = (\epsilon_a/\epsilon_b) r_1^{s_1 - t_1} \cdots r_n^{s_n - t_n}.$$

Now, omitting the primes such that $s_i - t_i = 0$ from the list, calling the remaining primes $p_1, \dots, p_m$, and letting $\epsilon = \epsilon_a/\epsilon_b$, we obtain an expression as desired.

Suppose that we have two such expressions for q. Again, by allowing zero exponents, it is enough to consider the following situation:

$$q = \epsilon p_1^{a_1} \dots p_m^{a_m} = \epsilon' p_1^{a_1'} \dots p_m^{a_m'}.$$

Since ϵ, ϵ' determine the sign of q, they must be equal and we need to show that $a_i = a_i'$ for all i. Dividing through, we get an expression of the form,

$$1 = p_1^{c_1} \dots p_m^{c_m},$$

where $c_i = a_i - a_i'$ and we need to show all the c_i are zero. By rearranging the primes, we may assume that $c_1, \dots, c_t$ are negative and $c_{t+1}, \dots, c_m$ are non-negative. We conclude that

$$p_1^{-c_1} \dots p_t^{-c_t} = p_{t+1}^{c_{t+1}} \dots p_m^{c_m}.$$

In this expression there are no negative exponents. Thus, from unique factorization for integers, since all the primes are distinct all the powers must be zero. □

Proposition 10.2.4. *$\sqrt{2}$ is not a rational number.*

Proof. Suppose it is, and write $\sqrt{2} = p_1^{a_1} \cdots p_m^{a_m}$, distinct primes with non-zero exponents (possibly negative). Then

$$2 = p_1^{2a_1} \cdots p_m^{2a_m}$$

must be the unique factorization of 2. However, 2 is prime. So there must be only one prime on the right hand side, i.e. $m = 1$. Then $2 = p_1^{2a_1}$ and we must have $p_1 = 2$ and $2a_1 = 1$. But this contradicts the fact that a_1 is an integer.

Here is another proof:
Suppose that $\sqrt{2}$ is rational and write $\sqrt{2} = m/n$, where $(m, n) = 1$. Then

$$2n^2 = m^2.$$

This implies that $2 | m^2 = m \cdot m$. As 2 is prime, it follows that $2 | m$. Thus, m is even, say $m = 2k$. It follows that $2n^2 = 4k^2$ and so $n^2 = 2k^2$. Therefore, $2 | n^2$ and by the same considerations $2 | n$. This means that 2 divides both n and m, contrary to our assumption. Thus, assuming $\sqrt{2}$ is rational leads to a contradiction and so $\sqrt{2}$ is not a rational number. □

Example 10.2.5. We can draw the conclusion that also $\alpha = \sqrt{1 + \sqrt{2}}$ is irrational. Indeed, if α was rational, say equal to m/n, then $\alpha^2 = m^2/n^2$ is also rational and so is $\alpha^2 - 1 = (m^2 - n^2)/n^2$. But $\alpha^2 - 1 = \sqrt{2}$, which we know to be irrational.

Some problems sound very hard, but turn out to have elementary solutions. For example, consider the statement: "there exists two irrational numbers a, b such that a^b is rational". This seems hard, but a shrewd observation due to Dov Jarden, gives a proof: if $\sqrt{2}^{\sqrt{2}}$ is rational then we are done. Otherwise, consider

a^b, where $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$. As $a^b = 2$, we are done. This clever argument doesn't tell us which is the correct example, $\sqrt{2}^{\sqrt{2}}$, or $(\sqrt{2}^{\sqrt{2}})^{\sqrt{2}}$. In fact, by the Gelfond-Schneider theorem, we know $\sqrt{2}^{\sqrt{2}}$ is transcendental (namely, it doesn't solve any non-zero polynomial with rational coefficients; being irrational just means it doesn't solve any linear polynomial with rational coefficients), so the example is actually $(\sqrt{2}^{\sqrt{2}})^{\sqrt{2}}$. See also Exercise 20.

Although the second proof of Proposition 10.2.4 is more elementary, the first proof is better. It is much easier to generalize and indeed in a similar way, one can show $\sqrt{3}, \sqrt[5]{18}$ etc. are irrational.

It is also known that e is irrational (see Appendix B) and π is irrational (hard). But it is still an open question (Fall 2023) whether Euler's constant

$$\gamma = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + + \frac{1}{n} - \log(n)\right) \approx 0.57721$$

is rational or not (it is believed to be irrational; if γ is rational, it was proved that its denominator has to have more than 10^{242080} digits!).

11. Exercises

(1) Find the quotient and remainder when a is divided by b:
   (a) $a = 302, b = 19$.
   (b) $a = -302, b = 19$.
   (c) $a = 0, b = 19$.
   (d) $a = 2000, b = 17$.
   (e) $a = 2001, b = 17$.
   (f) $a = 2002, b = 17$.

(2) Prove that the square of any integer a is either of the form $3k$ or of the form $3k+1$ for some integer k. (Hint: write a in the form $3q + r$, where $r = 0, 1$ or 2.)

(3) Prove of disprove: If $a|(b+c)$ then $a|b$ or $a|c$. If $a|b^c$ then $a|b$. If $a|c$ and $(a+b)|c$ then $b|c$.

(4) If $r \in \mathbb{Z}$ and r is a solution of $x^2 + ax + b$ (where $a, b \in \mathbb{Z}$) prove that $r|b$.

(5) If $n \in \mathbb{Z}$, what are the possible values of
   (a) $(n, n+2)$;
   (b) $(n, n+6)$.

(6) Find the following gcd's. In each case also express (a, b) as $ua + vb$ for suitable integers $u, v \in \mathbb{Z}$.
   (a) $(56, 72)$.
   (b) $(24, 138)$.
   (c) $(143, 227)$.
   (d) $(314, 159)$.

(7) If $a|c$ and $b|c$, must ab divide c? What if $(a, b) = 1$?

(8) The least common multiple of nonzero integers a, b is the smallest positive integer m such that $a|m$ and $b|m$. We denote it by $\mathrm{lcm}(a, b)$ or $[a, b]$. Prove that:
   (a) If $a|k$ and $b|k$ then $[a, b]|k$.
   (b) $[a, b] = \frac{ab}{(a,b)}$ if $a > 0, b > 0$.

(9) Let $a = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}$ and $b = p_1^{s_1} p_2^{s_2} \cdots p_k^{s_k}$, where $p_1, p_2, \ldots, p_k$ are distinct positive primes and each $r_i, s_i \geq 0$. Prove that
   (a) $(a, b) = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$, where $n_i = \min(r_i, s_i)$.
   (b) $[a, b] = p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k}$, where $t_i = \max(r_i, s_i)$.

(10) Prove or disprove: If n is an integer and $n > 2$, then there exists a prime p such that $n < p < n!$.

(11) Find all the primes between 1 and 150. The solution should consist of a list of all the primes + giving the last prime used to sieve + explanation why you didn't have to sieve by larger primes.

(12) Unlike twin primes, prove that there is a unique triplet of primes; namely, if $n, n+2, n+4$ are primes for some $n \geq 1$ then $n = 3$.

(13) Prove that $\sqrt{2 + \sqrt{3}}$ is irrational.

(14) Prove that if p is a prime then $\sqrt{p}$ is irrational.

(15) Prove that if p is a prime then $\sqrt[3]{p}$ is irrational.

(16) Let $k \geq 1$ be an integer. Prove that if a positive integer n is not a k-th power of another integer than $\sqrt[k]{n}$ is irrational.

(17) Prove that there are no rational numbers a, b such that $\sqrt{3} = a + b\sqrt{2}$.

(18) If the ratio of the frequencies of two musical notes is $3 : 2$, we say the notes form a fifth. For example, the notes C and G (where G is higher) form a fifth, as do the notes G and D (where D is higher). In fact, in the following sequence any two consecutive notes are supposed to form a fifth:

$$C, G, D, A, E, B, F\sharp, C\sharp, G\sharp, D\sharp, A\sharp, F, C.$$

(The last C is 7 octaves higher than the first C).

On the other hand, the ratio between two consecutive C's is an octave and is $2:1$. Explain why this leads to a contradiction; the sequence above cannot really be a sequence of fifths. This is solved, in tuning a piano, by the so-called equal temperament method; the fifths are not quite in ratio $3:2$. In fact, they are in ratio $x:1$, where x is chosen so to ensure that the octaves stay in ratio $2:1$. What is x and how close it is to $3:2$?

(19) Prove that there are infinitely many primes congruent to 3 modulo 4.

(20) Let I be the interval $((1/e)^{1/e}, \infty)$. Then every rational number in I is either of the form a^a where a is irrational, or of the form n^n where n is an integer. This gives infinitely many examples of irrational numbers a such that a^a is rational.

Here are some suggestions as to how to prove this statement. First, using analysis show that the function $f(x) = x^x$ is bijection between the interval $(1/e, \infty)$ and I. Now, let r be a rational number in I and assume that $r = a^a$, where a is rational too. Write $a = n/m$ and $r = b/c$ where b, c, m, n are positive integers and $(b, c) = (m, n) = 1$. The goal is to show $m = 1$.

By analysing $(n/m)^{n/m} = b/c$ conclude that we must have $c^m = m^n$. If $m > 1$, take a prime p dividing both c and m, say $p^i \| c$ and $p^j \| m$ (the notation $p^i \| c$ means that $p^i | c$, but $p^{i+1} \not| c$); deduce that $im = jn$ and from that $p^j | j$. Show that this is not possible.

Part 3. **Congruences and modular arithmetic**

12. Congruence modulo n

Let n be a positive integer. Define a relation on the set of integers by $x \sim y$ if $n|(x-y)$ (we shall also write that as $x \equiv y \pmod{n}$, or simply $x \equiv y$, if n is clear from the context). We say that x **is congruent to** y **modulo** n**.**

Lemma 12.0.1. *Congruence modulo n is an equivalence relation on $\mathbb{Z}$. The set $\{0,1,\dots,n-1\}$ is a complete set of representatives.*

Proof. First $n|(x-x)$ so $x \equiv x$ and the relation is thus reflexive. If $n|(x-y)$ then $n|-(x-y) = y-x$, so the relation is symmetric. Suppose $n|(x-y), n|(y-z)$ then $n|((x-y)+(y-z)) = x-z$ and so the relation is transitive too.

Let x be any integer and write $x = qn+r$ with $0 \le r < n$. Then $x-r = qn$ and so $x \equiv r$. It follows that every equivalence class is represented by some $r \in \{0,1,\dots,n-1\}$. The equivalence classes defined by elements of $\{0,1,\dots,n-1\}$ are disjoint. If not, then for some $0 \le i < j < n$ we have $i \equiv j$, that is, $n|(j-i)$. But $0 < j-i < n$ and we get a contradiction. □

Theorem 12.0.2. *Denote the equivalence classes of congruence modulo n by $\bar{0},\bar{1},\dots,\overline{n-1}$ instead of $[0],[1],\dots,[n-1]$. Denote this set by $\mathbb{Z}/n\mathbb{Z}$. The set $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring under the following operations:*

$$\bar{i}+\bar{j} = \overline{i+j}, \qquad \bar{i}\cdot\bar{j} = \overline{ij}.$$

The neutral element for addition is $\bar{0}$, for multiplication $\bar{1}$, and the inverse of $\bar{i}$ with respect to addition is $\overline{-i} = \overline{n-i}$.

Before proving the theorem we illustrate the definitions in a numerical example:

Example 12.0.3. We take $n = 13$ and calculate $\bar{5}\cdot\bar{6}-\bar{5}$. First, $\bar{5}\cdot\bar{6} = \overline{30} = \bar{4}$. Then $\bar{4}-\bar{5} = \bar{4}+\overline{-5} = \overline{4-5} = \overline{-1} = \overline{12}$. Note that we could have also calculated $\bar{4}-\bar{5} = \bar{4}+\bar{8} = \overline{12}$, or $\bar{5}\cdot\bar{6}-\bar{5} = \bar{5}(\bar{6}-\bar{1}) = \bar{5}\cdot\bar{5} = \overline{25} = \overline{12}$.

Modular arithmetic, that is calculating in the ring $\mathbb{Z}/n\mathbb{Z}$, is some times called "clock arithmetic". The reason is the following. The usual clock is really displaying hours modulo 12. When 5 hours pass from the time 10 o'clock, the clock shows 3 o'clock. Note that $3 \equiv 15 \pmod{12}$. We are used to adding hours modulo 12 (or modulo 24, for that matter), but we are not used to multiplying hours as that doesn't quite make sense. However, if we think about multiplication as repeated addition then $5\cdot 3 = 5+5+5$. So, in that sense, we are already familiar with the operations modulo 12 and the definitions above are a generalization.

Continuing with our numerical example, let us solve the equation $4x+2 = 7$ in $\mathbb{Z}/13\mathbb{Z}$. From now on we are just writing $4,2,7$ etc. for $\bar{4},\bar{2},\bar{7}$. So we need to solve $4x = 5$. Let us first look for a residue class r modulo 13 so that $4r \equiv 1 \pmod{13}$. We guess (but later we shall develop methods for that) that $r = 10$ and check: $4\cdot 10 = 40 \equiv 1 \pmod{13}$. We now multiply both sides of the equation $4x = 5$ by 10. Then, $4x = 5$ implies $10\cdot 4x \equiv x \equiv 50 \equiv 11 \pmod{13}$. Thus, the only possibility is $x = 11$. We go back to the original equation $4x = 5$ and verify that $4\cdot 11 \equiv 5 \pmod{13}$. We found the solution $x = 11$.

The idea of the calculation was to use that for a given $a \neq 0$ we can find r such that $ra = 1$. This is used to solve the equation $ax = b$ by reducing to $x = rax = rb$, and we had done that for the particular case $a = 4$ and $b = 5$. We remark that in general such an r need not exists if the modulos n is not a prime. These issues will be discussed later. In particular, as we shall see, there are efficient ways to find the solution to $ax \equiv 1 \pmod{n}$, but the situation for equations of the form $ax \equiv b \pmod{n}$ is more involved.

Example 12.0.4. As another example, we give the addition and multiplication table of the ring $\mathbb{Z}/5\mathbb{Z}$.

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Proof. (Of Theorem 12.0.2) We first prove that the operations do not depend on the representatives for the equivalence classes that we have chosen.

Suppose $\bar{i} = \bar{i'}$, $\bar{j} = \bar{j'}$, where i, i', j, j' need not be in the set $\{0, 1, 2, \ldots, n-1\}$. We have defined $\bar{i} + \bar{j} = \overline{i+j}$ and so we need to check that this is the same as $\overline{i'+j'}$. Since $\bar{i} = \bar{i'}, n|(i-i')$ and similarly $n|(j-j')$. Therefore, $n|((i+j)-(i'+j'))$; that is, $\overline{i+j} = \overline{i'+j'}$.

We also need to show that $\overline{ij} = \overline{i'j'}$. But, $ij - i'j' = ij - ij' + ij' - i'j' = i(j-j') + j'(i-i')$ and so $n|(ij - i'j')$.

The verification of the axioms is now easy if we make use of the fact that $\mathbb{Z}$ is a commutative ring:

(1) $\bar{i} + \bar{j} = \overline{i+j} = \overline{j+i} = \bar{j} + \bar{i}$. (The first and third equalities are by definition and the second equality follows from $\mathbb{Z}$ being a commutative ring.)

(2) $(\bar{i} + \bar{j}) + \bar{k} = \overline{i+j} + \bar{k} = \overline{(i+j)+k}$. Note that at this point we have used the simplification that we can use *any* representatives of the equivalence classes to carry out the operations. Had we insisted on always using representatives in the set $\{0, 1, 2, \ldots, n-1\}$ we would usually have needed to replace $i+j$ by its representative in that set and things would be turning messy. Now, $\overline{(i+j)+k} = \overline{i+(j+k)} = \bar{i} + \overline{j+k} = \bar{i} + (\bar{j} + \bar{k})$.

(3) $\bar{0} + \bar{i} = \overline{0+i} = \bar{i}$.

(4) $\bar{i} + \overline{-i} = \overline{i+(-i)} = \bar{0}$. Note that $\overline{-i} = \overline{n-i}$.

(5) $(\bar{i} \cdot \bar{j})\bar{k} = \overline{ij} \cdot \bar{k} = \overline{(ij)k} = \overline{i(jk)} = \bar{i} \cdot \overline{jk} = \bar{i}(\bar{j} \cdot \bar{k})$. (A dot is added occasionally merely to make it easier to read.)

(6) $\bar{1} \cdot \bar{i} = \overline{1 \cdot i} = \bar{i}$ and $\bar{i} \cdot \bar{1} = \overline{i \cdot 1} = \bar{i}$.

(7) $\bar{i}(\bar{j} + \bar{k}) = \bar{i} \cdot \overline{j+k} = \overline{i(j+k)} = \overline{ij + ik} = \overline{ij} + \overline{ik} = \bar{i} \cdot \bar{j} + \bar{i} \cdot \bar{k}$. Similarly, $(\bar{j} + \bar{k})\bar{i} = \overline{j+k} \cdot \bar{i} = \overline{(j+k)i} = \overline{ji + ki} = \overline{ji} + \overline{ki} = \bar{j} \cdot \bar{i} + \bar{k} \cdot \bar{i}$.

Furthermore, this is a commutative ring: $\bar{i} \cdot \bar{k} = \overline{ik} = \overline{ki} = \bar{k} \cdot \bar{i}$. □

In the proof we saw that the ring properties of $\mathbb{Z}/n\mathbb{Z}$, the set of equivalence classes modulo n, all follow from the ring properties of $\mathbb{Z}$. We shall later see that this can be generalized to any ring R: if we impose a correct notion of an equivalence relation, the equivalence classes themselves will form a ring and the fact that the ring axioms hold for it follows from the fact that they hold for R.

Theorem 12.0.5. *$\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime.*

Before providing the proof we introduce some terminology. Let R be a ring, $x \in R$ a non-zero element. x is called a **zero divisor** if there is an element $y \neq 0$ such that either $xy = 0$ or $yx = 0$ (or both).

Lemma 12.0.6. *Let R be a commutative ring. If R has zero divisors then R is not a field.*

Proof. Let $x \neq 0$ be a zero divisor and let $y \neq 0$ be an element such that $xy = 0$. If R is a field then there is an element $z \in R$ such that $zx = 1$. But then $z(xy) = z \cdot 0 = 0$ and also $z(xy) = (zx)y = 1 \cdot y = y$. So $y = 0$, and that is a contradiction. □

Proof. (Of Theorem) If $n = 1$ then $\mathbb{Z}/n\mathbb{Z}$ has a single element and so $0 = 1$ in that ring. Therefore, it is not a field. Suppose that $n > 1$ and n is not prime, $n = ab$ where $1 < a < n, 1 < b < n$. Then $\bar{a} \neq \bar{0}, \bar{b} \neq \bar{0}$ but $\bar{a} \cdot \bar{b} = \overline{ab} = \bar{n} = \bar{0}$. So $\mathbb{Z}/n\mathbb{Z}$ has zero divisors and thus is not a field.

Suppose now that n is prime and let $\bar{a} \neq \bar{0}$. That is, $n \nmid a$, which, since n is prime, means that $\gcd(n,a) = 1$. Consider the list of elements

$$\bar{0} \cdot \bar{a}, \bar{1} \cdot \bar{a}, \dots, \overline{n-1} \cdot \bar{a}.$$

We claim that they are distinct elements of $\mathbb{Z}/n\mathbb{Z}$. Suppose that $\bar{i} \cdot \bar{a} = \bar{j} \cdot \bar{a}$, for some $0 \leq i \leq j \leq n-1$ then $\overline{ia} = \overline{ja}$, which means that $n|(ia - ja) = (i-j)a$. Since $(n,a) = 1$, it follows that $n|(i-j)$ but that means $i = j$. Thus, the list $\bar{0} \cdot \bar{a}, \bar{1} \cdot \bar{a}, \dots, \overline{n-1} \cdot \bar{a}$ contains n distinct elements of $\mathbb{Z}/n\mathbb{Z}$ and so it must contain $\bar{1}$. That is, there's an i such that $\bar{i} \cdot \bar{a} = \bar{1}$ and therefore $\bar{a}$ is invertible.

Here is another proof for the invertibility of $\bar{a}$. Since $(a,n) = 1$, for suitable integers u, v we have $ua + vn = 1$. But that means that $n|(ua - 1)$. That is, $ua \equiv 1 \pmod{n}$.

The first proof has the advantage that a minor variant shows that every finite commutative ring with no zero-divisors is a field; the second proof has the advantage that in our particular situation one can use the Euclidean algorithm to calculate a u such that $ua \equiv 1 \pmod{n}$, namely to calculate a^{-1} in the field. We shall see that this is a tremendously useful fact. □

Let p be a prime number. We denote $\mathbb{Z}/p\mathbb{Z}$ also by $\mathbb{F}_p$. It is a field with p elements. It is a fact that any finite field (that is, any field with finitely many elements) has cardinality a power of a prime and for any prime power there is a field with that cardinality - we learn ways to construct such fields in this course. Finite fields, such as $\mathbb{F}_p$, play an important role in coding and cryptography as well as in pure mathematics. Interestingly enough, these fields were used to be called Galois fields and a notation such as $GF(p,n)$ (Galois field of p^n elements) was used. Initially, such fields were the height of abstraction; today we routinely use them in various cryptographic and coding theory implementations.

12.1. **Fermat's little theorem.**

Theorem 12.1.1. *(Fermat[17]) Let p be a prime number. Let $a \not\equiv 0 \pmod{p}$ then*

$$a^{p-1} \equiv 1 \pmod{p}.$$

Before proving the theorem we state two auxiliary statements whose proofs are left as an exercise.

Lemma 12.1.2. *Let p be a prime number. We have $p|\binom{p}{i}$ for every $1 \leq i \leq p-1$.*[18]

Lemma 12.1.3. *Let R be a commutative ring and $x, y \in R$. Interpret $\binom{n}{i}$ as adding the element 1 to itself $\binom{n}{i}$ times. Then the binomial formula holds in R:*

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}.$$

[17] Pierre de Fermat, 1601 - 1665, was a French lawyer and government official most remembered for his work in number theory; in particular for Fermat's Last Theorem that states the equation $x^n + y^n = z^n$ does not have solutions in positive integers x, y, z for any $n \geq 3$. He famously scribbled in the margin of his copy of Diophantus' *Arithmetica* book that he had discovered a marvellous proof that the margin is too small to contain. Fermat's last theorem was proved by Sir Andrew Wiles in 1994, ushering in the process a whole new area of number theory. Wiles' proof is so difficult, and builds on so much mathematics that didn't exist at Fermat's time, that Fermat's claim cannot be accepted as such. Fermat had also made important contributions to calculus and diophantine equations by introducing the so-called method of infinite descent. This method, still in use today, assumes that a given diophantine equation (namely, $f(x) = 0, f(x) \in \mathbb{Z}[x]$) has any solution in integers and attempts to use those to provide a solution in smaller integers (in absolute value). The process can then be repeated, yielding eventually a contradiction as one arrives at integer solutions of negative absolute value.

[18] Recall that the binomial coefficient $\binom{n}{a}$ is defined as follows. Let n, a be natural numbers. Define $\binom{n}{a} = \frac{n!}{a!\,(n-a)!}$, where $0! := 1$. This is in fact an integer and has the combinatorial interpretation that $\binom{n}{a}$ is the number of ways to choose a distinct objects from a set of n objects, where the order in which we choose the object doesn't matter. That is, it is the number of subsets of cardinality a of a set of cardinality n. This, incidentally, is a proof that $\binom{n}{a}$ is indeed an integer, and the reason the binomial number $\binom{n}{a}$ is often read "n choose a".

Proof. (of Fermat's little theorem) We prove the statement by induction on $1 \leq a \leq p-1$. For $a = 1$ the result is clear. Suppose the result for a and consider $a+1$, provided $a+1 < p$. We have, by the binomial formula,

$$\begin{aligned}(a+1)^p &= \sum_{i=0}^{p} \binom{p}{i} a^i \\ &= 1 + \binom{p}{1} a + \binom{p}{2} a^2 + \cdots + \binom{p}{p-1} a^{p-1} + a^p \\ &= 1 + a^p \qquad \text{(using the lemma)} \\ &= 1 + a \qquad \text{(using the induction hypothesis)}\end{aligned}$$

Since $1 + a \not\equiv 0 \pmod{p}$ it has an inverse y in $\mathbb{F}_p$, $y(1+a) \equiv 1$. Then, $y(1+a)^p \equiv y(1+a) \equiv 1$. As also $y(1+a)^p = y(1+a)(1+a)^{p-1} \equiv (1+a)^{p-1}$, we find that $(1+a)^{p-1} \equiv 1$. □

Example 12.1.4. We calculate 2^{100} modulo 13. We have $2^{100} = 2^{96}2^4 = (2^{12})^8 2^4 \equiv 2^4 \equiv 3$ modulo 13.

Fermat's little theorem gives a criterion for numbers to be composite. Let n be a positive integer. If there is $1 \leq a \leq n-1$ such that $a^{n-1} \not\equiv 1 \pmod{n}$ then n is not prime. Unfortunately, it is possible that for every $1 \leq a \leq n-1$ such that $(a,n) = 1$, one has $a^{n-1} \equiv 1 \pmod{n}$ and yet n is not prime. Thus, this test fails to recognize such n as composite numbers. Such numbers are called **Carmichael numbers**. Alford, Granville and Pomerance provved in 1994 that there are infinitely many such numbers. The first are 561, 1105, 1729, 2465, 2821, 6601, 8911, 10585, 15841, 29341, ... In 2022, David Larsen proved that a Carmichael number must always appear between x and $2x$ for any sufficiently large integer x.

Primality testing routines first test divisibility by small primes available to the program as pre-computed data and then use various methods that are more sophisticated version of the following method: choose randomly some $1 \leq a < n$: if $(a,n) \neq 1$ then n is not prime. If $(a,n) = 1$ the program calculated $a^{n-1} \pmod{n}$. If the result is not $1 \pmod{n}$ then n is not prime. If the result is 1, the program chooses another a. After a certain number of tests, say 10, if n passed all the tests it is declared as "prime", though there is no absolute reassurance it is indeed a prime; it could be a Carmichael number, or we could have just been unlucky in choosing our elements a. We remark that calculating $a^{n-1} \pmod{n}$ can be done quickly. One calculates $a, a^2, a^4, a^8, a^{16}, \cdots$ modulo n, as long as the power is less than n. This can be done rapidly. One then expresses n in base 2 to find the result. Here is an example: Let us calculate $3^{54} \pmod{55}$ (random choice of numbers). We have $3, 3^2 = 9, 3^4 = 81 = 26, 3^8 = 26^2 = 676 = 16, 3^{16} = 16^2 = 256 = 36, 3^{32} = 36^2 = 1296 = 31$. Now, $54 = 2 + 4 + 16 + 32$ and so $3^{54} = 9 \cdot 26 \cdot 36 \cdot 31 = 4$. In particular, 55 is not a prime – not that this is a particularly shrewd observation...

It is important to note that there is a polynomial-time algorithm (this means that the running time of the algorithm is at most some constant times the number of digits of the input) to decide, without any doubt, if an integer is prime. Such an algorithm was discovered by Agrawal, Kayal and Saxena in 2002 and was a real sensation at the time. It is important to note that the algorithm does not produce a decomposition of n in case n is composite. Such an algorithm will compromise the very backbone of e-commerce and military security.

12.2. **Solving equations in $\mathbb{Z}/n\mathbb{Z}$.**

12.2.1. *Linear equations.* We want to consider the equation $ax + b = 0$ in the ring $\mathbb{Z}/n\mathbb{Z}$. Let us assume that $\gcd(a,n) = 1$. Then, there are integers u, v such that $1 = ua + vn$. We remark that u, v are found by the Euclidean algorithm. Note that this implies that $ua \equiv 1 \pmod{n}$. Thus, if x solves $ax + b = 0$ in $\mathbb{Z}/n\mathbb{Z}$ then x solves the equation $uax + ub = 0 \pmod{n}$, that is $x + ub = 0$ and so $x = -ub$ in $\mathbb{Z}/n\mathbb{Z}$. Conversely, if $x = -ub$ in $\mathbb{Z}/n\mathbb{Z}$ where $ua = 1$ in $\mathbb{Z}/n\mathbb{Z}$ then $ax = a(-ub) = -aub = -b$ in $\mathbb{Z}/n\mathbb{Z}$.

We summarize: if $(a,n) = 1$ then the equation

$$ax + b = 0 \pmod{n},$$

has a unique solution $x = -ub$, where u is such that $ua = 1 \pmod{n}$.

Example 12.2.1. Here is a specific example: Let us solve $12x + 3 = 0 \pmod{17}$. First $17 = 12 + 5, 12 = 2 \cdot 5 + 2, 5 = 2 \cdot 2 + 1$, so $(12, 17) = 1$ and, moreover, $1 = 5 - 2 \cdot 2 = 5 - 2 \cdot (12 - 2 \cdot 5) = 5 \cdot 5 - 2 \cdot 12 = 5 \cdot (17 - 12) - 2 \cdot 12 = 5 \cdot 17 - 7 \cdot 12$. We see that $-7 \cdot 12 \equiv 1 \pmod{17}$. Thus, the solution is $x = 7 \cdot 3 = 21 = 4 \pmod{17}$.

More generally, if $(a, n) = d$, then the equation $ax + b \equiv 0 \pmod{n}$ has a solution if and only if $d|b$. Indeed, if we have a solution, then $b = kn - ax$ for some integer k and we see that $d|b$. Conversely, if $d|b$ consider the equation $(a/d)x + (b/d) \equiv 0 \pmod{n/d}$. A solution to this equation also solves the original equation $ax + b \equiv 0 \pmod{n}$. Now $(a/d, n/d) = 1$ and we are in the case already discussed above.

Unlike the situation $(a, n) = d$ where $d = 1$, if $d > 1$ there is more than one solution to $ax + b \equiv 0 \pmod{n}$. In fact, one can prove that if x_0 is one solution, then a complete set of solutions modulo n is

$$x_0,\ x_0 + \frac{n}{d},\ x_0 + 2 \cdot \frac{n}{d}, \ldots,\ x_0 + (d-1) \cdot \frac{n}{d}.$$

12.2.2. *Quadratic equations.* Consider the equation $ax^2 + bx + c = 0$ in $\mathbb{Z}/n\mathbb{Z}$ and assume n is a prime greater than 2. In that case, assuming that $a \neq 0$ modulo n, there is an element $(2a)^{-1}$. One can prove that the equation has a solution if and only if $b^2 - 4ac$ is a square in $\mathbb{Z}/n\mathbb{Z}$ (which may or may not be the case). In case it is a square, the solutions of this equation are given by the usual formula:

$$(2a)^{-1}(-b \pm \sqrt{b^2 - 4ac}).$$

For example, the equation $x^2 + x + 1$ has no solution in $\mathbb{Z}/5\mathbb{Z}$ because the discriminant $b^2 - 4ac$ is in this case $1^2 - 4 = -3 = 2$ in $\mathbb{Z}/5\mathbb{Z}$ and 2 can be checked not to be a square in $\mathbb{Z}/5\mathbb{Z}$ (one just tries: $0^2 = 0, 1^2 = 1, 2^2 = 4, 3^2 = 4, 4^2 = 1$ in $\mathbb{Z}/5\mathbb{Z}$). On the other hand, $x^2 + x + 1$ can be solved in $\mathbb{Z}/7\mathbb{Z}$. The solutions are $4(-1 \pm \sqrt{-3}) = -4 \pm 4\sqrt{4} = -4 \pm 8 = -4 \pm 1 = \{2, 4\}$.

When n is not prime, we shall not study the problem in this course, beyond remarking that one can proceed by trying all possibilities if n is small and that the number of solutions can be very large. For example: consider the equation $x^3 - x$ in $\mathbb{Z}/8\mathbb{Z}$. We can verify that its solutions are $0, 1, 3, 5, 7$. There are 5 solutions but the equation has degree three. We shall later see that in any **field** a polynomial equation of degree n has at most n roots. As the example of $x^3 - x$ in $\mathbb{Z}/8\mathbb{Z}$ suggests, this may fail spectacularly in a general commutative ring.

12.3. **Public key cryptography; the RSA method.** [19] We cannot go here too much into the cryptographical practical aspects. Suffices to say that in many cryptographical applications two parties X and Y wish to exchange a secret. Given any large integer n that secret can be represented as a number modulo n, and we leave it to the reader's imagination to devise methods for that. The method proceeds as follows:

X chooses two large primes $p < q$.

X calculates $n = pq$.

X calculates $k = (p-1)(q-1)$.

X chooses an integer d such that $(d, k) = 1$.

X finds an integer e such that $ed \equiv 1 \pmod{k}$.

X publishes for anyone to see the data e, n. This is called the **public key**.

The rest of the data p, q, k, d is kept secret. In fact, p, q, k can be destroyed altogether and only d is kept, and kept secret!

[19] The RSA method described above is named after Ron Rivest, Adi Shamir and Len Adleman, who discovered it in 1977.

d is called the **private key**.

Y, wishing to send a secret, writes it as an integer b modulo n, which is also relatively prime to n, and sends $b^e \pmod n$ to X, allowing anyone interested to see that message. The point is, and this is called the **discrete log problem**, that it is very difficult to find what b is, even when one knows b^e and n. Thus, someone seeing Y's message cannot find the secret b from it.

X, upon receiving Y's message b^e, calculates $(b^e)^d$.

Lemma 12.3.1. *We have* $b^{ed} \equiv b \pmod n$.

Proof. We need to show that $b^{ed} \equiv b \pmod p$ and $b^{ed} \equiv b \pmod q$. Then $p|(b^{ed}-b)$ and $q|(b^{ed}-b)$ and so (using the p,q are primes and distinct), $n = pq|(b^{ed}-b)$.

The argument being symmetric, we just show $b^{ed} \equiv b \pmod p$. We have modulo p

$$\begin{aligned} b^{ed} &= b^{1+vk} \\ &= b \cdot ((b^{p-1})^{q-1})^v \\ &= b \cdot (1^{q-1})^v \qquad \text{(Fermat's little theorem)} \\ &= b. \end{aligned}$$

□

We have shown that X can retrieve Y's secret.

Here is a numerical example:

```
p = 10007, q = 10009;
n = p*q = 100160063
k = (p-1)*(q-1) = 100140048
d = 10001
e = 88695185
b = 3
b^e = 33265563
33265563^10001 = 3 Mod n.
```

RSA rests on several security assumptions. Besides the belief that the discrete log problem is inherently difficult, it also supposes that the difficulty of factoring an integer of the form pq, where p and q are primes of similar size is a difficult problem computationally whose solution has running time comparable to the size of p. All evidence so far indicates that this is so. Suppose that p is a prime of the order of magnitude of 10^{1000} (at this point in time (2023) this is still considered very secure). Let us calculate how long it will take to factor n by brute force trial and error. Cray's supercomputer *Titan* does 20,000 trillion (2×10^{15}) flops (calculations) per second. It had cost 97 million dollars to be built. To simplify, let us assume that this many flops boils down to trying 2×10^{15} integers as factors of n per second.

We can assume that such a computer can't be built for less that 10 million dollars. A computer 10000 times faster will probably cost today in the excess of 10 billion dollars, if it can be constructed at all. The existence of anything bigger would probably be public knowledge for budgetary reasons and the size of resources needed for constructing and running such a computer. Even with dedicated architecture it wouldn't perform more than 10^{25} operations per second. The time to factor an integer greater than 10^{1000} by brute force will take more than 10^{975} seconds, which is much much more than a billion billion years. Even relaxing our assumptions greatly shows that for all practical purposes, as long as our security assumptions are valid, it is not feasible to break RSA based on a key of this size in any feasible time.

13. Exercises

(1) Given an integer N, we write N in decimal expansion as $N = n_k n_{k-1} \dots n_0$, the n_i being the digits of N. Note that this means that $N = n_0 + 10n_1 + 10^2 n_2 + \dots + 10^k n_k$. In the following you are asked to show certain divisibility criteria that can be proved by using congruences.
   (a) Prove that a positive integer $N = n_k n_{k-1} \dots n_0$ is divisible by 3 if and only if the sum of its digits $n_0 + n_1 + \dots + n_k$ is divisible by 3. (Hint: show that in fact N and $n_0 + n_1 + \dots + n_k$ are congruent to the same number modulo 3.) Example: 34515 is divisible by 3 because $3 + 4 + 5 + 1 + 5 = 18$ is divisible by 3.
   (b) Prove that a positive integer $N = n_k n_{k-1} \dots n_0$ is divisible by 11 if and only if the sum of its digits with alternating signs $n_0 - n_1 + n_2 - \dots + (-1)^k n_k$ is divisible by 11. Example: 1234563 is divisible by 11 since $1 - 2 + 3 - 4 + 5 - 6 + 3 = 0$ is divisible by 11.
   (c) Prove that a positive integer $N = n_k n_{k-1} \dots n_0$ is divisible by 7 if and only if when we let $M = n_k n_{k-1} \dots n_1$, we have that $M - 2n_0$ is divisible by 7. Example: take the number $7 * 11 * 13 * 17 = 17017$. It is clearly divisible by 7. Let us check the criterion against this example. We form the number $1701 - 2 * 7 = 1687$ and then the number $168 - 2 * 7 = 154$ and then the number $15 - 2 * 4 = 7$. So it works. Let us also check the number 82. It is not divisible by 7, in fact it's residue modulo 7 is 5. Also $8 - 2 * 2 = 4$, so the criterion shows that it's not divisible. Note though that in this case the number $N = 82$ and the number $M = 8 - 2 * 2 = 4$ don't have the same residue modulo 7. So you need to construct your argument a little differently than in the previous exercises.

(2) To check if you had multiplied correctly two large numbers A and B, $A \times B = C$, you can make the following check: sum the digits of A; keep doing it repeatedly until you get a single digit number a. Do the same for B and C and get numbers b, c. If you have multiplied correctly, the sum of digits of ab is c. Prove that this is so. This is called in French "preuve par neuf".

   Example: I have multiplied $A = 367542$ by $B = 687653$ and got $C = 252741358926$. To check (though this doesn't prove the multiplication is correct) I do: $3 + 6 + 7 + 5 + 4 + 2 = 27, 2 + 7 = 9$ and $a = 9$. Also $6 + 8 + 7 + 6 + 5 + 3 = 35, 3 + 5 = 8$ and $b = 8$. $ab = 72$ and its sum of digits is 9. On the other hand $2 + 5 + 2 + 7 + 4 + 1 + 3 + 5 + 8 + 9 + 2 + 6 = 54, 5 + 4 = 9$. So it checks.

(3) Calculate the expression $\frac{3 \cdot 5 - 3^3}{2 \cdot 6 + 10}$ in $\mathbb{Z}/5\mathbb{Z}$ and in $\mathbb{Z}/7\mathbb{Z}$. Note: we write $\frac{1}{a}$ to denote the inverse a^{-1} of a with respect multiplication in the field. For example, in $\mathbb{Z}/5\mathbb{Z}$, $\frac{1}{3} = 2$ and in $\mathbb{Z}/7\mathbb{Z}$, $\frac{1}{3} = 5$. The expression $\frac{a}{b}$ means $a \cdot b^{-1}$.

(4) (a) Find all solutions to the equation $x^2 + x = 0$ in $\mathbb{Z}/p\mathbb{Z}$, where p is prime.
   (b) Find all solutions to the equation $x^2 + x = 0$ in $\mathbb{Z}/6\mathbb{Z}$.

(5) Solve the following equations:
   (a) $12x = 2$ in $\mathbb{Z}/19\mathbb{Z}$.
   (b) $7x = 2$ in $\mathbb{Z}/24\mathbb{Z}$.
   (c) $31x = 1$ in $\mathbb{Z}/50\mathbb{Z}$.
   (d) $34x = 1$ in $\mathbb{Z}/97\mathbb{Z}$.
   (e) $27x = 2$ in $\mathbb{Z}/40\mathbb{Z}$.
   (f) $15x = 5$ in $\mathbb{Z}/63\mathbb{Z}$.

(6) (a) Let $p > 2$ be a prime. Prove that an equation of the form $ax^2 + bx + c$ (where $a, b, c \in \mathbb{F}_p, a \neq 0$) has a solution in $\mathbb{Z}/p\mathbb{Z}$ if and only if $b^2 - 4ac$ is a square in $\mathbb{Z}/p\mathbb{Z}$. If this is so, prove that the solutions are given by the familiar formula.
   (b) Determine for which values of a the equation $x^2 + x + a$ has a solution in $\mathbb{Z}/7\mathbb{Z}$.

(7) Calculate the following:
   (a) $(2^{19808} + 6)^{-1} + 1 \pmod{11}$.
   (b) $12, 12^2, 12^4, 12^8, 12^{16}, 12^{25}$ all modulo 29. (Hint: think how to proceed most efficiently before computing).

(8) Let p be a prime number, prove that $p|\binom{p}{i}$ for every $1 \leq i \leq p-1$.

(9) Let R be a commutative ring and $x, y \in R$. Interpret $\binom{n}{i}$ as the sum in R of $\binom{n}{i}$ times the element 1. Then the binomial formula holds in R:

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}.$$

(10) Find all values a for which the system of equations $xy = a, x + y = 1$ has a solution in $\mathbb{Z}/19\mathbb{Z}$. Is there such an a for which there is a unique solution?

(11) Prove Wilson's theorem: Let p be a prime number then $(p-1)! \equiv -1 \pmod{p}$.

(12) Let p be an odd prime and consider the sum

$$\frac{m}{n} := 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{p-1}.$$

Prove that $p|m$. One method starts by multiplying this sum by $(p-1)!$. The other method considers the ring $\mathbb{Z}[1/p] := \{\frac{a}{b} : p \nmid b, a, b \in \mathbb{Z}\}$. As p is an element of this ring, we can talk about divisibility by p and define for two elements x, y in this ring that $x \equiv y \pmod{p}$ if $p|(x-y)$. Show that there is a well-defined commutative ring structure on the set of congruence classes. Show that the residue classes are represented once more by $0, 1, \ldots, p-1$. Let a be an integer not congruent to 0 modulo p. There is an integer b such that $ab \equiv 1 \pmod{p}$; namely, the congruence class of b is the inverse of a in the field $\mathbb{Z}/p\mathbb{Z}$. Show that $b \equiv \frac{1}{a} \pmod{p}$, where this congruence is taking place in the ring $\mathbb{Z}[1/p]$. With these preparations, prove that $p|m$.

(13) Provide another proof of Fermat's little theorem: *Let p be a prime number and $a \not\equiv 0 \pmod{p}$ then* $a^{p-1} \equiv 1 \pmod{p}$.

Start the proof by showing that the residue classes $a, 2a, 3a, \ldots, (p-1)a$ are distinct and non-zero, and deduce that $(p-1)! \cdot a^{p-1} \equiv (p-1)! \pmod{p}$.

Part 4. Polynomials and their arithmetic

Some of our main achievements in studying the arithmetic of $\mathbb{Z}$ are the following:

- Division with residue, greatest common divisor and the Euclidean algorithm.
- Primes and unique factorization.
- Construction of new rings, the rings $\mathbb{Z}/n\mathbb{Z}$, out of $\mathbb{Z}$. In particular, the construction of fields with p elements, for every prime number p.

In this section and the following one, we shall see that these constructions, notions and theorems hold true also for the ring of polynomials over a field. One particular application will be, once we sufficiently develop the theory, the construction of finite fields with p^n elements, where p is a prime and $n \geq 1$ an integer. One should note that the only ring with p^2 elements that we know so far, the ring $\mathbb{Z}/p^2\mathbb{Z}$, is never a field. And so at this point it is not clear at all whether fields with p^2 elements, say, even exist, let alone how to calculate in them. All that we will be able to do.

14. The ring of polynomials

Let R be a commutative ring. A good example to keep in mind is $R = \mathbb{Z}$ or $R = \mathbb{C}$, but our discussion allows any commutative ring, so rings like $\mathbb{Z}/n\mathbb{Z}$ and $\mathbb{Z}[i] = \{a + bi \in \mathbb{C} : a, b \in \mathbb{Z}\}$ are perfectly good examples as well. We define the **ring of polynomials** over R as

$$R[x] = \{a_n x^n + \cdots + a_1 x + a_0 : a_i \in R\}.$$

In the definition n is any non-negative integer. $a_n x^n + \cdots + a_1 x + a_0$ is called a polynomial with coefficients in R and the a_i are called the coefficients. Note that we allow some, or even all, coefficients to be zero. We implicitly identify polynomials of the form $a_n x^n + \cdots + a_1 x + a_0$ and $0 \cdot x^{n+m} + \cdots + 0 \cdot x^{n+1} + a_n x^n + \cdots + a_1 x + a_0$, and we often write, say, $x^5 + x^3 + x$ instead of $x^5 + 0 \cdot x^4 + 1 \cdot x^3 + 0 \cdot x^2 + 1 \cdot x + 0$. The **zero polynomial** 0 is the choice $n = 0$ and $a_0 = 0$. We define addition by adding coefficients (assume $n \geq m$; a similar formula holds when $n \leq m$)

$$(a_n x^n + \cdots + a_1 x + a_0) + (b_m x^m + \cdots + b_1 x + b_0) \\ = a_n x^n + \cdots + (a_m + b_m) x^m + \cdots + (a_1 + b_1) x + (a_0 + b_0).$$

If you wish, by adding zero coefficients, we can always assume that our polynomials are both of the form $a_n x^n + \cdots + a_1 x + a_0$ and $b_n x^n + \cdots + b_1 x + b_0$, in which case we can write their sum as

$$(a_n x^n + \cdots + a_1 x + a_0) + (b_n x^n + \cdots + b_1 x + b_0) = (a_n + b_n) x^n + \cdots + (a_1 + b_1) x + (a_0 + b_0).$$

We also define multiplication by

$$(a_n x^n + \cdots + a_1 x + a_0)(b_m x^m + \cdots + b_1 x + b_0) = c_{n+m} x^{n+m} + \cdots + c_1 x + c_0,$$

where

$$c_i = a_0 b_i + a_1 b_{i-1} + \cdots + a_{i-1} b_1 + a_i b_0.$$

Note that in the formula for c_i it is entirely possible that some a_j or b_j, $0 \leq j \leq i$, are not defined; this happens if $j > n$ or $j > m$, respectively. In this case we understand a_j, or b_j, as zero.

Example 14.0.1. Take $R = \mathbb{Z}$. Then,

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

and

$$(2x^2 + x - 2)(x^3 + x - 1) = 2x^5 + x^4 - x^2 - 3x + 2.$$

A polynomial $f(x) = a_n x^n + \cdots + a_1 x + a_0$ is called **monic** if $a_n = 1$. It is called of **degree** n if $a_n \neq 0$. If f has degree 0, that is $f(x) = a, a \in R, a \neq 0$, then f is called a **constant polynomial**. The degree of the zero polynomial is not defined, but it is also considered as a constant polynomial.

Proposition 14.0.2. *With the operations defined above $R[x]$ is a commutative ring, with zero being the zero polynomial and 1 being the constant polynomial 1. The additive inverse of $a_n x^n + \cdots + a_1 x + a_0$ is $-a_n x^n - \cdots - a_1 x - a_0$.*

Since the proof is straightforward we leave it as an exercise.

A commutative ring R is called an **integral domain** if it is not the zero ring and whenever $x, y \in R$ satisfy $xy = 0$ then either $x = 0$ or $y = 0$ (or both).

Proposition 14.0.3. *If R is an integral domain then $R[x]$ is an integral domain. If $f(x), g(x) \in R[x]$ are non-zero polynomials,*

$$\deg(f(x) \cdot g(x)) = \deg(f(x)) + \deg(g(x)).$$

Proof. Say $\deg(f(x)) = n$, $\deg(g(x)) = m$. Then, by definition, $f(x) = a_n x^n + \cdots + a_1 x + a_0$ with $a_n \neq 0$ and $g(x) = b_m x^m + \cdots + b_1 x + b_0$ with $b_m \neq 0$. Therefore,

$$f(x)g(x) = a_n b_m x^{n+m} + (a_n b_{m-1} + a_{n-1} b_m) x^{n+m-1} + \cdots .$$

Since R is an integral domain $a_n b_m \neq 0$, and so $f(x)g(x) \neq 0$ and $\deg(f(x) \cdot g(x)) = n + m$. □

Let $\mathbb{F}$ be a field. We shall see that there are many similarities between the ring of integers $\mathbb{Z}$ and the ring of polynomials with coefficients in $\mathbb{F}$, $\mathbb{F}[x]$. Initially, we shall see that they share many notions concerning divisibility. The notion of absolute value for $\mathbb{Z}$ which measures the size of an integer is replaced by the notion of degree for polynomials, which is the "size" of a polynomial. Using this notion we will define division with residue and a Euclidean algorithm. Later on, after introducing some ideas from ring theory, we will also construct the analog of the rings of congruence classes $\mathbb{Z}/n\mathbb{Z}$ ($n \geq 1$ an integer) and in particular of the fields $\mathbb{Z}/p\mathbb{Z}$ (p a prime number).

There is a class of rings R called **Euclidean rings** that includes both $\mathbb{Z}$ and $\mathbb{F}[x]$, where $\mathbb{F}$ is a field. For such rings one can likewise divide division with residue, gcd's, Euclidean algorithm, unique factorization and so on. Such rings are commutative integral domains with a "size function", with properties analogous to the absolute value and degree, and that what makes everything "tick". We will not develop the general theory here, except for remarking that the proofs go exactly the same.

15. Division with residue

Let $\mathbb{F}$ be a field. We have defined the ring of polynomials $\mathbb{F}[x]$; it is an integral domain but is never a field; for example x does not have an inverse with respect to multiplication.

Theorem 15.0.1. *Let $f(x), g(x)$ be two polynomials in $\mathbb{F}[x]$, $g(x) \neq 0$. Then, there exist unique polynomials $q(x), r(x)$ in $\mathbb{F}[x]$ such that*

$$f(x) = q(x)g(x) + r(x), \qquad r(x) = 0 \text{ or } \deg r(x) < \deg g(x).$$

Proof. We first show the existence and later the uniqueness. Consider the set

$$S = \{f(x) - q(x)g(x) : q(x) \in \mathbb{F}[x]\}.$$

If $0 \in S$ then there is a $q(x)$ such that $f(x) = q(x)g(x)$ and we take $r(x) = 0$. Else, choose an element $r(x)$ in S of minimal degree. Since $r(x)$ is in S we can write $r(x) = f(x) - q(x)g(x)$ for some $q(x)$.
Claim. $\deg r(x) < \deg g(x)$.
Let us write $r(x) = r_n x^n + \cdots r_1 x + r_0$ and $g(x) = g_m x^m + \cdots + g_1 x + g_0$, with $r_n \neq 0, g_m \neq 0$. Assume, by contradiction, that $n \geq m$. Then, the polynomial $r_1(x) = r(x) - r_n g_m^{-1} x^{n-m} g(x) = (r_{n-1} - r_n g_m^{-1} g_{m-1}) x^{n-1} + \cdots$ has degree smaller then $r(x)$. On the other hand, the expression $r_1(x) = r(x) - r_n g_m^{-1} x^{n-m} g(x) = f(x) - q(x)g(x) - r_n g_m^{-1} x^{n-m} g(x) = f(x) - (q(x) + r_n g_m^{-1} x^{n-m}) g(x)$ shows that $r_1(x) \in S$. Contradiction. We have therefore established the existence of $q(x), r(x)$ such that

$$f(x) = q(x)g(x) + r(x), \qquad r(x) = 0 \text{ or } \deg r(x) < \deg g(x).$$

We now prove uniqueness. Suppose that also

$$f(x) = q_1(x)g(x) + r_1(x), \qquad r_1(x) = 0 \text{ or } \deg r_1(x) < \deg g(x).$$

We need to show that $q(x) = q_1(x), r(x) = r_1(x)$. We have,

$$(q(x) - q_1(x))g(x) = r_1(x) - r(x).$$

The right hand side is either zero or has degree less than $g(x)$. If it's zero then, since $\mathbb{F}[x]$ is an integral domain, we also have $q(x) = q_1(x)$. If $r(x) \neq r_1(x)$ then also $q(x) \neq q_1(x)$ but then the degree of the left hand side is $\deg(q(x) - q_1(x)) + \deg(g(x)) \geq \deg(g(x)) > \deg(r_1(x) - r(x))$ and we get a contradiction. □

16. Arithmetic in $\mathbb{F}[x]$

In this section $\mathbb{F}$ is a field. We denote by $\mathbb{F}^\times$ the set of non-zero elements of $\mathbb{F}$.

16.1. **Some remarks about divisibility in a commutative ring T.** The definitions we made in § 9.2 can be made in general and the same basic properties hold. Let T be a commutative ring and $a, b \in T$. We say that a **divides** b if $b = ac$ for some $c \in T$. We have the following properties:

(1) $a|b \Rightarrow a| - b$.
(2) $a|b \Rightarrow a|bd$ for any $d \in T$.
(3) $a|b, a|d \Rightarrow a|(b \pm d)$.

In particular, the definition and properties hold for the ring of polynomials $R[x]$, where R is a commutative ring.

16.2. **GCD of polynomials.**

Definition 16.2.1. Let $f(x), g(x) \in \mathbb{F}[x]$, not both zero. The **greatest common divisor** of $f(x)$ and $g(x)$, denoted $\gcd(f(x), g(x))$ of just $(f(x), g(x))$, is a monic polynomial of largest degree dividing both $f(x)$ and $g(x)$. (We shall see below that there is a unique such polynomial.)

Theorem 16.2.2. *Let $f(x), g(x)$ be polynomials, not both zero. The gcd of $f(x)$ and $g(x)$, $h(x) = (f(x), g(x))$, is unique and can be expressed as*

$$h(x) = u(x)f(x) + v(x)g(x), \qquad u(x), v(x) \in \mathbb{F}[x].$$

It is the monic polynomial of minimal degree having such an expression. If $t(x)$ divides both $g(x)$ and $f(x)$ then $t(x)|h(x)$.

Proof. Consider the following set of monic polynomials

$$S = \{a(x) : a(x) = u(x)f(x) + v(x)g(x) \text{ for some } u(x), v(x) \in \mathbb{F}[x], a(x) \text{ monic}\}.$$

S contains a non-zero polynomial, because if $f(x) \neq 0$, $f(x) = bx^n + l.o.t.$[20], then $b^{-1}f(x) \in S$; if $f(x) = 0$ then $g(x)$ is not zero and the same argument can be applied to $g(x)$. Let $h(x)$ be an element of minimal degree of S. We claim that $h(x)$ divides both $f(x)$ and $g(x)$. Since the situation is symmetric, we just prove $h(x)|f(x)$. Suppose not, then we can write $f(x) = q(x)h(x) + r(x)$, where $r(x)$ is a non-zero polynomial of degree smaller than $h(x)$. Then $r(x) = f(x) - q(x)(u(x)f(x) + v(x)g(x)) = (1 - q(x)u(x)) \cdot f(x) - q(x)v(x) \cdot g(x)$ and so, if we let $r_1(x)$ be $r(x)$ divided by its leading coefficient, we see that $r_1(x) \in S$ and has degree smaller than $h(x)$, which is a contradiction.

By construction, $h(x)$ is the monic polynomial of minimal degree having such an expression. If $t(x)$ divides both $g(x)$ and $f(x)$ then $t(x)|(u(x)f(x) + v(x)g(x)) = h(x)$. Therefore, $h(x)$ is a monic polynomial of the largest possible degree dividing both $f(x), g(x)$. Suppose that $h_1(x)$ is another monic polynomial dividing $f(x)$ and $g(x)$ having the largest possible degree, i.e., the degree of $h(x)$. Then, we have $h(x) = h_1(x)b(x)$ by what we proved. Since both polynomials have the same degree $b(x)$ must be a constant polynomial, and, then, since both are monic, $b(x) = 1$. We have shown the gcd is unique. □

[20] l.o.t. = lower order terms.

16.3. **The Euclidean algorithm for polynomials.**

Theorem 16.3.1. *Let $f(x), g(x) \in \mathbb{F}[x]$ be non-zero polynomials, $g(x) = a_n x^n + l.o.t.$ If $g(x)|f(x)$ then $(f(x), g(x)) = a_n^{-1} g(x)$. Else, define inductively,*

$$\begin{aligned}
&f(x) = q_0(x)g(x) + r_0(x), && \deg(r_0) < \deg(g)\\
&g(x) = q_1(x)r_0(x) + r_1(x), && \deg(r_1) < \deg(r_0)\\
&r_0(x) = q_2(x)r_1(x) + r_2(x), && \deg(r_2) < \deg(r_1)\\
&\vdots\\
&r_{t-2}(x) = q_t(x)r_{t-1}(x) + r_t(x), && \deg(r_t) < \deg(r_{t-1})\\
&r_{t-1}(x) = q_{t+1}(x)r_t(x).
\end{aligned}$$

This is indeed possible, and the process always terminates.
Letting $r_t(x) = c_m x^m + \cdots + c_0$, we have

$$(f(x), g(x)) = c_m^{-1} r_t(x).$$

Moreover, this algorithm also allows expressing $(f(x), g(x))$ in the form $u(x)f(x) + v(x)g(x)$.

Proof. Each step in the process is done based on Theorem 15.0.1. The process must terminate because the degrees decrease and they are natural numbers.

It is easy to see that $r_t | r_{t-1}$. Suppose we know r_t divides $r_{t-1}, r_{t-2}, \ldots, r_a$ then, since $r_{a-1} = q_{a+1}r_a + r_{a+1}$ we get also that $r_t | r_{a-1}$. We conclude that r_t divides $r_0, r_1, \ldots, r_t$. Exactly the same argument gives that r_t divides $g(x)$ and $f(x)$.

Conversely, if $a(x)$ divides $f(x)$ and $g(x)$ then $a(x)|(f(x) - q_0(x)g(x)) = r_0(x)$ and therefore $a(x)|(g(x) - q_1(x)r_0(x)) = r_1(x)$, etc. We see that $a(x)|r_t(x)$ and so $r_t(x)$, once divided by its leading coefficient, must be the greatest common divisor of $f(x)$ and $g(x)$. □

Example 16.3.2. (1) $f(x) = x^2 + 1$, $g(x) = x^2 + 2ix - 1$, complex polynomials. We have

$$\begin{aligned}
f(x) &= 1 \cdot (x^2 + 2ix - 1) + (-2ix + 2)\\
(x^2 + 2ix - 1) &= (\frac{1}{-2i}x - \frac{1}{2})(-2ix + 2).
\end{aligned}$$

It follows that $(f(x), g(x)) = \frac{1}{-2i}(-2ix + 2) = x + i$. This implies that $-i$ is a root of both polynomials, as one can verify.

(2) Now we choose $\mathbb{F} = \mathbb{Z}/3\mathbb{Z}$, the field with 3 elements. We take $f(x) = x^3 + 2x + 1, g(x) = x^2 + 1$. We then have,

$$\begin{aligned}
f(x) &= x \cdot (x^2 + 1) + (x + 1)\\
(x^2 + 1) &= (x - 1) \cdot (x + 1) + 2\\
x + 1 &= (2x + 2) \cdot 2.
\end{aligned}$$

This implies that $(f(x), g(x)) = 1$. We have

$$\begin{aligned}
2 &= (x^2 + 1) - (x - 1) \cdot (x + 1)\\
&= g(x) - (x - 1)(f(x) - xg(x))\\
&= (-x + 1)f(x) + (x^2 - x + 1)g(x).
\end{aligned}$$

And so we find (note that $1 = -2$ in $\mathbb{F}$)

$$1 = (f(x), g(x)) = (x - 1)f(x) - (x^2 - x + 1)g(x).$$

(3) Consider the polynomials $f(x) = x^3 + 5x^2 + 4x, g(x) = x^3 + x^2 - x - 1$, as rational polynomials. Then

$$f(x) = 1 \cdot g(x) + 4x^2 + 5x + 1$$
$$x^3 + x^2 - x - 1 = (\frac{1}{4}x - \frac{1}{16})(4x^2 + 5x + 1) - \frac{15}{16}x - \frac{15}{16}$$
$$(4x^2 + 5x + 1) = \frac{-16}{15}(4x + 1)(-\frac{15}{16}x - \frac{15}{16}).$$

It follows that $(f(x), g(x)) = x + 1$.

To express $x + 1$ as $u(x)f(x) + v(x)g(x)$ we work backwards:

$$\begin{aligned}\frac{-15}{16}(x+1) &= g(x) - (\frac{1}{4}x - \frac{1}{16})(4x^2 + 5x + 1)\\ &= g(x) - (\frac{1}{4}x - \frac{1}{16})(f(x) - g(x))\\ &= -(\frac{1}{4}x - \frac{1}{16}) \cdot f(x) + (\frac{15}{16} + \frac{1}{4}x) \cdot g(x).\end{aligned}$$

Thus,

$$x + 1 = (f(x), g(x)) = (-\frac{1}{15} + \frac{4}{15}x) \cdot f(x) - (1 + \frac{4}{15}x) \cdot g(x).$$

(4) Now consider the same polynomials over the field $\mathbb{F} = \mathbb{Z}/3\mathbb{Z}$. We now have:

$$f(x) = 1 \cdot g(x) + x^2 + 2x + 1$$
$$x^3 + x^2 - x - 1 = (x - 1)(x^2 + 2x + 1).$$

Therefore, now we have $(f(x), g(x)) = x^2 + 2x + 1 = (x + 1)^2$.

One interesting application of the Euclidean algorithm is that it allows to check whether two polynomials $f, g \in \mathbb{C}[x]$ (for example) have a common root, although it would not find the root. The point is that if $f(\alpha) = g(\alpha) = 0$ then $(x - \alpha)|f(x)$ and $(x - \alpha)|g(x)$. Indeed, divide $f(x)$ by $x - \alpha$ with residue: $f(x) = q(x)(x - \alpha) + r(x)$. Division with residue gives that $r(x)$ is a constant polynomial. On the other hand, substitute $x = \alpha$ to see that $r(\alpha) = f(\alpha) - q(\alpha)(\alpha - \alpha) = 0$. Thus, $r(x) = 0$ and $(x - \alpha)|f(x)$; similarly, $(x - \alpha)|g(x)$. Thus, if f and g have a common root α then $(x - \alpha)|(f, g)$ and so $(f, g) \neq 1$. Conversely, suppose that $d(x) := (f, g) \neq 1$. As $d(x)$ is a non-constant polynomial it has a root $\alpha \in \mathbb{C}$ and as $f(x) = d(x)q_1(x), g(x) = d(x)q_2(x)$ for some $q_i(x) \in \mathbb{C}[x]$, it follows that $f(\alpha) = g(\alpha) = 0$.

The interesting point about this use of the Euclidean algorithm is that it provides a *quick and efficient answer* to the question of whether two complex polynomials f and g have a common root. For a general field $\mathbb{F}$, it reduces the question to whether $\gcd(f, g)$ has a root in $\mathbb{F}$. It should be stressed the straightforward approach, namely "calculate the roots of f and g and check if one of them is the same" usually fails, because there is no general method to find in exact form the roots of polynomials of high degree.

16.4. **Irreducible polynomials and unique factorization.** Let $\mathbb{F}$ be a field. We define a relation on polynomials $f(x) \in \mathbb{F}[x]$. We say that $f(x) \sim g(x)$ if there is an element $a \in \mathbb{F}, a \neq 0$ such that $f(x) = ag(x)$.

Lemma 16.4.1. *This relation is an equivalence relation. Related polynomials are called* **associates**. *The associates of* 1 *are the non-zero scalars* $\mathbb{F}^\times$.

Proof. The relation is reflexive because $f(x) = 1 \cdot f(x)$ and symmetric, because $f(x) = ag(x)$ implies $g(x) = a^{-1}f(x)$. It is also transitive since $f(x) = ag(x)$ and $g(x) = bh(x)$ implies $f(x) = abh(x)$ and $ab \neq 0$. Finally, the last claim follows immediately from the definition. □

The motivation for this definition of being "associated" is that as far as division goes two associated polynomials behave the same. Indeed, suppose that $f \sim g$, say $f = ag$. If $f|h$, that is, if $h = ff_1$ for some polynomial f_1, then $h = g(af_1)$ and that shows that $g|h$ as well. The argument can be reversed and one conclude that if $g|h$ then $f|h$. It is also easy to check that if $h|f$ then $h|g$ and, conversely, if $h|g$ then $h|f$.

Why did this issue not come up for integers?! Well, in a sense it was always there but it was so "visible" that it required no special definition. The units of $\mathbb{Z}$ are just ± 1 and in this case the correct definition is to say that two integers a, b are associate if $a = \pm b$. Under this definition it is clear that $a|c \Leftrightarrow b|c$ and $c|a \Leftrightarrow c|b$.

Let us return to polynomials. A non-constant polynomial f is called **irreducible** if $g|f$ implies that $g \sim 1$ or $g \sim f$. As we have noted, if $g|f$ and $g_1 \sim g$ then $g_1|f$. Therefore, trying to define "irreducible" by $g|f$ implies that $g(x) = 1$, or $g(x) = f(x)$, will not make sense; it will be in general a too strong requirement.

Proposition 16.4.2. *Let $f(x) \in \mathbb{F}[x]$ be a non-constant polynomial. The following are equivalent:*

(1) f is irreducible.

(2) If $f|gh$ then $f|g$ or $f|h$.

Proof. Suppose that f is irreducible, $f|gh$ and $f \nmid g$. The only monic polynomials dividing f are 1 and $a^{-1}f$, where a is the leading coefficient of f. Therefore, $(f, g) = 1$ and so, for suitable polynomials u, v we have $uf + vg = 1$. Then $ufh + vgh = h$. Since f divides the left hand side, it also divides the right hand side, i.e., $f|h$.

Suppose now that f has the property $f|gh \Rightarrow f|g$ or $f|h$. Let g be a divisor of f. Then $f = gh$ for some h and so $f|gh$. Therefore, $f|g$ or $f|h$. Since $h|f$ too, the situation concerning whether $f|g$ or $f|h$ is entirely symmetric and we can assume that $g|f$ and $f|g$. This implies that $\deg(g) \leq \deg(f)$ and $\deg(f) \leq \deg(g)$, and so $\deg(f) = \deg(g)$. But then $\deg(h) = \deg(f) - \deg(g) = 0$ and so h is a constant polynomial. We find that $f \sim g$. □

Example 16.4.3. Here are some comments on irreducible polynomials.

(1) Every linear polynomial is irreducible.

(2) If $f = ax^2 + bx + c$ is reducible then it is divisible by some linear polynomial, hence by some monic linear polynomial, say $(x + \alpha)$. Thus, $f = (x + \alpha)(ax + \delta)$, where $\alpha, \delta \in \mathbb{F}$. It follows then that f has a root in $\mathbb{F}$, for example $x = -\alpha$.

Conversely, suppose that f has a root $\alpha \in \mathbb{F}$ then, as we shall see shortly (Theorem 16.5.1), $f = (x - \alpha)g(x)$ for some polynomial $g(x) \in \mathbb{F}[x]$, and degree considerations dictate that $g(x)$ is a linear polynomial.

Therefore, for quadratic polynomials one can say that f is reducible if and only if f has a root in $\mathbb{F}$. If, furthermore, $2 \neq 0$ in the field $\mathbb{F}$ then one can prove that f has a root if and only if $b^2 - 4ac$ is a square in $\mathbb{F}$. In fact, in that case, the unique factorization of f is

$$ax^2 + bx + c = a\left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a}\right).$$

(3) If f has degree 3 it is still true that f is reducible if and only if f has a root. But if f has degree 4 or higher this may fail. For example, the polynomial $x^2 - 2$ is irreducible over $\mathbb{Q}$ because $\sqrt{2}$ is irrational. Same for $x^2 - 3$. Thus, for example, the polynomials $(x^2 - 2)^2, (x^2 - 2)(x^2 - 3), (x^2 - 3)^2$ are reducible over $\mathbb{Q}$ but don't have a root. (Indeed, if α is a root of $(x^2 - 2)(x^2 - 3)$ then in $\mathbb{C}$ we have $(\alpha - \sqrt{2})(\alpha + \sqrt{2})(\alpha - \sqrt{3})(\alpha + \sqrt{3}) = 0$ and so α is $\pm\sqrt{2}$ or $\pm\sqrt{3}$ and, in any case, is not rational.)

(4) The property of f being irreducible depends on the field. It is not an absolute property. For example, $x^2 - 2$ is irreducible in $\mathbb{Q}[x]$ but is reducible in $\mathbb{C}[x]$ because there we can write $x^2 = (x - \sqrt{2})(x + \sqrt{2})$.

Theorem 16.4.4. *(Unique factorization for polynomials) Let $f(x) \in \mathbb{F}[x]$ be a non-zero polynomial. Then there is an $a \in \mathbb{F}^\times$ and distinct monic irreducible polynomials $f_1, \cdots, f_g$ and positive integers $r_1, \ldots, r_g$ such that*

$$f = af_1^{r_1} \cdots f_g^{r_g}. \tag{3}$$

Moreover, if

$$f = bh_1^{s_1} \cdots h_t^{s_t},$$

where $b \in \mathbb{F}^\times$, h_i distinct monic irreducible polynomials and $s_i > 0$, then $a = b$, $g = t$, and after re-naming the h_i's we have $h_i = f_i$ for all i, and $r_i = s_i$ for all i.

Proof. The proof is very similar to the proof for integers. We first prove the existence of factorization. Suppose that there is a non-zero polynomial $f(x)$ with no such factorization. Choose then a non-zero polynomial $f(x)$ of minimal degree for which no such factorization exists. Then $f(x)$ is not a constant polynomial and is not an irreducible polynomial either, else $f(x) = a_n x^n + \cdots + a_0 = a_n \cdot (a_n^{-1} f(x))$ is a suitable factorization. It follows that $f(x) = f_1(x) f_2(x)$, where each $f_i(x)$ has degree less than that of $f(x)$.

Therefore, each $f_i(x)$ has a factorization

$$f_1(x) = c_1 a_1(x) \cdots a_m(x), \qquad f_2(x) = c_2 b_1(x) \cdots b_n(x),$$

with $c_i \in \mathbb{F}$ and a_i, b_j monic irreducible polynomials, not necessarily distinct. It follows that

$$f(x) = (c_1 c_2) a_1(x) \cdots a_m(x) b_1(x) \cdots b_n(x),$$

has also a factorization as claimed, by collecting factors. Contradiction. Thus, no such $f(x)$ exists, and every polynomial has a factorization as claimed.

We now show the uniqueness of the factorization. Suppose that

$$f(x) = c_1 a_1(x) \cdots a_m(x) = c_2 b_1(x) \cdots b_n(x),$$

with $c_i \in \mathbb{F}$ and a_i, b_j monic irreducible polynomials, not necessarily distinct (the polynomials are not related to those appearing in the previous part of the proof). We prove the result by induction on degree f. Since c_i is the leading coefficient of f, we have $c_1 = c_2$. In particular, the case of $\deg(f) = 0$ holds. Assume that we proved uniqueness for all polynomials of degree $\leq d$ and $\deg(f) = d + 1 \geq 1$. Since $a_1(x) | c_2 b_1(x) \cdots b_n(x)$ and $a_1(x)$ is irreducible, it follows that either $a_1(x) | c_2$ (which is impossible because c_2 is a constant) or $a_1(x) | b_i(x)$ for some i. But since $b_i(x)$ is irreducible it then follows that that $a_1(x) \sim b_i(x)$ and so, both polynomials being monic, $a_1(x) = b_i(x)$.

Let us re-number the b_i so that $a_1 = b_1$. Then, dividing by $a_1(x)$ we have

$$c_1 a_2(x) \cdots a_m(x) = c_2 b_2(x) \cdots b_n(x).$$

Induction gives that $m = n$ and, after re-numbering the b_i, $a_i(x) = b_i(x), i = 2, 3, \ldots, n$. □

Example 16.4.5. Here are some examples.

(1) $f(x) = ax + b, a \neq 0$ has unique factorization $a(x + a^{-1}b)$.

(2) $f(x) = ax^2 + bx + c$ is irreducible if and only if it has no root in $\mathbb{F}$, as we have seen above. If this is the case,

$$f(x) = a(x^2 + a^{-1}bx + a^{-1}c)$$

is the unique factorization. Otherwise, f has two roots, say α and β (and if $2 \neq 0$ in $\mathbb{F}$ we have a formula for them) and

$$f(x) = a(x - \alpha)(x - \beta)$$

is the unique factorization.

(3) Consider the polynomial $f(x) = (x^2 - 2)(x^2 - 3)$ over the field $K = \mathbb{Q}(\sqrt{2})$. We claim that $x^2 - 3$ is irreducible over K. Indeed, if not, $\sqrt{3} \in K$ and so $\sqrt{3} = a + b\sqrt{2}$ for some rational numbers a, b. Squaring, we get

$$3 = a^2 + 2b^2 + 2ab\sqrt{2}.$$

But this implies that $2ab\sqrt{2}$ is rational and so $ab = 0$. If $b = 0$ we get that $\sqrt{3} = a$ is rational, which is a contradiction. If $a = 0$ we get that $3 = 2b^2$, which is a contradiction, because of unique factorization of rational numbers: The power of 2 in the left hand side (i.e., in 3) is 0, while in the right hand side (i.e. in $2a^2$) is odd, whatever it may be. Therefore,

$$f(x) = (x - \sqrt{2})(x + \sqrt{2})(x^2 - 3)$$

is the unique factorization of f over $\mathbb{Q}(\sqrt{2})$.

(4) What is the unique factorization of x^4+x over $\mathbb{F}_2[x]$? An obvious factor is x and then we are left with x^3+1. We note that $x=1$ is a root ($2=0$ now!) and so $x^3+1=(x+1)(x^2+x+1)$ (cf. Theorem 16.5.1); this is correct as $-1=1$ in $\mathbb{F}_2$. The polynomial x^2+x+1 is quadratic and so is reducible over $\mathbb{F}_2$ if and only if it has a root in $\mathbb{F}_2$. We check by calculation that neither $x=0$ nor $x=1$ are roots. We conclude that the unique factorization is given in this case by

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

(5) Over the complex numbers any non-constant polynomial f factors as

$$f(x)=a_n\prod_{i=1}^{n}(x-z_i),$$

(see Proposition 6.3.3) and this is precisely its unique factorization, except if some of the z_i are equal and then we may wish to collect them together to get a factorization as in (3).

We can now deduce from Theorem 16.4.4 the analogues of Proposition 10.2.1 and Corollary 10.2.2. The proofs are the same.

Proposition 16.4.6. *Let f,g be non-zero polynomials in $\mathbb{F}[x]$. Then $f|g$ if and only if $f=ap_1^{a_1}\cdots p_m^{a_m}$ and $g=bp_1^{a_1'}\cdots p_m^{a_m'}q_1^{b_1}\cdots q_t^{b_t}$ (products of distinct irreducible monic polynomials p_i; a,b non-zero scalars) with $a_i'\geq a_i$ for all $i=1,\ldots,m$.*

Corollary 16.4.7. *Let $f=ap_1^{a_1}\cdots p_m^{a_m}, g=bp_1^{b_1}\cdots p_m^{b_m}$ with p_i distinct irreducible monic polynomials, a,b non zero scalars and a_i,b_i non-negative integers. (Any two non-zero polynomials can be written this way). Then*

$$\gcd(f,g)=p_1^{\min(a_1,b_1)}\cdots p_m^{\min(a_m,b_m)}.$$

16.5. **Roots.** Let $\mathbb{F}$ be a field and let $f(x)\in\mathbb{F}[x]$ be a non-zero polynomial. Recall that an element $a\in\mathbb{F}$ is called a **root** (or **zero**, or **solution**) of f if $f(a)=0$.

Theorem 16.5.1. *Let $f(x)\in\mathbb{F}[x]$ be a non-zero polynomial.*

(1) If $f(a)=0$ then $f(x)=(x-a)g(x)$ for a unique polynomial $g(x)\in\mathbb{F}[x]$. In particular, if f is irreducible of degree greater than 1 then f has no roots in $\mathbb{F}$.
(2) Let $\deg(f)=d$ then f has at most d roots.

Proof. Suppose that $f(a)=0$ and divide f by $x-a$ with a residue, getting

$$f(x)=g(x)(x-a)+r(x),$$

where $r(x)$ is either zero or a polynomial of degree less than that of $x-a$. That is, in either case, $r(x)$ is a constant. Substitute $x=a$. We get $0=f(a)=g(a)(a-a)+r=r$ and so $f(x)=(x-a)g(x)$.

Consider the factorization of f into irreducible monic polynomials:

$$f=A(x-a_1)^{s_1}\cdots(x-a_m)^{s_m}f_1(x)^{r_1}\ldots f_n(x)^{r_n},$$

where the f_i are irreducible polynomials of degree larger than 1, the r_i,s_i are positive and $A\in\mathbb{F}^\times$. Note that if $f(a)=0$ then, since $f_i(a)\neq 0$ (else $f_i(x)=(x-a)g_i(x)$, for some polynomial g_i, hence reducible), we must have $a=a_i$ for some i. It follows that the number of roots of f, counting multiplicities, is $s_1+s_2+\cdots+s_m=\deg((x-a_1)^{s_1}\cdots(x-a_m)^{s_m})\leq\deg(f)=d$. □

A field $\mathbb{F}$ is called **algebraically closed** if any non-constant polynomial $f(x)\in\mathbb{F}[x]$ has a root in $\mathbb{F}$. Recall the following important result.

Theorem 16.5.2 (The Fundamental Theorem of Algebra)**.** *The field of complex numbers is algebraically closed.*

It is a fact (proven in Algebra III) that every field is contained in an algebraically closed field. If $\mathbb{F}$ is algebraically closed, then the only irreducible polynomials over $\mathbb{F}$ are the linear polynomials, viz. $x-a, a\in\mathbb{F}$. It follows then that

$$f(x)=A(x-a_1)^{s_1}\cdots(x-a_m)^{s_m},$$

where A is the leading coefficient of f and $a_1,\dots,a_m$ are the roots (with multiplicities $s_1,\dots,s_m$).

A natural question is, for a given field $\mathbb{F}$ and a given polynomial $f(x)$, to tell if f has a root in $\mathbb{F}$ or not. It is the first check one can do when trying to determine whether $f(x)$ is irreducible or not. Unfortunately, in general this is impossible to decide; but we have some partial answers in special cases.

Proposition 16.5.3. *Let $f(x)=a_nx^n+\cdots+a_1x+a_0$ be a non-constant polynomial with integer coefficients. If $a=s/t$, $(s,t)=1$, is a rational root of f then $s|a_0$ and $t|a_n$.*

Proof. We have $a_n(s/t)^n+\cdots+a_1(s/t)+a_0=0$ and so

$$a_ns^n+a_{n-1}s^{n-1}t+\cdots+a_1st^{n-1}+a_0t^n=0.$$

Since s divides $a_ns^n+a_{n-1}s^{n-1}t+\cdots+a_1st^{n-1}$, it follows that $s|a_0t^n$. Then, since $(s,t)=1$, we get that $s|a_0$. Similarly, t divides $a_{n-1}s^{n-1}t+\cdots+a_1st^{n-1}+a_0t^n$, so t divides a_ns^n. Now $(s,t)=1$ implies that $t|a_n$. □

Example 16.5.4. *Problem: Find the rational roots of the polynomial $x^4-\frac{7}{2}x^3+\frac{5}{2}x^2-\frac{7}{2}x+\frac{3}{2}$.*
The roots are the same as for the polynomial $2x^4-7x^3+5x^2-7x+3$. They are thus of the form s/t, where $s=\pm1,\pm3, t=\pm1,\pm2$. We have the possibilities $\pm1,\pm1/2,\pm3,\pm3/2$. By checking each case, we find the roots are $1/2$ and 3. We remark that after having found the root $1/2$ we can divide the polynomial $2x^4-7x^3+5x^2-7x+3$ by $x-1/2$ finding $2x^3-6x^2+2x-6$, whose roots are the roots of x^3-3x^2+x-3. So, in fact, the only possibilities for additional roots are ±3. We saved this way the need to check if $\pm3/2$ are roots.

Here is another example. *Is the polynomial x^3+2x^2+5 irreducible over $\mathbb{Q}$?*
In this case, if it is reducible then one of the factors would have to have degree 1 (this type of argument only works for degrees $1,2,3$ polynomials. For higher degree, we might have a reducible polynomial with no linear factor, e.g., $(x^2+1)(x^2+3)$). Namely, the polynomial would have a rational root. But the rational roots can only be $\pm1,\pm5$ and one verifies those are not roots. Thus, the polynomial is irreducible.

Proposition 16.5.5. *If $f(x)\in\mathbb{R}[x]$ is a polynomial of odd degree then f has a root in $\mathbb{R}$.*

Proof. Since the roots of f are the roots of $-f$, we may assume that $f(x)=a_nx^n+\cdots+a_1x+a_0$, $a_i\in\mathbb{R}$, $a_n>0$. An easy estimate shows that there is an $N>0$ such that $f(N)>0$ and $f(-N)<0$. By the intermediate value theorem there is some a, $-N\le a\le N$ such that $f(a)=0$. (Another proof appears in the exercises.) □

16.6. **Eisenstein's criterion.**

Theorem 16.6.1. *Let $f(x)=x^n+\cdots+a_1x+a_0\in\mathbb{Z}[x]$ be a monic polynomial with integer coefficients. Suppose that for some prime p we have $p|a_i,\forall i$ and $p^2\nmid a_0$ then $f(x)$ is irreducible over $\mathbb{Q}$.*

Proof. We first prove that $f(x)$ is irreducible over $\mathbb{Z}$. Namely, suppose $f(x)=g(x)h(x)$, where $g(x),h(x)\in\mathbb{Z}[x]$ and both polynomials are not constant. Let us write $g(x)=c_ax^a+\cdots+c_1x+c_0, h(x)=d_bx^b+\cdots+d_1x+d_0$. Note that c_a,d_b are ±1. Reduce the identity $f(x)=g(x)h(x)$ modulo p to get $x^{a+b}=\overline{g(x)}\cdot\overline{h(x)}$. By unique factorization for $\mathbb{Z}/p\mathbb{Z}[x]$ we conclude that $\overline{g(x)}=\overline{c_a}x^a, \overline{h(x)}=\overline{d_b}x^b$, and in particular, that $p|c_0, p|d_0$. It follows that $p^2|c_0d_0=a_0$ and that's a contradiction.

We next prove that if $f(x)$ is reducible over $\mathbb{Q}$ it is reducible over $\mathbb{Z}$. Suppose that $f(x)=g(x)h(x)$, where $g(x),h(x)$ are in $\mathbb{Q}[x]$ are non-constant polynomials. Multiply by a suitable integer to get that $F(x)=G(x)H(x)$, where $F(x)=Nf(x)$ is in $\mathbb{Z}[x]$ and all its coefficients are divisible by N, and $G(x),H(x)\in\mathbb{Z}[x]$ are non-constant polynomials. It is enough to prove that if a prime p divides all the coefficients of $F(x)$, it

either divides all the coefficients of $G(x)$ or all the coefficients of $H(x)$, because then, peeling off one prime at the time, we find a factorization of $f(x)$ into polynomials with integer coefficients.

Reduce the equation $F(x) = G(x)H(x)$ modulo p, for a prime p that divides all the coefficients of $F(x)$, to find $0 = \overline{G(x)} \cdot \overline{H(x)}$. Using that $\mathbb{Z}/p\mathbb{Z}[x]$ is an integral domain, we conclude that either $\overline{G(x)} = 0$ or $\overline{H(x)} = 0$, which means that either p divides all the coefficients of G, or all the coefficients of H. □

16.7. **Roots of polynomials in $\mathbb{Z}/p\mathbb{Z}$.** Let p be a prime and let $\mathbb{Z}/p\mathbb{Z}$ be the field with p elements whose elements are congruence classes modulo p. By Fermat's little theorem, every element of $\mathbb{Z}/p\mathbb{Z}^\times$ is a root of $x^{p-1} - 1$. This gives $p-1$ distinct roots of $x^{p-1} - 1$ and so these must be all the roots and each must appear with multiplicity one. It follows that the roots of $x^p - x$ are precisely the elements of $\mathbb{Z}/p\mathbb{Z}$, again each with multiplicity one. That is,

$$x^p - x = \prod_{a=0}^{p-1} (x - \bar{a}).$$

Proposition 16.7.1. *Let $f(x)$ be any polynomial in $\mathbb{Z}/p\mathbb{Z}[x]$. Then $f(x)$ has a root in $\mathbb{Z}/p\mathbb{Z}$ if and only if $\gcd(f(x), x^p - x) \neq 1$.*

Proof. If $f(a) = 0$ for some $a \in \mathbb{Z}/p\mathbb{Z}$ then $(x-a)|f(x)$, but also $(x-a)|(x^p - x)$. It follows that $\gcd(f(x), x^p - x) \neq 1$. Conversely, if $h(x) = \gcd(f(x), x^p - x) \neq 1$ then, since $h(x)|x^p - x = \prod_{a=0}^{p-1}(x - \bar{a})$, by unique factorization we must have $h(x) = \prod_{i=1,\ldots,n}(x - a_i)$ for some distinct elements $a_1, \ldots, a_n$ of $\mathbb{Z}/p\mathbb{Z}$. In particular, each such a_i is a root of $f(x)$. □

The straightforward way to check if $f(x)$ has a root in $\mathbb{Z}/p\mathbb{Z}$ is just to try all possibilities for x. Suppose that $f(x)$ has a small degree relative to p. Even then, except in special cases, we still have to try p residue classes, each in its turn, to see if any of which is a root. But p may be very large, much too large for this method to be feasible. For example, p might be of cryptographic size $\approx 2^{2048}$ (according to the RSA company, this should be secure until 2030). Even with a computer doing 10^{10} operations per second, which is about what a good laptop does these days (2016), checking all these possibilities will take about 10^{600} years!

Proposition 16.7.1 suggests a different method: Calculate $\gcd(f(x), x^p - x)$. Note that except for the first step

$$x^p - x = q_0(x)f(x) + r_0(x),$$

all the polynomials involved in the Euclidean algorithm would have very small degrees (smaller than f's for example) and so the Euclidean algorithm will terminate very quickly. The first step, though, could be very time consuming given what we know at this point. Later we shall see that it can, in fact, be done quickly (in order of magnitude $\log(p)$). For example, using the software GP/PARI, it took my laptop 69 microseconds to determine that for $p = 2^{2203} - 1$, the polynomial $f(x) = x^3 + x + 1$ is irreducible. This p is an example of a Mersenne prime; for numbers of the form $2^n - 1$ we have special methods to ascertain their primality. Even a prime of the form $p = 2^{9941} - 1$ was no problem; it took 3.582 seconds to determine that $f(x)$ is reducible modulo p. It even took only 7.710 seconds to find the root. As the root has close to 3000 digits, I am not listing it here.

We have seen that many of the features of arithmetic in $\mathbb{Z}$ can be carried out in $\mathbb{F}[x]$. We still don't have an analogue of passing from $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ in the context of $\mathbb{F}[x]$. This is one motivation for studying rings in much more detail; we'd like to be able to emulate the process of $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ for general rings, not just $\mathbb{F}[x]$.

17. Exercises

(1) In each case, divide $f(x)$ by $g(x)$ with residue:
   (a) $f(x) = 3x^4 - 2x^3 + 6x^2 - x + 2$, $g(x) = x^2 + x + 1$ in $\mathbb{Q}[x]$.
   (b) $f(x) = x^4 - 7x + 1$, $g(x) = 2x^2 + 1$ in $\mathbb{Q}[x]$.
   (c) $f(x) = 2x^4 + x^2 - x + 1$, $g(x) = 2x - 1$ in $\mathbb{Z}/5\mathbb{Z}[x]$.
   (d) $f(x) = 4x^4 + 2x^3 + 6x^2 + 4x + 5$, $g(x) = 3x^2 + 2$ in $\mathbb{Z}/7\mathbb{Z}[x]$.

(2) Use the Euclidean algorithm to find the gcd of the following pairs of polynomials and express it as a combination of the two polynomials.
   (a) $x^4 - x^3 - x^2 + 1$ and $x^3 - 1$ in $\mathbb{Q}[x]$.
   (b) $x^5 + x^4 + 2x^3 - x^2 - x - 2$ and $x^4 + 2x^3 + 5x^2 + 4x + 4$ in $\mathbb{Q}[x]$.
   (c) $x^4 + 3x^3 + 2x + 4$ and $x^2 - 1$ in $\mathbb{Z}/5\mathbb{Z}[x]$.
   (d) $4x^4 + 2x^3 + 3x^2 + 4x + 5$ and $3x^3 + 5x^2 + 6x$ in $\mathbb{Z}/7\mathbb{Z}[x]$.
   (e) $x^3 - ix^2 + 4x - 4i$ and $x^2 + 1$ in $\mathbb{C}[x]$.
   (f) $x^4 + x + 1$ and $x^2 + x + 1$ in $\mathbb{Z}/2\mathbb{Z}[x]$.

(3) Consider the polynomial $x^2 + x = 0$ over $\mathbb{Z}/n\mathbb{Z}$.
   (a) Find an n such that the equation has at least 4 solutions.
   (b) Find an n such that the equation has at least 8 solutions.

(4) Is the given polynomial irreducible:
   (a) $x^2 - 3$ in $\mathbb{Q}[x]$? In $\mathbb{R}[x]$?
   (b) $x^2 + x - 2$ in $\mathbb{F}_3[x]$? In $\mathbb{F}_7[x]$? (For any prime p we denote $\mathbb{Z}/p\mathbb{Z}$ also by $\mathbb{F}_p$. This notation is used only for primes! Namely, one does *not* use a notation as $\mathbb{F}_n$ if n is not a prime.)

(5) Find the rational roots of the polynomial $2x^4 + 4x^3 - 5x^2 - 5x + 2$.

(6) For a polynomial $g(x) = a_n x^n + \cdots + a_1 x + a_0$ with complex coefficients, let $\overline{g(x)} = \overline{a_n} x^n + \cdots + \overline{a_1} x + \overline{a_0}$ be the polynomial obtained by taking the complex conjugate of the coefficients. Check that $\overline{g_1(x)g_2(x)} = \overline{g_1(x)} \cdot \overline{g_2(x)}$. Let $f(x)$ be a polynomial with real coefficients and
$$f(x) = a(x - \alpha_1)^{a_1}(x - \alpha_2)^{a_2} \cdots (x - \alpha_r)^{a_r},$$
its unique factorization over $\mathbb{C}$. Apply complex conjugation to both sides. Deduce that if α is a root of f with multiplicity d then $\bar{\alpha}$ is a root of f with the same multiplicity. Deduce that if f has odd degree then f has a real root.

(7) Let $p > 2$ be a prime. Calculate the gcd of $x^{p-1} - 1$ and $x^2 + 1$ in the ring $\mathbb{Z}/p\mathbb{Z}[x]$, using the Euclidean algorithm method, and conclude that -1 is a square in $\mathbb{Z}/p\mathbb{Z}$ if and only $p \equiv 1 \pmod 4$.

(8) Let p be a prime number. Use the factorization of $x^p - x$ to deduce Wilson's theorem: $(p-1)! \equiv -1 \pmod p$

Part 5. **Rings**

18. Some basic definitions and examples

Recall our definition of a ring.

Definition 18.0.1. A **ring** R is a non-empty set together with two operations, called "addition" and "multiplication" that are denoted, respectively, by

$$(x,y) \mapsto x+y, \qquad (x,y) \mapsto xy.$$

One requires the following axioms to hold:

(1) $x+y = y+x$, $\forall x,y \in R$. (Commutativity of addition)
(2) $(x+y)+z = x+(y+z)$, $\forall x,y,z \in R$. (Associativity of addition)
(3) There exists an element in R, denoted 0, such that $0+x = x$, $\forall x \in R$. (Neutral element for addition)
(4) $\forall x \in R$, $\exists y \in R$ such that $x+y = 0$. (Inverse with respect to addition)
(5) $(xy)z = x(yz)$, $\forall x,y,z \in R$. (Associativity of multiplication)
(6) There exists an element $1 \in R$ such that $1x = x1 = x$, $\forall x \in R$. (Neutral element for multiplication)
(7) $z(x+y) = zx+zy, (x+y)z = xz+yz$, $\forall x,y,z \in R$. (Distributivity)

Recall also that a ring R is called a **division ring** (or sometimes a **skew-field**) if $1 \neq 0$ in R and any non-zero element of R has an inverse with respect to multiplication. A commutative division ring is precisely what we call a field.

Example 18.0.2. The ring of integers $\mathbb{Z}$ is a commutative ring. It is not a division ring and so is not a field. The rational numbers $\mathbb{Q}$ form a field. The real numbers $\mathbb{R}$ form a field. The complex numbers $\mathbb{C}$ form a field.

We have also noted some useful formal consequences of the axioms defining a ring:

(1) The element 0 appearing in axiom (3) is unique.
(2) Given x, the element y appearing in axiom (4) is unique. We shall denote y by $-x$.
(3) We have $-(-x) = x$ and $-(x+x') = -x-x'$, where, technically $-x-x'$ means $(-x)+(-x')$.
(4) We have $x \cdot 0 = 0, 0 \cdot x = 0$.

Here are some further examples. We do not prove that the ring axioms hold; this is left as an exercise.

Example 18.0.3. Let $\mathbb{F}$ be a field and $n \geq 1$ an integer. Consider the set of $n \times n$ matrices:

$$M_n(\mathbb{F}) = \left\{ \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} : a_{ij} \in \mathbb{F} \right\}.$$

For example:

(1) for $n = 1$ we just get $(a_{11}), a_{11} \in \mathbb{F}$;
(2) for $n = 2$, $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$;
(3) for $n = 3$ we get $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$.

In general we shall write an $n \times n$ matrix as (a_{ij}), or $(a_{ij})_{i,j=1}^n$ if we need to be very clear about the dimensions of the matrix. The index i is the row index and the index j is the column index. We then define

$$(a_{ij}) + (b_{ij}) = (a_{ij}+b_{ij}), \qquad (a_{ij})(b_{ij}) = (c_{ij}),$$

where

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

We can say that the ij entry of the product AB, is the dot product of the i-th row of A with the j-th column of B.

$$\begin{pmatrix} \cdots & & \cdots \\ & c_{ij} & \\ \cdots & & \cdots \end{pmatrix} = \begin{pmatrix} \vdots & & \vdots \\ a_{i1} & \cdots & a_{in} \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} \cdots & b_{1j} & \cdots \\ & \vdots & \\ \cdots & b_{nj} & \cdots \end{pmatrix}$$

For example:

(1) for $n = 1$ we get $(a) + (b) = (a + b)$ and $(a)(b) = (ab)$. Namely, we just get $\mathbb{F}$ again!
(2) for $n = 2$, we have

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix},$$

and

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}.$$

Under these definitions $M_n(\mathbb{F})$ is a ring, called the **ring of $n \times n$ matrices with entries in $\mathbb{F}$**, with identity given by the identity matrix

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \\ & & \ddots & \\ 0 & \dots & & 1 \end{pmatrix},$$

and zero given by the zero matrix (the matrix all whose entries are zero). For $n \geq 2$ this is a non-commutative ring. For example, for $n = 2$ we have,

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

These are never equal, else $2 = 1$ in $\mathbb{F}$, which implies $1 = 0$ in $\mathbb{F}$, which is never the case, by definition.

Example 18.0.4. Let ϵ be a formal symbol and $\mathbb{F}$ a field. The **ring of dual numbers**, $\mathbb{F}[\epsilon]$, is defined as

$$\mathbb{F}[\epsilon] = \{a + b\epsilon : a, b \in \mathbb{F}\},$$

with the following addition and multiplication:

$$(a + b\epsilon) + (c + d\epsilon) = a + c + (b + d)\epsilon, \quad (a + b\epsilon)(c + d\epsilon) = ac + (ad + bc)\epsilon.$$

Note that ϵ is a zero divisor: $\epsilon \neq 0$ but $\epsilon^2 = 0$.

Example 18.0.5. Let R_1, R_2 be rings. Then $R_1 \times R_2$ is a ring with the following operations:

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2), \quad (a_1, b_1)(a_2, b_2) = (a_1 a_2, b_1 b_2).$$

The zero element is $(0_{R_1}, 0_{R_2})$ and the identity element is $(1_{R_1}, 1_{R_2})$. The ring $R_1 \times R_2$ is called the **direct product** of R_1 and R_2.

Some of the examples of rings are rather complicated (for example, rings of matrices) or exotic looking (like the ring of dual numbers) and one would be justified to ask why one would want to consider such rings. As one progresses in mathematics, the need for the concept of ring becomes more and more clear. In linear algebra we learn that linear transformations of $\mathbb{R}^3$, such as rotations fixing the origin, can be described by 3×3 matrices with real entries and composition of linear transformations matches multiplication of matrices in $M_3(\mathbb{R})$. Modern algebraic geometry associated to every commutative ring R a space $\mathrm{Spec}(R)$ (it is a set, there is a notion of open subsets and of functions, and so on) whose points are the prime ideals of R (see

below for the concept of a prime ideal) and, in this setting, the ring of dual numbers is used to study the tangent space at a point of $\mathrm{Spec}(R)$. If $\mathbb{F}$ is a field, we understand the importance of the ring of polynomials $\mathbb{F}[x]$ and similarly $\mathbb{F}[x,y]$, but if we want to considers polynomials in x "modulo a fixed polynomial $f(x)$", or polynomials in x and y "modulo a collection of polynomials $f_1(x,y),\dots,f_n(x,y)$" etc., we get new rings and ring theory is the best way to have a rigorous language to discuss those. Thus, with no more apologies, we proceed to develop the very basics of this huge area of mathematics.

Definition 18.0.6. Let $S \subset R$ be a subset. S is called a **subring** of R if the following holds:
(1) $0_R, 1_R$ belong to S;
(2) $s_1, s_2 \in S \Rightarrow s_1 \pm s_2 \in S$;
(3) $s_1, s_2 \in S \Rightarrow s_1 s_2 \in S$.

Note that in this case S is a ring in its own right.

Example 18.0.7. The easiest examples are $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ being subrings of $\mathbb{C}$. We've already seen examples of subrings of the ring of 2×2 matrices in Exercise 26, page 31.

Consider the subset $\{(r,0) : r \in \mathbb{R}\}$ of the ring $\mathbb{R}\times\mathbb{R}$. It is closed under addition and multiplication. It is even a ring because $(r,0)(1,0) = (r,0)$ and so $(1,0)$ serves as an identity element for this subset. Nonetheless, *it is not a subring of* $\mathbb{R}\times\mathbb{R}$ *according to our definition, because the identity element of* $\mathbb{R}\times\mathbb{R}$, *which is* $(1,1)$, *does not belong to the set* $\{(r,0) : r \in \mathbb{R}\}$. Some authors have other conventions (and thus would consider $\mathbb{R}\times\{0\}$ as a subring of $\mathbb{R}\times\mathbb{R}$), but by-and-large our conventions are the more widespread. One needs to be careful when considering other textbooks, though.

Example 18.0.8. Let $n \neq \pm 1$ be a square free integer (meaning, if $p|n$, p prime, then $p^2 \nmid n$). Then $\sqrt{n}$ is not a rational number. Indeed if $\sqrt{n}$ is rational, $\sqrt{n} = s/t$, $(s,t) = 1$, then $n = s^2/t^2$. Let p a prime dividing n and so $p^2 \nmid n$. Since $nt^2 = s^2$, $p|s^2$. But then $p|s$. Looking at the power of p in the unique factorization of both sides, it follows that $p|t$ and thus $p|(s,t)$ – a contradiction.

Consider

$$\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a, b \in \mathbb{Z}\}.$$

This is a subset of $\mathbb{C}$, containing 0 and 1 and is closed under addition and multiplication:

$$(a + b\sqrt{n}) + (c + d\sqrt{n}) = (a + c) + (b + d)\sqrt{n},$$

and

$$(a + b\sqrt{n})(c + d\sqrt{n}) = (ac + bdn) + (ad + bc)\sqrt{n}.$$

We remark that any element of this ring has a unique expression as $a + b\sqrt{n}$. Indeed, if $a + b\sqrt{n} = c + d\sqrt{n}$, either $b = d$ (and then obviously $a = c$) or $\sqrt{n} = (a - c)/(d - b)$ is a rational number, which it's not.

19. Ideals

Definition 19.0.1. Let R be a ring. A (two-sided) **ideal** I of R is a subset of R such that
(1) $0 \in I$;
(2) if $a, b \in I$ then $a + b \in I$;
(3) if $a \in I$, $r \in R$, then $ra \in I$ and $ar \in I$.

Remark 19.0.2. Note that if $a \in I$ then $-1 \cdot a = -a \in I$.

We shall use the notation $I \lhd R$ to indicate that I is an ideal of R.

Example 19.0.3. $I = \{0\}$ and $I = R$ are always ideals. They are called the **trivial ideals**.

Example 19.0.4. Suppose that R is a division ring (e.g., a field) and $I \lhd R$ is a non-zero ideal. Then $I = R$. Indeed, there is an element $a \in I$ such that $a \neq 0$. Then $1 = a^{-1}a \in I$ and so for every $r \in R$ we have $r = r \cdot 1 \in I$. That is, $I = R$. We conclude that a division ring has only the trivial ideals. (Note also that the argument shows for any ring R that if an ideal I contains an invertible element of R then $I = R$.)

Example 19.0.5. Let R be a commutative ring. Let $r \in R$. The **principal ideal** (r) is defined as

$$(r) = \{ra : a \in R\} = \{ar : a \in R\}.$$

We also denote this ideal by rR or Rr. This is indeed an ideal: First $0 = r \cdot 0$ is in (r). Second, given two elements ra_1, ra_2 in (r) we have $ra_1 + ra_2 = r(a_1 + a_2) \in (r)$ and for every $s \in R$ we have $s(ra_1) = (sr)a_1 = (rs)a_1 = r(sa_1) \in rR$ (using commutativity!), $(ra_1)s = r(a_1 s) \in rR$.

Definition 19.0.6. Let R be a commutative ring. If every ideal of R is principal, one calls R a **principal ideal ring.**

Example 19.0.7. Let $\mathbb{F}$ be a field. It is a principal ideal ring. The only ideals are $\{0\} = (0)$ and $\mathbb{F} = (1)$. One can also show that all the ideals of $\mathbb{F}[\epsilon]$ are $\{0\} = (0), \mathbb{F}[\epsilon] = (1)$ and $(\epsilon) = \{b\epsilon : b \in \mathbb{F}\}$ and so the ring of dual numbers is also a principal ideal ring.

Example 19.0.8. The ring of polynomials $\mathbb{C}[x, y]$ in two variables with complex coefficients is not a principal ideal ring. We claim that the set of polynomials $I = \{f(x, y) : f(0, 0) = 0\}$, namely, polynomials with zero constant term, is an ideal that is not principal. We leave that as an exercise.

Theorem 19.0.9. *$\mathbb{Z}$ is a principal ideal ring. In fact, the list*

$$(0), (1), (2), (3), (4), \ldots$$

is a complete list of the ideals of $\mathbb{Z}$. (Note that another notation is $0, 1\mathbb{Z}, 2\mathbb{Z}, 3\mathbb{Z}, 4\mathbb{Z}, \ldots$.)

Proof. We already know that these are ideals, in fact principal ideals, and we note that for $i > 0$ the minimal positive number in the ideal (i) is i. Thus, these ideals are distinct.

Let I be an ideal of $\mathbb{Z}$. If $I = \{0\}$ then I appears in the list above. Else, there is some non-zero element $a \in I$. If $a < 0$ then $-a = -1 \cdot a \in I$ and so I has a positive element in it. Choose the smallest positive element in I and call it i.

First, since $i \in I$ so is ia for any $a \in \mathbb{Z}$ and so $(i) \subset I$. Let $b \in I$. Divide b by i with residue: $b = qi + r$, where $0 \leq r < i$. Note that $r = b - qi$ is an element of I, smaller than i and non-negative. The only possibility is that $r = 0$ and so $b = qi \in (i)$. Thus, $I = (i)$. □

Theorem 19.0.10. *Let $\mathbb{F}$ be a field. The ring $\mathbb{F}[x]$ is a principal ideal ring. Two ideals $(f(x)), (g(x))$ are equal if and only if $f \sim g$.*

Proof. The proof is very similar to the case of $\mathbb{Z}$. Let I be an ideal. If $I = \{0\}$ then $I = (0)$, the principal ideal generated by 0. Else, let $f(x) \in I$ be a non-zero polynomial whose degree is minimal among all non-zero elements of I. On the one hand $I \supseteq (f(x))$. On the other hand, let $g(x) \in I$ and write $g(x) = q(x)f(x) + r(x)$, where $r(x)$ is either zero or of degree smaller than $f's$. But $r(x) = g(x) - q(x)f(x) \in I$. Thus, we must have $r(x) = 0$ and so $g(x) = q(x)f(x) \in (f(x))$. That is, $I \subseteq (f(x))$.
At this point we need a definition and a lemma.

Let R be any ring. The **units** of R are denoted $R^\times$ and defined as follows:

$$R^\times = \{x \in R : \exists y \in R, xy = yx = 1\}.$$

For example, 1_R is always a unit. If R is a field then, by definition, $R^\times = R - \{0\}$. Given $x \in R^\times$ the y such that $xy = yx = 1$ is unique and we denote it x^{-1}. Indeed, if also $xy_1 = 1$ then $y(xy_1) = y$ and on the other hand $y(xy_1) = (yx)y_1 = 1 \cdot y_1 = y_1$. Thus, $y = y_1$.

Lemma 19.0.11. *Let R be a commutative integral domain and $a, b \in R$. We say that $a \sim b$ (a and b are associates) if for some unit u of R, $au = b$. This is an equivalence relation. If $a|b$ and $b|a$ then $a \sim b$.*

Proof. (Lemma) As $a = 1 \cdot a$ and $au = b$ implies $a = u^{-1}b$ if u is a unit (else u^{-1} doesn't even make sense), this is a reflexive and symmetric relation. If $au = b$ and $bv = c$, where $u, v \in R^\times$ then $a(uv) = c$ and uv is a unit (the inverse is $v^{-1}u^{-1}$). This shows transitivity.

Now suppose $a|b$ and $b|a$. If $a = 0$ then $b = 0$ and conversely, and there's nothing to prove. Else, write $au = b$ for some $u \in R$ and $bv = a$ for some $v \in R$. We need to show that u, v are units. But, we have $a(1 - uv) = a - (au)v = a - bv = a - a = 0$. Since R is an integral domain and $a \neq 0$ it must be that $1 - uv = 0$. Thus $uv = 1$. Starting from $b(1 - vu) = 0$ we get in the same way that $vu = 1$. It follows that u, v are units of R. □

Note that for $R = \mathbb{F}[x]$, the notion of being associate is precisely the one we have previously defined. Indeed, the units of R are just the non-zero scalars and two polynomials are associate precisely if they differ by multiplication by a non-zero scalar. As R is also an integral domain, we may apply the Lemma. So, suppose that $(f(x)) \supset (g(x))$ then $g(x) = f(x)h(x)$ for some polynomial $h(x) \in \mathbb{F}[x]$. That is $f(x)|g(x)$. Thus, if $(f(x)) = (g(x))$ then $f|g$ and $g|f$ and so $f \sim g$.

If $f|g$, say $g(x) = f(x)h(x)$ then any multiple of $g(x)$, say $g(x)t(x)$ is equal to $f(x)[h(x)t(x)]$ and so $(g(x)) \subset (f(x))$. If $f \sim g$ then $f|g$ and $g|f$ and so, by the argument above, $(f(x)) = (g(x))$. □

Proposition 19.0.12. *Let R be a commutative ring and r, s associate elements of R then $(r) = (s)$.*

Proof. Let u be a unit such that $ur = s$ then $(s) = Rs = (Ru)r \subseteq Rr = (r)$. The same way, $(r) \subseteq (s)$. □

Example 19.0.13. Let R_1, R_2 be rings with ideals I_1, I_2, respectively. Then $I_1 \times I_2$ is an ideal of $R_1 \times R_2$.

Example 19.0.14. Let us consider the ring $\mathbb{F}[x]$ and in it the set

$$S = \{f(x) : f(x) = a_0 + a_2x^2 + a_3x^3 + \dots\},$$

of polynomials with no x term. Note that $0 \in S$ and $s_1, s_2 \in S \Rightarrow s_1 + s_2 \in S$ and even $s_1s_2 \in S$. However, S is not an ideal. We have $1 \in S$ but $x = x \cdot 1 \notin S$.

Example 19.0.15. Consider the ring $\mathbb{Z}[\sqrt{5}]$. In this ring we consider

$$I = \{5a + b\sqrt{5} : a, b \in \mathbb{Z}\}.$$

We claim that I is an ideal. This can be verified directly, but it is easier to note that I is in fact the principal ideal $(\sqrt{5})$.

Example 19.0.16. Let R be a ring and I_1, I_2 two ideals of R. Then

$$I_1 + I_2 = \{i_1 + i_2 : i_1 \in I_1, i_2 \in I_2\}$$

is an ideal of R. Inductively, the sum of n ideals $I_1 + I_2 + \dots + I_n$ is an ideal. A particular case is the following: Let R be a commutative ring and $I_i = r_iR$ a principal ideal. Then

$$r_1R + r_2R + \dots + r_nR = \{\sum_{i=1}^{n} r_ia_i : a_i \in R\}$$

is an ideal of R; we often denote it by $(r_1, r_2, \dots, r_n)$ or $\langle r_1, r_2, \dots, r_n\rangle$ (so in particular a principal ideal (r) may also be denoted $\langle r\rangle$).

Let us consider the situation of the ring $R = \mathbb{Z}[\sqrt{-5}]$ and the ideal $\langle 2, 1 + \sqrt{-5}\rangle$. We know abstractly that this is an ideal. We claim that this ideal is not principal. In particular, this shows that this ideal is not $\mathbb{Z}[\sqrt{-5}]$ and, more importantly, gives us an example of a ring with non-principal ideals.

Suppose that $\langle 2, 1 + \sqrt{-5}\rangle = \langle a + b\sqrt{-5}\rangle$. It follows that $2 = (a + b\sqrt{-5})(c + d\sqrt{-5})$ and so that $2 = (a - b\sqrt{-5})(c - d\sqrt{-5})$ (check!). Therefore, by multiplying these two equations, $4 = (a^2 + 5b^2)(c^2 + 5d^2)$. This is an equation in integers and so (because $0 < a^2 + 5b^2 \leq 4$) $a \in \{\pm 1, \pm 2\}, b = 0$ and we conclude that $\langle 2, 1 + \sqrt{-5}\rangle = \langle a\rangle$ is equal to $\langle 1\rangle$ or $\langle 2\rangle$. Now, if $\langle 2, 1 + \sqrt{-5}\rangle = \langle 2\rangle$ this implies that $1 + \sqrt{-5} = 2(c + d\sqrt{-5})$, which is a contradiction. If $\langle 2, 1 + \sqrt{-5}\rangle = \langle 1\rangle$ then $1 = 2(c_1 + d_1\sqrt{-5}) + (1 + \sqrt{-5})(c_2 + d_2\sqrt{-5}) = (2c_1 + c_2 - 5d_2) + \sqrt{-5}(2d_1 + c_2 + d_2)$. Therefore, $2d_1 + c_2 + d_2 = 0$, that is, $-2d_1 - c_2 = d_2$ and we get $1 = 2c_1 + c_2 - 5d_2 = 2c_1 + c_2 + 10d_1 + 5c_2 = 2(c_1 + 3c_2 + 5d_1)$. This is an equation in integers and it implies that 1 is even. Contradiction.

20. Homomorphisms

When we discussed sets, we also considered functions from one set to another. It was the functions, really, that made the subject much more deep and useful. Indeed, without functions, we wouldn't even be able to compare cardinalities of sets. Moreover, functions on sets are as natural as things get in mathematics. I can consider the set of guests of a wedding. A well-known headache-inducing problem is try and find a function from the set of guests to the set of tables that associates to each guest a table so as to maximize the number of friends around each table.

It is a general principle in mathematics that when sets are endowed with more structure, the maps should take these structures into account; we say, "respect" the structures. We are about to make the definition for rings and later on we will see it for groups. In Algebra II (or other linear algebra courses) you will see it for vector spaces. In each case, the way to define correctly functions should be evident; one should be very careful not to ignore all the special features of the special sets (rings, groups, vector spaces, ...) that we are considering.

Definition 20.0.1. Let R, S be rings. A function $f\colon R \to S$ is a **ring homomorphism** if the following holds:
(1) $f(1_R) = 1_S$;
(2) $f(r_1 + r_2) = f(r_1) + f(r_2)$;
(3) $f(r_1 r_2) = f(r_1)f(r_2)$.

Here are some formal consequences (that are nonetheless very useful).

- $f(0_R) = 0_S$. Indeed, $f(0_R) = f(0_R + 0_R) = f(0_R) + f(0_R)$. Let $y = f(0_R)$ then $y = y + y$. Adding $-y$ to both sides we find $0_S = y = f(0_R)$.
- We have $f(-r) = -f(r)$. Indeed: $0_S = f(0_R) = f(r + (-r)) = f(r) + f(-r)$ and so $f(-r) = -f(r)$ (just because it sums with $f(r)$ to 0_S!)
- We have $f(r_1 - r_2) = f(r_1) - f(r_2)$, because $f(r_1 - r_2) = f(r_1 + (-r_2))$ (this, by definition) and so $f(r_1 - r_2) = f(r_1) + f(-r_2) = f(r_1) - f(r_2)$.

Note, in particular, that $f(0_R) = 0_S$ is a consequence of axioms (2), (3). On the other hand $f(1_R) = 1_S$ does not follow from (2), (3) and we therefore include it as an axiom (though not all authors do that). Here is an example. Consider,

$$f : \mathbb{R} \to \mathbb{R} \times \mathbb{R}, \quad f(r) = (r, 0).$$

This map satisfies $f(r_1 + r_2) = f(r_1) + f(r_2)$ and $f(r_1 r_2) = f(r_1)f(r_2)$, but $f(1) = (1, 0)$ is not the identity element of $\mathbb{R} \times \mathbb{R}$. So this is **not** a ring homomorphism.

On the other hand, if $S \subset R$ is a subring then the inclusion map $i\colon S \to R, i(s) = s$, is a ring homomorphism. Note that this explains why in the definition of a subring we insisted on $1_R \in S$.

Proposition 20.0.2. *Let $f : R \to S$ be a homomorphism of rings. The image of f is a subring of S.*

Proof. As we have seen, $f(0_R) = 0_S$. Also, by definition $f(1_R) = 1_S$ and so $0_S, 1_S \in \mathrm{Im}(f)$. Let now $s_1, s_2 \in \mathrm{Im}(f)$, say $s_i = f(r_i)$. Then, $s_1 \pm s_2 = f(r_1) \pm f(r_2) = f(r_1 \pm r_2)$ and so $s_1 \pm s_2 \in \mathrm{Im}(f)$. Similarly, $s_1 s_2 = f(r_1 r_2)$ and so $s_1 s_2 \in \mathrm{Im}(f)$. □

Definition 20.0.3. Let $f : R \to S$ be a homomorphism of rings. The **kernel** of f, $\mathrm{Ker}(f)$, is defined as follows:

$$\mathrm{Ker}(f) = \{r \in R : f(r) = 0\}.$$

Proposition 20.0.4. $\mathrm{Ker}(f)$ *is an ideal of R. The map f is injective if and only if* $\mathrm{Ker}(f) = \{0\}$.

Proof. First, since $f(0_R) = 0_S$ we have $0_R \in \mathrm{Ker}(f)$. Suppose that $r_1, r_2 \in \mathrm{Ker}(f)$ then $f(r_i) = 0_S$ and we find that $f(r_1 + r_2) = f(r_1) + f(r_2) = 0_S + 0_S = 0_S$, so $r_1 + r_2 \in \mathrm{Ker}(f)$.

Now suppose that $r_1 \in \mathrm{Ker}(f)$ and $r \in R$ is any element. We need to show that $rr_1, r_1 r \in \mathrm{Ker}(f)$. We calculate $f(rr_1) = f(r)f(r_1) = f(r)0_S = 0_S$, so $rr_1 \in \mathrm{Ker}(f)$. Similarly for $r_1 r$.

So far we proved that $\mathrm{Ker}(f)$ is an ideal. Suppose now that f is injective. Then $f(r) = 0_S$ implies $f(r) = f(0_R)$ and so $r = 0_R$. That is, $\mathrm{Ker}(f) = \{0_R\}$.

Suppose conversely that $\mathbf{Ker}(f) = \{0_R\}$. If $f(r_1) = f(r_2)$ then $0_S = f(r_1) - f(r_2) = f(r_1 - r_2)$ and so $r_1 - r_2 \in \mathbf{Ker}(f)$. Since $\mathbf{Ker}(f) = \{0_R\}$, we must have $r_1 - r_2 = 0_R$; that is, $r_1 = r_2$. We proved that f is injective. □

We now look at some examples:

Example 20.0.5. Let $n \geq 1$ be an integer. Define a function,

$$f\colon \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z},$$

by $f(a) = \bar{a}$ (the congruence class of a modulo n). Then f is a homomorphism:

(1) $f(1) = \bar{1}$ and $\bar{1}$ is the indeed the identity element of $\mathbb{Z}/n\mathbb{Z}$;
(2) $f(a+b) = \overline{a+b} = \bar{a} + \bar{b} = f(a) + f(b)$;
(3) $f(ab) = \overline{ab} = \bar{a}\,\bar{b} = f(a)f(b)$.

We observe that f is surjective and its kernel is $\{a : \bar{a} \equiv 0 \pmod{n}\} = \{a : n|a\} = n\mathbb{Z}$.

Example 20.0.6. Let R, S, be any rings and define

$$f\colon R \times S \to R, \qquad f((r,s)) = r.$$

This is a homomorphism:

(1) $f(1_{R\times S}) = f((1_R, 1_S)) = 1_R$;
(2) $f((r_1,s_1) + (r_2,s_2)) = f((r_1 + r_2, s_1 + s_2)) = r_1 + r_2 = f((r_1,s_1)) + f((r_2,s_2))$;
(3) $f((r_1,s_1)(r_2,s_2)) = f((r_1r_2, s_1s_2)) = r_1r_2 = f((r_1,s_1))f((r_2,s_2))$.

The kernel of f is $\{(r,s) : r = 0\} = \{(0,s) : s \in S\} = \{0\} \times S$.

Example 20.0.7. Let $\mathbb{F}$ be a field and $\mathbb{F}[\epsilon]$ the ring of dual numbers. Define

$$f\colon \mathbb{F}[\epsilon] \to \mathbb{F}, \qquad f(a + b\epsilon) = a.$$

Then f is a homomorphism:

(1) $f(1) = 1$;
(2) $f((a+b\epsilon) + (c+d\epsilon)) = f(a + c + (b+d)\epsilon) = a + c = f(a+b\epsilon) + f(c+d\epsilon)$;
(3) $f((a+b\epsilon)(c+d\epsilon)) = f(ac + (ad+bc)\epsilon) = ac = f(a+b\epsilon)f(c+d\epsilon)$.

The kernel of f is $\{a + b\epsilon : a = 0, b \in \mathbb{F}\} = \{b\epsilon : b \in \mathbb{F}\}$. We claim that this is the ideal (ϵ). On the one hand $b\epsilon$ certainly is in (ϵ) for any b. That is $\mathbf{Ker}(f) \subseteq (\epsilon)$. On the other hand $(c+d\epsilon)\epsilon = c\epsilon$ and that shows $(\epsilon) \subseteq \mathbf{Ker}(f)$.

Example 20.0.8. Let $\mathbb{F}$ be a field. Let $a \in \mathbb{F}$ be a fixed element. Define

$$\alpha\colon \mathbb{F}[x] \to \mathbb{F}, \qquad \alpha(g(x)) = g(a).$$

Then α is a homomorphism:

(1) $\alpha(1)$ is the value of the constant polynomial 1 at a which is just 1, so $\alpha(1) = 1$.
(2) We have $\alpha(f+g) = (f+g)(a) = f(a) + g(a) = \alpha(f) + \alpha(g)$;
(3) Similarly, $\alpha(fg) = (fg)(a) = f(a)g(a) = \alpha(f)\alpha(g)$.

Therefore, α is a homomorphism. It is called a **specialization homomorphism** or an **evaluation homomorphism**. The kernel of α is $\{f \in \mathbb{F}[x] : f(a) = 0\}$ and it is equal to the principal ideal $(x-a)$. Indeed: if $g(x) \in (x-a)$ then $g(x) = (x-a)g_1(x)$ and so $g(a) = (a-a)g_1(a) = 0$. Conversely, if $g(a) = 0$, Theorem 16.5.1 says that $g(x) = (x-a)g_1(x)$ for some polynomial $g_1(x)$ and so $g(x) \in (x-a)$.

Example 20.0.9. Let A be the set of all continuous functions $f\colon [0,1] \to \mathbb{R}$. Define the sum (respectively, product) of two functions f, g to be the function $f+g$ (resp. fg) whose value at any x is $f(x) + g(x)$ (resp. $f(x)g(x)$). That is:

$$(f+g)(x) = f(x) + g(x), \quad (fg)(x) = f(x)g(x).$$

This is a ring (in particular, these are *operations* – the sum and product of continuous functions is continuous!). Its zero element is the constant function zero and its identity element is the constant function 1. Let $a \in [0,1]$ be a fixed element. Define

$$\varphi\colon A \to \mathbb{R}, \qquad \varphi(f) = f(a).$$

Then φ is a ring homomorphism whose kernel are all the functions vanishing at the point a.

20.1. **Units.** Let R be any ring. Recall the definition of units: The **units** of R are denoted $R^\times$ and defined as follows:

$$R^\times = \{x \in R : \exists y \in R, xy = yx = 1\}.$$

For example, 1_R is always a unit. If R is a field then, by definition, $R^\times = R - \{0\}$.

Lemma 20.1.1. *We have the following properties:*

(1) If $r_1, r_2 \in R^\times$ then $r_1 r_2 \in R^\times$.
(2) Let $f: R \to S$ be a homomorphism of rings then $f(R^\times) \subseteq S^\times$.

Proof. Suppose that $r_1, r_2 \in R^\times$ and $r_1 y_1 = y_1 r_1 = 1, r_2 y_2 = y_2 r_2 = 1$. Let $y = y_2 y_1$ then $(r_1 r_2)y = r_1(r_2 y_2)y_1 = r_1 \cdot 1 \cdot y_1 = r_1 y_1 = 1$. A similar computation gives $y(r_1 r_2) = 1$ and so $r_1 r_2 \in R^\times$.

Let now $f: R \to S$ be a homomorphism and $r \in R^\times$ with $ry = yr = 1_R$. Then $f(r)f(y) = f(ry) = f(1_R) = 1_S$ and $f(y)f(r) = f(yr) = f(1_R) = 1_S$. It follows that $f(r) \in S^\times$. □

Example 20.1.2. We have $\mathbb{Z}^\times = \{\pm 1\}$. We have $\mathbb{Q}^\times = \mathbb{Q} - \{0\}$.

Example 20.1.3. We have $\mathbb{F}[\epsilon]^\times = \{a + b\epsilon : a \neq 0\}$. Indeed, if $a \neq 0$ then $(a + b\epsilon)(a^{-1} - a^{-2}b\epsilon) = 1$ (where a^{-2} is by definition $(a^2)^{-1}$. It satisfies $a^{-2}a = a^{-1}$). Conversely, if $(a + b\epsilon)(c + d\epsilon) = 1$ then $ac = 1$ and so $a \neq 0$.

Example 20.1.4. Let $n \neq 0, 1$ be a square free integer. Recall that $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a, b \in \mathbb{Z}\}$ and every element of this ring has a unique expression as $a + b\sqrt{n}$. We claim that

$$\mathbb{Z}[\sqrt{n}]^\times = \{a + b\sqrt{n} : a^2 - b^2 n = \pm 1\}.$$

Indeed, if $a^2 - b^2 n = \pm 1$ then $(a + b\sqrt{n})(a - b\sqrt{n}) = \pm 1$ and so $a + b\sqrt{n}$ is invertible with inverse $\pm(a - b\sqrt{n})$. Conversely, if $a + b\sqrt{n}$ is invertible, say $(a + b\sqrt{n})(c + d\sqrt{n}) = 1$ (for some $c, d \in \mathbb{Z}$) then $ad + bc = 0$ and so also $(a - b\sqrt{n})(c - d\sqrt{n}) = 1$. We get that

$$(a + b\sqrt{n})(a - b\sqrt{n})(c + d\sqrt{n})(c - d\sqrt{n}) = 1.$$

But $(a + b\sqrt{n})(a - b\sqrt{n}) = a^2 - b^2 n$ and $(c + d\sqrt{n})(c - d\sqrt{n}) = c^2 - d^2 n$ are integers. So

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

If n is negative, then it is easy to see that the unique solutions are $a = \pm 1, b = 0$, except when $n = -1$ where also $b = \pm 1$ is a solution. Namely, in general, only ± 1 are units for $n < 0$, but for $\mathbb{Z}[i]$ the units are $\pm 1, \pm i$. On the other hand, for $n > 1$ it turns out that there are infinitely many units; there are infinitely many solutions (a, b) to the so-called Pell equation

$$a^2 - b^2 n = 1.$$

This is not an easy statement. It is interesting to try and prove this, but don't be discouraged if you can't. The Pell equation has been studied much and is related to the Archimedes Cattle problem that asks for "the number of the cattle of the sun which once grazed upon the plains of Sicily". It is stated in the form of a poem, ultimately reducing to a Pell equation, the solutions of which are just enormous. Finding the smallest solution to a Pell equation is a complicated problem, related to the continued fraction expression of $\sqrt{n}$. It behaves rather erratically. For example, the smallest solution to $a^2 - 60 \cdot b^2 = 1$ is $a = 31, b = 4$. In contrast, the smallest solution to $a^2 - 61 \cdot b^2 = 1$ is $a = 1766319049, b = 226153980$.

Example 20.1.5. Let $\mathbb{F}$ be a field. The units of the ring $M_2(\mathbb{F})$ are the matrices

$$\mathrm{GL}_2(\mathbb{F}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc \neq 0 \right\}.$$

Indeed, suppose that for the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ we have $ad - bc \neq 0$. Consider the matrix

$$(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

(where by $t\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ we mean $\begin{pmatrix} ta & tb \\ tc & td \end{pmatrix}$. It is equal to $\begin{pmatrix} a & b \\ c & d \end{pmatrix}t$). We claim that this is the inverse. We have

$$\begin{aligned}(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} &= (ad-bc)^{-1}\begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.\end{aligned}$$

Similarly, one checks that $\begin{pmatrix} a & b \\ c & d \end{pmatrix}(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Suppose now that $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is invertible. The expression $ad - bc$ is called the **determinant** of the matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and is denoted $\det(M)$. One can verify by a laborious but straightforward calculation that for any two matrices M, N we have

$$\det(MN) = \det(M)\det(N).$$

If the matrix M has an inverse, say $MN = NM = I_2$, then

$$\det(MN) = \det(M)\det(N) = \det(I_2) = 1,$$

and that shows that $\det(M) \neq 0$. One can then show that N is necessarily $(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. In fact, a more general fact is true.

Let R be a ring and $x \in R^\times$, $yx = xy = 1$. Suppose also that $zx = xz = 1$. Then $(y-z)x = 1 - 1 = 0$ and so $(y-z)(xy) = 0 \cdot y = 0$. Therefore, since $xy = 1$, we have $y - z = 0$ that is $y = z$.

Applying this to $MN = I_2$ and $M(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = I_2$, we conclude that

$$N = (ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

21. Quotient rings

In this section we construct quotient rings. These are rings constructed out of a given ring by a process of "moding out by an ideal". Our main motivation is constructing finite fields whose cardinality is a power of a prime number, generalizing the fields $\mathbb{Z}/p\mathbb{Z}$. Finite fields, historically one of the most esoteric aspects of algebra, should be considered today also as part of applied mathematics. Their usefulness to computer

science and engineering, mainly through the subjects of cryptography, coding theory and complexity theory, is enormous. Knowing to construct and compute in finite fields is one of the main goals of this course.

Consider a surjective ring homomorphism $f\colon R \to S$. Given an element $s \in S$ let r be an element of R such that $f(r) = s$. How unique is r? If $a \in I := \mathrm{Ker}(f)$ then $f(r+a) = f(r) + f(a) = f(r) + 0 = f(r)$. Conversely, if $f(r_1) = s$ then $f(r_1 - r) = f(r_1) - f(r) = s - s = 0$ so $a := r_1 - r \in I$ and $r_1 = r + a$. Let us use the notation

$$r + I = \{r + i : i \in I\}.$$

We proved that if r is any element of R such that $f(r) = s$ then

$$f^{-1}(s) = r + I.$$

Thus, in a sense, we may identify elements of S with cosets of R and from this point of view we may say that the cosets (thought of as being the elements of S) form a ring.

In this section we perform a key construction that eliminates the need in S. Given a ring R and a two-sided ideal $I \lhd R$ we construct a new ring R/I, whose elements are cosets of I.

Definition 21.0.1. Let R be a ring and $I \lhd R$ a two sided ideal. A **coset** of I is a subset of R of the form

$$a + I := \{a + i : i \in I\},$$

where is a is an element of R.

Example 21.0.2. Suppose that $R = \mathbb{Z}$ and $I = (n)$ for some positive integer n. Then,

$$a + (n) = \{\dots, a - n, a, a + n, a + 2n, \dots\},$$

are precisely the integers congruent to a modulo n. Thus, a coset of the ideal (n) is *the same thing* as a congruence class modulo n.

Lemma 21.0.3. *We have the following facts:*

(1) Every element of R belongs to a coset of I.
(2) Two cosets are either equal or disjoint.
(3) The following are equivalent: (i) $a + I = b + I$; (ii) $a \in b + I$; (iii) $a - b \in I$.

Proof. The first claim is easy: the element r belongs to the coset $r + I$, because $r = r + 0$ and $0 \in I$.

Suppose that $a + I \cap b + I \neq \emptyset$. Then, there is an element of R that can be written as

$$a + i_1 = b + i_2,$$

for some $i_1, i_2 \in I$. We show that $a + I \subset b + I$; by symmetry we have the opposite inclusion and so the cosets are equal. An element of $a + I$ has the form $a + i$ for some $i \in I$. We have $a + i = b + (i_2 - i_1) + i = b + (i_2 - i_1 + i)$. Note that $i_2 - i_1 + i \in I$ and so $a + i \in b + I$.

We next prove the equivalence of (i), (ii) and (iii). Clearly (i) implies (ii) because $a \in a + I$. If (ii) holds then $a = b + i$ for some $i \in I$ and so $a - b = i \in I$ and (iii) holds. If (iii) holds then $a - b = i$ for some $i \in I$, and so $a \in a + I$ and also $a = b + i \in b + I$. That is, $a + I \cap b + I \neq \emptyset$ and so $a + I = b + I$. □

Theorem 21.0.4. *Let R be a ring and $I \lhd R$ a two-sided ideal. Denote the collection of cosets of I in R by R/I. Define addition by*

$$(a + I) + (b + I) = (a + b) + I,$$

and multiplication by

$$(a + I)(b + I) = ab + I.$$

These operations are well defined and make R/I into a ring (a **quotient ring***) with zero element $0 + I = I$ and identity element $1 + I$.*

Proof. First, our definition of the operations makes use of writing a coset as $a + I$. This way of writing is not unique and so we should check that our definitions are independent of the choice of the element a such that the coset is equal to $a + I$. Namely, if

$$a + I = a' + I, \quad b + I = b' + I,$$

we need to check that

$$a+b+I=a'+b'+I, \qquad ab+I=a'b'+I.$$

Now, $(a+b)-(a'+b')=(a-a')+(b-b')$. By Lemma 21.0.3, $a-a'\in I, b-b'\in I$ and so $(a+b)-(a'+b')\in I$. Therefore, by the same lemma, $a+b+I=a'+b'+I$. Also $ab-a'b'=(a-a')b+a'(b-b')$. Now, $a-a'\in I, b-b'\in I$ and so $(a-a')b\in I, a'(b-b')\in I$ and thus $ab-a'b'\in I$. Therefore, $ab+I=a'b'+I$.

We now verify the ring axioms. It will be convenient to write $\bar{a}$ for $a+I$. With this notation we have

$$\bar{a}+\bar{b}=\overline{a+b}, \quad \bar{a}\,\bar{b}=\overline{ab}.$$

The axioms follow from the definition of the operations and the fact that they hold for R. To make clear at what point we use that the axioms hold in R, we use the notation $\stackrel{!}{=}$ to draw attention to this.

(1) $\bar{a}+\bar{b}=\overline{a+b}\stackrel{!}{=}\overline{b+a}=\bar{b}+\bar{a}$.

(2) $\bar{a}+(\bar{b}+\bar{c})=\bar{a}+\overline{b+c}=\overline{a+(b+c)}\stackrel{!}{=}\overline{(a+b)+c}=\overline{a+b}+\bar{c}=(\bar{a}+\bar{b})+\bar{c}$.

(3) We have $\bar{0}+\bar{a}=\overline{0+a}\stackrel{!}{=}\bar{a}$. (We remark that $\bar{0}=I$.)

(4) We have $\bar{a}+\overline{-a}=\overline{a+(-a)}\stackrel{!}{=}\bar{0}$.

(5) $\bar{a}(\bar{b}\,\bar{c})=\bar{a}\,\overline{bc}=\overline{a(bc)}\stackrel{!}{=}\overline{(ab)c}=\overline{ab}\,\bar{c}=(\bar{a}\,\bar{b})\,\bar{c}$.

(6) We have $\bar{a}\,\bar{1}=\overline{a\,1}\stackrel{!}{=}\bar{a}$ and $\bar{1}\,\bar{a}=\overline{1\,a}\stackrel{!}{=}\bar{a}$.

(7) $(\bar{a}+\bar{b})\bar{c}=\overline{a+b}\;\bar{c}=\overline{(a+b)c}\stackrel{!}{=}\overline{ac+bc}=\overline{ac}+\overline{bc}=\bar{a}\;\bar{c}+\bar{b}\;\bar{c}$. Also, $\bar{c}(\bar{a}+\bar{b})=\bar{c}\;\overline{a+b}=\overline{c(a+b)}\stackrel{!}{=}\overline{ca+cb}=\overline{ca}+\overline{cb}=\bar{c}\,\bar{a}+\bar{c}\,\bar{b}$.

□

Proposition 21.0.5. *The natural map,*

$$\pi\colon R\to R/I, \qquad a\mapsto\pi(a):=\bar{a}$$

is a surjective ring homomorphism with kernel I. Thus, every ideal $I\lhd R$ is the kernel of some ring homomorphism from R to some other ring.

Proof. Note that $1\mapsto\bar{1}$, which is the identity element of R/I. We have $\pi(a+b)=\overline{a+b}=\bar{a}+\bar{b}=\pi(a)+\pi(b)$. Also, $\pi(ab)=\overline{ab}=\bar{a}\,\bar{b}=\pi(a)\,\pi(b)$. We have shown that π is a ring homomorphism and it is clearly surjective.

The kernel of π are the elements $a\in R$ such that $\pi(a)=\bar{a}=\bar{0}$, namely, the elements a such that $a+I=0+I$. By Lemma 21.0.3 this is the set of elements a such that $a-0\in I$, namely, the kernel is precisely I. □

Example 21.0.6. Consider the ring $\mathbb{Z}$. If we take the ideal $\{0\}$ then $\mathbb{Z}/\{0\}$ can be identified with $\mathbb{Z}$; the map $\mathbb{Z}\to\mathbb{Z}/\{0\}$ is a bijective ring homomorphism. Let $n>0$ then. The ring $\mathbb{Z}/(n)$ has as elements the cosets $a+(n)$. Two cosets $a+(n), b+(n)$ are equal if and only if $a-b\in(n)$, that is, precisely when $n|(a-b)$. We see that the elements of $\mathbb{Z}/(n)$ are just the congruence classes modulo n, as we have in fact noted before, and the operations on $\mathbb{Z}/(n)$ are just the operations we defined on congruence classes.

Thus, the quotient rings of $\mathbb{Z}$ are $\mathbb{Z}/n\mathbb{Z}$ (either $\mathbb{Z}$, if $n=0$, or the familiar rings of congruences when $n\neq 0$). In particular, if p is a prime number we get the field $\mathbb{Z}/p\mathbb{Z}$ of p elements. Following on the analogy between $\mathbb{Z}$ and $\mathbb{F}[x]$, it is natural to examine next the quotient rings of $\mathbb{F}[x]$. We shall see that in fact we can get this way fields and in particular fields whose cardinality is any power of a prime (in contrast $\mathbb{Z}/p^a\mathbb{Z}$ is never a field for $a>1$). It is a fact that any finite field has cardinality a power of a prime, so the methods we develop in this course produce **all** finite fields. We shall not prove, though, in this course that any finite field has cardinality a power of a prime, or that we get all finite fields this way. This is done usually in Algebra IV.

21.1. **The quotient ring $\mathbb{F}[x]/(f(x))$.** Let $\mathbb{F}$ be a field , $f(x)\in\mathbb{F}[x]$ a non-constant polynomial and $(f(x))=f(x)\cdot\mathbb{F}[x]$ the principal ideal it defines. Consider the quotient ring $\mathbb{F}[x]/(f(x))$. Suppose that $f(x)=x^n+a_{n-1}x^{n-1}+\cdots+a_0$ is a monic polynomial of degree n. The following lemma is an analogue of Lemma 12.0.1.

Lemma 21.1.1. *Every element of $\mathbb{F}[x]/(f(x))$ is of the form $\overline{g(x)} := g(x) + (f(x))$ for a unique polynomial $g(x)$ which is either zero or of degree less than n.*

Proof. Let $h(x)$ be a polynomial. To say that we have equality of cosets, $h(x) + (f(x)) = g(x) + (f(x))$, is to say that $h(x) = q(x)f(x) + g(x)$. The requirement that $\deg(g) < \deg(f)$ amounts to the assertion that the expression

$$h(x) = q(x)f(x) + g(x)$$

is the one gotten by dividing h by f with residue. We know that this is always possible and in a unique fashion. □

Theorem 21.1.2. *Let $\mathbb{F}$ be a field, $f(x) \in \mathbb{F}[x]$ a non-constant irreducible polynomial of degree n. The quotient ring $\mathbb{F}[x]/(f(x))$ is a field. If $\mathbb{F}$ is a finite field of cardinality q then $\mathbb{F}[x]/(f(x))$ is a field with q^n elements.*

Proof. We already know that $\mathbb{F}[x]/(f(x))$ is a commutative ring. We note that $\bar{0} \neq \bar{1}$ because $1 \notin (f)$ (if it were, f would be a constant polynomial). Thus, we only need to show that a non-zero element has an inverse. Let $\overline{g(x)}$ be a non-zero element. That means that $g(x) \notin (f(x))$, thus $f(x) \nmid g(x)$, and so $\gcd(f,g) = 1$ (here is where we use that f is irreducible). Therefore, there are polynomials $u(x), v(x)$ such that

$$u(x)f(x) + v(x)g(x) = 1.$$

Passing to the quotient ring, that means that $\bar{v}\bar{g} = \bar{1}$, and $\bar{1}$ is the identity of the quotient ring. Therefore $\bar{g}$ is invertible (and its inverse is $\bar{v}$).

Finally, by the Lemma, every element of $\mathbb{F}[x]/(f(x))$ has a unique representative of the form $a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, where $a_0, a_1, \ldots, a_{n-1}$ are elements of $\mathbb{F}$. If $\mathbb{F}$ has q elements, we get q^n such polynomials as q^n is the number of choices for the coefficients $a_0, a_1, \ldots, a_{n-1}$. □

Example 21.1.3. A field with 4 elements. Take the field $\mathbb{F}$ to be $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ and consider the polynomial $x^2 + x + 1$ over that field. Because it is of degree 2 and has no root in $\mathbb{F}_2$ it must be irreducible. Therefore, $\mathbb{F}_2[x]/(x^2 + x + 1)$ is a field $\mathbb{K}$ with 4 elements. Let us list its elements:

$$\mathbb{K} = \{\bar{0}, \bar{1}, \bar{x}, \overline{x+1}\}.$$

(This is the list of polynomials $a_0 + a_1x$ with $a_0, a_1 \in \mathbb{F}_2$.) We can describe the addition and multiplication by tables:

$+$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{1}$	$\bar{1}$	$\bar{0}$	$\overline{x+1}$	$\bar{x}$
$\bar{x}$	$\bar{x}$	$\overline{x+1}$	$\bar{0}$	$\bar{1}$
$\overline{x+1}$	$\overline{x+1}$	$\bar{x}$	$\bar{1}$	$\bar{0}$

,

$\cdot$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{x}$	$\bar{0}$	$\bar{x}$	$\overline{x+1}$	$\bar{1}$
$\overline{x+1}$	$\bar{0}$	$\overline{x+1}$	$\bar{1}$	$\bar{x}$

Example 21.1.4. A field with 9 elements. Consider the polynomial $x^2 + 1$ over $\mathbb{F}_3 = \mathbb{Z}/3\mathbb{Z}$. It is quadratic and has no root in $\mathbb{F}_3$, hence is irreducible over $\mathbb{F}_3$. We conclude that $L = \mathbb{F}_3[x]/(x^2 + 1)$ is a field with 9 elements. Note that in $\mathbb{F}_3$ the element $-1 = 2$ is not a square. However, in L we have $x^2 = x^2 - (x^2 + 1) = -1$ and so -1 is a square now; its root is x (viewed as an element of L). In fact, any quadratic polynomial over $\mathbb{F}_3$ has a root in L, because the discriminant "$b^2 - 4ac$" is either $0, 1, 2$ and all those are squares in L.

For example, consider the polynomial $t^2 + t + 2$. It has discriminant $-7 \equiv -1 \pmod 3$, which is not a square in $\mathbb{F}_3$. In the field L we have $x^2 = -1$. The solutions of the polynomial are then $(-1 \pm \sqrt{-1})/2 = 2(-1 \pm x) = 1 \pm 2x$.

On the other hand, one can prove that the polynomial t^3+t^2+2 is irreducible in $\mathbb{F}_3$ and stays irreducible in L. In MATH 457 we learn a systematic theory for deciding which polynomials stay irreducible and which do not.

Example 21.1.5. Fields with 8 and 16 elements. A polynomial of degree 3 is irreducible if and only if it doesn't have a root. We can verify that x^3+x+1 doesn't have a root in $\mathbb{F}_2=\mathbb{Z}/2\mathbb{Z}$ and conclude that $\mathbb{F}_2[x]/(x^3+x+1)$ is a field with 8 elements. Consider the field $\mathbb{K}$ with 4 elements constructed above. We note that the polynomial $t^2+t+\bar{x}$ is irreducible over $\mathbb{K}$ (simply by substituting for t any of the four elements of $\mathbb{K}$ and checking). Thus, we get a field $\mathbb{L}$ with 16 elements

$$\mathbb{L}=\mathbb{K}[t]/(t^2+t+\bar{x}).$$

Remark 21.1.6. To construct a finite field of p^n elements, where p is a prime number, we need to find a polynomial of degree n which is irreducible over the field of $\mathbb{Z}/p\mathbb{Z}$. One can prove that such a polynomial always exists by a counting argument. Finding a specific one for a given p and n is harder. Nonetheless, given p and n one can find such an irreducible polynomial and so construct a field of p^n elements explicitly, for example in the sense that one can write a computer program that makes calculations in such a field.

21.2. **Every polynomial has a root in a bigger field.**

Theorem 21.2.1. *Let $\mathbb{F}$ be a field and $f(x)\in\mathbb{F}[x]$ a non-constant polynomial. There is a field L containing $\mathbb{F}$ and an element $\ell\in L$ such that $f(\ell)=0$.*

Proof. If $g|f$ and $g(\ell)=0$ then also $f(\ell)=0$, so we may assume that f is irreducible. Let $L=\mathbb{F}[x]/(f(x))$. This is a field. We have a natural map $\mathbb{F}\to L$, $a\mapsto\bar{a}$. This map is an injective ring homomorphism and we identify $\mathbb{F}$ with its image in L and say that $L\supset\mathbb{F}$.

Now, suppose that $f(x)=a_nx^n+\dots a_1x+a_0$. To say that f has a root in L is to say that for some element $\ell\in L$ we have

$$a_n\ell^n+\cdots+a_1\ell+a_0=0.$$

We check that this hold for the element $\ell=\bar{x}$. Indeed,

$$a_n\bar{x}^n+\cdots+a_1\bar{x}+a_0=\overline{a_nx^n+\cdots+a_1x+a_0}=\overline{f(x)}=0_L.$$

□

Example 21.2.2. According to this result, -1 has a square root in the field $\mathbb{R}[x]/(x^2+1)$. One can show that $\mathbb{R}[x]/(x^2+1)\cong\mathbb{C}$.

21.3. **Roots of polynomials over $\mathbb{Z}/p\mathbb{Z}$.**

We can now continue our discussion, begun in § 16.7, of the efficient determination of whether a small degree polynomial $f(x)$ over $\mathbb{Z}/p\mathbb{Z}$ has a root in $\mathbb{Z}/p\mathbb{Z}$. Recall that the only remaining point was whether the Euclidean algorithm step,

$$x^p-x=q(x)f(x)+r(x),$$

can be done rapidly. Now we can answer that affirmatively. Note that $r(x)+x$ is exactly the representative of x^p in the ring $\mathbb{F}[x]/(f(x))$. This representative can be calculated quickly by the method we already used for calculating powers. We need to calculate

$$x,x^2,x^4,x^8,\dots$$

and express p in base 2, $p=\sum a_i2^i, a_i\in\{0,1\}$, $x^p=\prod_{\{i:a_i\neq0\}}x^{2^i}$ and so on. We see that the slowing factor now is how quickly we can carry out multiplication in the ring $\mathbb{F}[x]/(f(x))$. It is not hard to see that this depends on the degree of f and not on p.

Let us illustrate this by finding if x^3+x+1 has a root in the field with 17 elements. We calculate x^{17} in the quotient ring $\mathbb{L}=\mathbb{F}_{17}[x]/(x^3+x+1)$. We have x,x^2,

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

$x^8=(x^2+x)^2=x^4+2x^3+x^2=-(x^2+x)+2(-x-1)+x^2=-3x-2$,

$x^{16}=(3x+2)^2=9x^2+12x+4$

and so

$x^{17} - x = x(9x^2 + 12x + 4) - x = 9(-x - 1) + 12x^2 + 3x = 12x^2 - 6x - 9.$

This is the residue of dividing $x^{17} - x$ in $x^3 + x + 1$. Now we continue with the Euclidean algorithm, in the way we are used to.

$x^3 + x + 1 = (10x + 5)(12x^2 - 6x - 9) + 2x + 12,$

$12x^2 - 6x - 9 = (6x - 5)(2x + 12)$

and it follows that

$$\gcd(x^{17} - x, x^3 + x + 1) = x + 6,$$

and so that $x^3 + x + 1$ has a unique root modulo 17, equal to 11. It is easy to verify this conclusion: $x^3 + x + 1 = (x + 6)(x^2 + 11x + 3)$, which shows that $x = 11$ is indeed a root. The polynomial $x^2 + 11x + 3$ has discriminant $121 - 12 = 109 \equiv 7 \pmod{17}$, which is not a square modulo 17 (by direct verification), hence it is irreducible modulo 17 and there are no other roots.

22. The First Isomorphism Theorem

22.1. **Isomorphism of rings.**

Definition 22.1.1. Let R, S be rings. A ring homomorphism $f\colon R \to S$ is called an **isomorphism** if f is bijective.

Lemma 22.1.2. *If $f\colon R \to S$ is a ring isomorphism then the inverse function $g = f^{-1}$, $g\colon S \to R$ is also a ring homomorphism, hence an isomorphism. (The inverse function is defined by $g(s) = r$, where r is the unique element such that $f(r) = s$.)*

Proof. First, because $f(1_R) = 1_S$ we have $g(1_S) = 1_R$. Next, let $s_1, s_2 \in S$. We need to prove $g(s_1 + s_2) = g(s_1) + g(s_2)$ and $g(s_1 s_2) = g(s_1)g(s_2)$. It is enough to prove that

$$f(g(s_1 + s_2)) = f(g(s_1) + g(s_2)), \qquad f(g(s_1 s_2)) = f(g(s_1)g(s_2)),$$

because f is injective. But $f(g(s_1) + g(s_2)) = f(g(s_1)) + f(g(s_2)) = s_1 + s_2 = f(g(s_1 + s_2))$ and $f(g(s_1)g(s_2)) = f(g(s_1))f(g(s_2)) = s_1 s_2 = f(g(s_1 s_2))$. □

Definition 22.1.3. Let R, S be rings. We say that R and S are isomorphic if there is a ring isomorphism $R \to S$.

Lemma 22.1.4. *Being isomorphic is an equivalence relation on rings.*

Proof. First, the identity function is always a ring homomorphism from R to R, so this relation is reflexive. Secondly, if $f\colon R \to S$ is an isomorphism then $g\colon S \to R$ is an isomorphism, where g is the inverse function to f. Thus, the relation is symmetric. Now suppose $f\colon R \to S$ and $g\colon S \to T$ are ring isomorphisms between the rings R, S, T. To show the relation is transitive we need to prove that $g \circ f\colon R \to T$ is an isomorphism. Indeed:

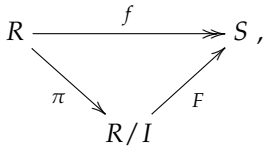
(1) $(g \circ f)(1_R) = g(f(1_R)) = g(1_S) = 1_T$;
(2) $(g \circ f)(r_1 + r_2) = g(f(r_1 + r_2)) = g(f(r_1) + f(r_2)) = g(f(r_1)) + g(f(r_2)) = (g \circ f)(r_1) + (g \circ f)(r_2)$;
(3) $(g \circ f)(r_1 r_2) = g(f(r_1 r_2)) = g(f(r_1)f(r_2)) = g(f(r_1)) \cdot g(f(r_2)) = (g \circ f)(r_1) \cdot (g \circ f)(r_2)$.

□

We shall denote that R is isomorphic to S by $R \cong S$.

22.2. **The First Isomorphism Theorem.**

Theorem 22.2.1. *Let $f\colon R \to S$ be a surjective homomorphism of rings. Let $I = \mathrm{Ker}(f)$ then there is an isomorphism $F\colon R/I \to S$, such that the following diagram commutes*



where $\pi\colon R \to R/I$ is the canonical map $g \mapsto \bar{g}$.

Proof. We define a function,

$$F\colon R/I \to S,$$

by

$$F(\bar{g}) = f(g).$$

We first prove that this map is well defined. Suppose that $\bar{g} = \overline{g_1}$. We need to show that $f(g) = f(g_1)$. This holds because $\bar{g} = \overline{g_1}$ means $g - g_1 \in I = \mathrm{Ker}(f)$. Now:

- $F(1_{R/I}) = F(\overline{1_R}) = f(1_R) = 1_S$;
- $F(\bar{g} + \bar{h}) = F(\overline{g+h}) = f(g+h) = f(g) + f(h) = F(\bar{g}) + F(\bar{h})$;
- $F(\bar{g}\,\bar{h}) = F(\overline{gh}) = f(gh) = f(g) \cdot f(h) = F(\bar{g}) \cdot F(\bar{h})$.

We also have

$$(F \circ \pi)(g) = F(\bar{g}) = f(g),$$

so $F \circ \pi = f$. Because of this we have that F is surjective. We next show F is injective. Suppose that $F(\bar{g}) = 0_S$ then $f(g) = 0_S$ and so $g \in I$. Thus, $\bar{g} = 0_{R/I}$. □

Example 22.2.2. We consider again the homomorphism $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$. It is a surjective ring homomorphism with kernel given by the principal ideal (n) and we conclude that

$$\mathbb{Z}/(n) \cong \mathbb{Z}/n\mathbb{Z},$$

a fact we have noticed somewhat informally before.

Example 22.2.3. We have $\mathbb{R}[x]/(x^2+1) \cong \mathbb{C}$. To show that, define a ring homomorphism

$$\mathbb{R}[x] \to \mathbb{C},$$

by $\sum_{j=0}^{n} a_j x^j \mapsto \sum_{j=0}^{n} a_j i^j$. This is a well defined function taking 1 to 1. It is easy to verify it is a homomorphism. In fact, recall that $\mathbb{C}[x] \to \mathbb{C}$, $f \mapsto f(i)$, is a homomorphism. Our map is the restriction of the evaluation-at-i homomorphism to the subring $\mathbb{R}[x]$ and so is also a homomorphism. It is surjective because, for example, $a + bx \mapsto a + bi$.

The kernel I definitely contains x^2+1, and so all its multiples. That is, I contains the ideal (x^2+1). Because every ideal of $\mathbb{R}[x]$ is principal, $I = (f)$ for some polynomial f. Because $x^2+1 \in (f)$, $f|(x^2+1)$. Since x^2+1 is irreducible over $\mathbb{R}$, either $f \sim 1$ or $f \sim x^2+1$. If $f \sim 1$ we have $(f) = \mathbb{R}[x]$ and so any polynomial is in the kernel, which is clearly not the case (for example, 1 is not in the kernel). Thus $f \sim x^2+1$ and $I = (x^2+1)$; by the First Isomorphism Theorem we have

$$\mathbb{R}[x]/(x^2+1) \cong \mathbb{C}.$$

22.3. **The Chinese Remainder Theorem.**

Theorem 22.3.1. *Let m, n be positive integers such that $(m,n) = 1$. Then*

$$\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}.$$

Proof. We define a function

$$f\colon \mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}, \quad f(a) = (a \pmod m, a \pmod n).$$

This function is a ring homomorphism:

- $f(1) = (1 \pmod m, 1 \pmod n) = (1_{\mathbb{Z}/m\mathbb{Z}}, 1_{\mathbb{Z}/n\mathbb{Z}})$;

- $f(a+b) = (a+b \pmod m, a+b \pmod n) = (a \pmod m, a \pmod n) + (b \pmod m, b \pmod n) = f(a)+f(b)$;
- $f(ab) = (ab \pmod m, ab \pmod n) = (a \pmod m, a \pmod n) \cdot (b \pmod m, b \pmod n) = f(a)f(b)$.

The kernel of the map is the set $\{a : m|a, n|a\} = \{a : mn|a\}$ (using that $(m,n) = 1$), that is, the kernel is the principal ideal (mn). That means that the integers $0, 1, \dots, mn-1$ all have different images in the target. Since the target has mn elements, we conclude that f is surjective. By the First Isomorphism Theorem

$$\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}.$$

□

This theorem is very useful. It says that to solve an equation modulo mn, $(m,n) = 1$, is the same as solving it modulo m and modulo n. That is, for given integers $a_0, \dots, a_n$ and an integer A we have $a_n A^n + \cdots + a_1 A + a_0 \equiv 0 \pmod{mn}$ if and only if we have $a_n A^n + \cdots + a_1 A + a_0 \equiv 0 \pmod m$ and $a_n A^n + \cdots + a_1 A + a_0 \equiv 0 \pmod n$. Here is an example:

Example 22.3.2. Solve the equation $5x + 2 = 0$ modulo 77.

We consider the equation modulo 7 and get $5x = -2 = 5 \pmod 7$ so $x = 1 \pmod 7$; we consider it modulo 11 and get $5x = -2 = 20 \pmod{11}$ and get that $x = 4 \pmod{11}$. There is an $x \in \mathbb{Z}$ such that $x \pmod 7 = 1, x \pmod{11} = 4$ and in fact x is unique modulo 77 (this is the CRT). We can guess that $x = 15$ will do in this case, but it raises the general problem of finding the inverse isomorphism to $\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, which we address next.

22.3.1. *Inverting $\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$.* Suppose we know how to find integers e_1, e_2 such that $e_1 = 1 \pmod m, e_1 = 0 \pmod n$ and e_2 such that $e_2 = 0 \pmod m, e_2 = 1 \pmod n$, then we would have solved our problem. Indeed, given now two congruence classes $a \pmod m, b \pmod n$ take the integer $ae_1 + be_2$. It is congruent to a modulo m and to b modulo n.

Since $(m,n) = 1$ we may find u, v such that $1 = um + vn$. Put

$$e_1 = 1 - um, \quad e_2 = 1 - vn.$$

These are the integers we are looking for.

Example 22.3.3. Solve the equation $56x + 23 = 0 \pmod{323}$.

We have $323 = 17 \cdot 19$.

- Solution modulo 17.

  We have the equation $5x + 6 = 0 \pmod{17}$. Or $x = -6 \cdot 5^{-1} = 11 \cdot 5^{-1}$. To find 5^{-1} we look for u, v such that $1 = u5 + v17$.

  $17 = 3 \cdot 5 + 2$, $5 = 2 \cdot 2 + 1$ so $1 = 5 - 2 \cdot 2 = 5 - 2 \cdot (17 - 3 \cdot 5) = 7 \cdot 5 - 2 \cdot 17$ and so $7 \cdot 5 = 1 \pmod{17}$. We conclude that $x = 11 \cdot 7 = 77 = 9 \pmod{17}$.
- Solution modulo 19.

  We have the equation $-x + 4 = 0$ so $x = 4 \pmod{19}$ is a solution.
- Finding e_1, e_2.

  We have $19 = 17 + 2$, $17 = 8 \cdot 2 + 1$ so $1 = 17 - 8 \cdot 2 = 9 \cdot 17 - 8 \cdot 19$. It follows that $e_1 = 1 - 9 * 17 = -152, e_2 = 1 + 8 * 19 = 153$.
- We conclude that the solution to the equation $56x + 23 = 0 \pmod{323}$ is $9 * e_1 + 4 * e_2 = -1368 + 612 = -756$ and modulo 323 this is 213.

With linear equations, there is in fact a quicker way to solve the equation that we already know. If the leading coefficient in $ax + b$ is prime to mn, there is a c such that $ca \equiv 1 \pmod{mn}$ (c can be found using the Euclidean algorithm). Then $x = -bc$.

Example 22.3.4. Solve the equation $x^2 = 118 \pmod{323}$. As before we reduce to solving $x^2 = 118 = 16 \pmod{17}$ and $x^2 = 118 = 4 \pmod{19}$. There are two solutions in each case, given by $x = \pm 4 \pmod{17}$ and $x = \pm 2 \pmod{19}$. We conclude that over all we have 4 solutions given by

$$\pm 4(-152) \pm 2(153) \pmod{323}.$$

One can then reduce those numbers to standard representatives and find that $21, 55, 268, 302 \pmod{323}$ are the four solutions.

One can be more precise about the connection between solutions mod mn and solutions mod m and mod n. First, let us generalize the Chinese Remainder Theorem:

Theorem 22.3.5. *Let $m_1, \ldots, m_k$ be relatively prime non-zero integers (that is $(m_i, m_j) = 1$ for $i \neq j$). Then there is an isomorphism*

$$\mathbb{Z}/m_1m_2 \ldots m_k\mathbb{Z} \cong \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_k\mathbb{Z},$$

given by

$$a \pmod{m_1m_2 \ldots m_k} \mapsto (a \pmod{m_1}, a \pmod{m_2}, \ldots, a \pmod{m_k}).$$

The theorem is not hard to prove by induction on k. The main case, $k = 2$, is the one we proved above. Now, let $g(x) = a_nx^n + \ldots a_1x + a_0$ be a polynomial with integer coefficients. Let S be the solutions of g in $\mathbb{Z}/m_1m_2 \ldots m_k\mathbb{Z}$ and S_i the solutions of g in $\mathbb{Z}/m_i\mathbb{Z}$. Then we have a bijection

$$S \leftrightarrow S_1 \times S_2 \times \cdots \times S_k,$$

given by

$$a \pmod{m_1m_2 \ldots m_k} \mapsto (a \pmod{m_1}, a \pmod{m_2}, \ldots, a \pmod{m_k}).$$

Indeed, $g(a) \pmod{m_1m_2 \ldots m_k}$ is mapped to $(g(a) \pmod{m_1}, g(a) \pmod{m_2}, \ldots, g(a) \pmod{m_k})$ and $g(a) \equiv 0 \pmod{m_1m_2 \ldots m_k}$ if and only if for every i we have $g(a) \equiv 0 \pmod{m_i}$. This shows that we have a map

$$S \to S_1 \times S_2 \times \cdots \times S_k.$$

But, conversely, given solutions r_i to $g(x)$ mod m_i, there is a unique $r \pmod{m_1m_2 \cdots m_k}$ such that $r \equiv r_i \pmod{m_i}$ and $g(r) \equiv 0 \pmod{m_1m_2 \cdots m_k}$ because it is true modulo every m_i.

In particular, we may draw the following conclusion.

Corollary 22.3.6. *Let $m_1, \ldots, m_k$ be relatively prime integers. Let s be the number of solutions to the equation $a_nx^n + \cdots + a_1x + a_0 = 0 \pmod{m_1m_2 \cdots m_k}$ and let s_i be the number of solutions modulo m_i. Then*

$$s = s_1s_2 \cdots s_k.$$

Example 22.3.7. The equation $x^2 = 1$ has 8 solutions modulo $2 \cdot 3 \cdot 5 \cdot 7 = 490$, because it has one solution mod 2, and 2 solutions mod 3,5 or 7.

Example 22.3.8. The equation $34x = 85 \pmod{17 \cdot 19}$ has 17 solutions, because it has 17 solutions modulo 17 (it is then the equation $0 \cdot x = 0 \pmod{17}$) and has a unique solution modulo 19 (it is then the equation $4x = 10 \pmod{19}$ and $x = 12$ is the unique solution).

On the other hand, the equation $34x = 5 \pmod{17 \cdot 19}$ has no solutions, because it has no solutions modulo 17.

There remains the question how to calculate a solution mod $m_1m_2 \cdots m_k$ from solutions mod m_1, mod m_2, ..., mod m_k. That is, how to find explicitly the inverse to the map

$$\mathbb{Z}/m_1m_2 \ldots m_k\mathbb{Z} \to \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_k\mathbb{Z}.$$

We explain how to do that for 3 numbers m_1, m_2, m_3, though the method is general.

We first find integers ϵ_1, ϵ_2 such that

$$\epsilon_1 \equiv 1 \pmod{m_1}, \qquad \epsilon_1 \equiv 0 \pmod{m_2m_3},$$

and

$$\epsilon_2 \equiv 0 \pmod{m_1}, \qquad \epsilon_2 \equiv 1 \pmod{m_2m_3}.$$

This we know how to do because we are only dealing with two relatively prime numbers, that is, m_1 and m_2m_3. Then, find λ_2, λ_3 such that

$$\lambda_2 \equiv 1 \pmod{m_2}, \qquad \lambda_2 \equiv 0 \pmod{m_3},$$

and

$$\lambda_3 \equiv 0 \pmod{m_2}, \qquad \lambda_3 \equiv 1 \pmod{m_3}.$$

Then, the numbers

$$\mu_1 = \epsilon_1, \quad \mu_2 = \epsilon_2\lambda_2, \quad \mu_3 = \epsilon_2\lambda_3,$$

are congruent to $(1,0,0),(0,1,0)$ and $(0,0,1)$ respectively in the ring $\mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \mathbb{Z}/m_3\mathbb{Z}$. To find an integer mod $m_1m_2m_3$ mapping to (a,b,c) in $\mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \mathbb{Z}/m_3\mathbb{Z}$ take $a\mu_1 + b\mu_2 + c\mu_3$:

$$a\mu_1 + b\mu_2 + c\mu_3 \mapsto (a,b,c) \in \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \mathbb{Z}/m_3\mathbb{Z}.$$

Let us illustrate all this with a numerical example.

Example 22.3.9. Find the solutions to the equation $x^2 + x + 2 = 0 \pmod{7 \cdot 11 \cdot 23}$.

The solutions modulo 7 are $S_1 = \{3\}$; the solutions modulo 11 are $S_2 = \{4,6\}$; the solutions modulo 23 are $S_3 = \{9,13\}$. (Those are found by brute computation.) To find the corresponding solutions modulo $7 \cdot 11 \cdot 23$ we first find the μ_i above.

First $11 \cdot 23 = 36 \cdot 7 + 1$ and so $\epsilon_1 = 253, \epsilon_2 = -252$ satisfy $\epsilon_1 \equiv 1 \pmod 7, \epsilon_1 \equiv 0 \pmod{253}$, $\epsilon_2 \equiv 0 \pmod 7, \epsilon_2 \equiv 1 \pmod{253}$. To find λ_1, λ_2 we note that $1 = 23 - 2 \cdot 11$ and so $\lambda_1 = 23$, $\lambda_2 = -22$ satisfy $\lambda_1 \equiv 1 \pmod{11}, \lambda_1 \equiv 0 \pmod{23}$ and $\lambda_2 \equiv 0 \pmod{11}, \lambda_2 \equiv 1 \pmod{23}$. The numbers μ_1, μ_2, μ_3 are then $253, -252 \cdot 23, -252 \cdot (-22)$. They should be understood as numbers modulo $7 \cdot 11 \cdot 23 = 1771$ and so we can replace $-252 \cdot 23$ by 1288 (which is congruent to it modulo 1771) and $-252 \cdot (-22)$ by 231. Then to get a number congruent to (a,b,c) in $\mathbb{Z}/7\mathbb{Z} \times \mathbb{Z}/11\mathbb{Z} \times \mathbb{Z}/23\mathbb{Z}$, we take $a \cdot 253 + b \cdot 1288 + c \cdot 231$. In particular, the solution $(3,4,9)$ produces $3 \cdot 253 + 4 \cdot 1288 + 9 \cdot 231 = 7990$ which we can replace by the number 906 which is congruent to it modulo 1771. By our theory 906 is one of the 4 solutions to the equation $x^2 + x + 2 = 0 \pmod{1771}$ and this can be verified.

Do you think you could have found this solution in an easier way?! (I can't see how).

Example 22.3.10. Although the method described to find μ_1, μ_2, μ_3 is perhaps theoretically the quickest, in practice you may find it less confusing to perform the following variant. First, as $1 = am_1 + bm_1m_2$ for some integers a, b, if we let $\mu_1 = bm_1m_2$ then

$$\mu_1 \equiv 1 \pmod{m_1}, \quad \mu_1 \equiv 0 \pmod{m_2}, \quad \mu_1 \equiv 0 \pmod{m_3}.$$

Similarly, write now (for different a, b) $1 = am_2 + bm_1m_3$ and we let $\mu_2 = bm_1m_3$. Then $\mu_2 \equiv 1 \pmod{m_2}$ and $\mu_2 \equiv 0 \pmod{m_1}$, $\mu_2 \equiv 0 \pmod{m_3}$. Finally, write $1 = am_3 + bm_1m_2$ and let $\mu_3 = bm_1m_2$. Then $\mu_3 \equiv 1 \pmod{m_3}$ and $\mu_3 \equiv 0 \pmod{m_1}$, $\mu_3 \equiv 0 \pmod{m_2}$.

Consider the following numerical example. Let us solve the equation $x^2 + x \pmod{60}$. The solutions mod 4 are $\{0,3\}$, mod 3 are $\{0,2\}$ and mod 5 are $\{0,4\}$. Using that $\gcd(4,15) = 1$ we find that $1 = 3 \cdot 15 - 11 \cdot 4$ and so $\mu_1 = 45$. Using that $\gcd(3,20) = 1$ we find that $1 = 2 \cdot 20 - 13 \cdot 3$ and so that $\mu_2 = 40$. Finally, as $\gcd(5,12) = 1$ we find that $1 = 3 \cdot 12 - 7 \cdot 5$ and so $\mu_3 = 36$.

The general solution to $x^2 + x = 0 \pmod{60}$ is thus,

$$45 \cdot \begin{cases} 0 \\ 3 \end{cases} + 40 \cdot \begin{cases} 0 \\ 2 \end{cases} + 36 \cdot \begin{cases} 0 \\ 4 \end{cases} = \{0, 144, 80, 224, 135, 279, 215, 359\} = \{0, 15, 20, 24, 35, 39, 44, 59\}.$$

23. Prime and maximal ideals

Let R be a commutative ring and I an ideal of R.

Definition 23.0.1. The ideal I is called **maximal**, if $I \neq R$ and the only ideals of R containing I are I and R, namely there is no ideal J such that $I \subsetneq J \subsetneq R$. The ideal I is called **prime** if $I \neq R$ and if $ab \in I$ implies $a \in I$ or $b \in I$.

Theorem 23.0.2. *The following hold:*

(1) I is a prime ideal if and only if R/I is an integral domain.
(2) I is a maximal ideal if and only if R/I is a field.
(3) A maximal ideal is a prime ideal.

Proof. (1) Let I be a prime ideal. The quotient ring R/I is a commutative ring. Let $\bar{a},\bar{b} \in R/I$ and suppose that $\bar{a}\cdot\bar{b} = \bar{0}$. This means that $\overline{ab} = \bar{0}$ and so that $ab \in I$. Thus, $a \in I$ or $b \in I$, which implies that $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$. Furthermore, since R/I is not zero, $0_{R/I} \neq 1_{R/I}$ (as a ring in which 0 is equal to 1 is the zero ring). This proves that R/I is an integral domain.

Suppose conversely that R/I is an integral domain and that $ab \in I$. Then $\bar{a}\cdot\bar{b} = \bar{0}$ and this implies $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$. Thus, $a \in I$ or $b \in I$ and so, as also $I \neq R$, I is a prime ideal.

(2) Suppose that I is a maximal ideal. If $1 \in I$ then $I = R$ and that is a contradiction. Thus, $1 \notin I$ and therefore in R/I we have $\bar{1} \neq \bar{0}$. Since R is commutative so is R/I and it remains to prove that every non-zero element $\bar{a}$ of R/I is invertible. Consider the ideal $\langle a, I\rangle = (a) + I$ of R. It strictly contains I because $a \notin I$. Therefore $(a) + I = R$ and in particular, for some $r \in R$ and $i \in I$ we have $ar + i = 1$. In the ring R/I we get $\bar{a}\cdot\bar{r} = \bar{1}$.

Suppose now that R/I is a field and let $\pi\colon R \to R/I$ be the canonical homomorphism. Let $J \supseteq I$ be an ideal. Then $\pi(J)$ is an ideal of R/I (if $f\colon R \to S$ is a ring homomorphism and J an ideal of R then $f(J)$ is an ideal of $\mathrm{Im}(f)$). Since R/I is a field, $\pi(J)$ is either the zero ideal of R/I or R/I. In the first case we have $\pi(J) = \{\bar{0}\}$, which means that every element of J belongs to I and so $J = I$. In the second case we have $\pi(J) = R/I$. Let $r \in R$ then for some $j \in J$ we have $r + I = j + I$ and that means that $r = j + i$ for some $i \in I$. But $J \supseteq I$ and so $r \in J$. It follows that $J = R$.

(3) Since a field is an integral domain, if I is maximal it is prime.

□

Example 23.0.3. Let R be a commutative ring. Two elements f, g of R are called **associate** if $f = g\epsilon$ for some unit ϵ. We write $f \sim g$. Being associate is an equivalence relation.

Suppose that R is a commutative domain. Namely, a non-zero commutative ring with no zero divisors. Consider an ideal I of R of the form (f). When is I a prime ideal? If $I = \{0\}$, namely, if $f = 0$, then I is a prime ideal as $R/I \cong R$. If $f \neq 0$ and I is a prime ideal then $I \neq R$ and so f is also not a unit. It further has the property that $f|ab$ implies $f|a$ or $f|b$. Such a non-zero non-unit element f is called **prime**. It is easy to reverse the argument and we find that if $I = (f)$ is a prime ideal, either I is equal to zero, or f is a prime element and, vice-versa, if f is a prime element $I = (f)$ is a prime ideal.

Now, we may ask when is $I = (f)$ a maximal ideal? The simple answer is that this is the case if $I \neq R$, (namely, if f is not a unit) and whenever $f \nmid g$ (namely, whenever $g \notin (f)$) then $(g, f) = R$. Namely, one can write $1 = ag + bf$ for some $a, b \in R$. This is not very interesting. In fact, it is merely a restatement of the fact that $R/(f)$ is a field.

However, suppose that R is a **principal ideal domain**, namely, a non-zero commutative ring with no zero divisors in which every ideal is principal. One can check that $(f) = (g)$ if and only if $f \sim g$. When is an ideal $I = (f)$ maximal?

If $f = 0$ and I is maximal, then any ideal of the form (r), where $r \neq 0$, must be the whole ring (as this ideal contains I properly); that is, any $r \neq 0$ is a unit and that implies that R is a field. Conversely, if R is a field then $\{0\}$ is a maximal ideal. Now, suppose that $f \neq 0$. Then (f) is also a prime ideal and so f is a prime element. However, because R is a principal ideal domain, we can show that conversely, if f is prime then (f) is not just a prime ideal, but in fact a maximal ideal. Indeed, suppose that $J \supseteq (f)$ is an ideal. Write $J = (g)$, which is possible because R is a principal ideal domain. $(g) \supseteq (f)$ and that means that $f = gb$ for some $b \in R$. Thus, $f|gb$. If $f|g$ then $g \in (f)$ and so $(g) \subseteq (f)$ and therefore $(g) = (f)$ (and $g \sim f$). If $f|b$, as $b|f$, we get that $f \sim b$, which implies that g is a unit. That is, $(g) = R$.

To summarize. In any commutative integral domain, if $I = (f)$ is a prime ideal then either $f = 0$, or f is a prime element. If f is a prime element, or zero, $I = (f)$ is a prime ideal. If R is a principal ideal domain, then any prime ideal is maximal (and always a maximal ideal is prime).

24. Exercises

(1) Recall that for the ring $\mathbb{Z}$ a complete list of ideals is given by (0), (1), (2), (3), (4), $(5),\dots$, where (n) is the principal ideal generated by n, namely, $(n)=\{na : a\in\mathbb{Z}\}$. Find the complete list of ideals of the ring $\mathbb{Z}\times\mathbb{Z}$.

(2) Let R be a ring and let I and J be two ideals of R.
 (a) Prove that $I\cap J$ is an ideal of R, where
$$I\cap J=\{r : r\in I, r\in J\}.$$
 It is called the intersection of the ideals I and J.
 (b) Prove that
$$I+J=\{i+j : i\in I, j\in J\}$$
 is an ideal of R. It is called the sum of the ideals I and J.
 (c) Find for every two ideals of the ring $\mathbb{Z}$ their sum and intersection.

(3) Let $\mathbb{F}$ be a field. Prove that the ring $M_2(\mathbb{F})$ of 2×2 matrices with entries in $\mathbb{F}$ has no non-trivial (two-sided) ideals. That is, every ideal is either the zero ideal or $M_2(\mathbb{F})$ itself.
(Note: there is also a notion of a one-sided ideal that we don't discuss in this course. The ring $M_2(\mathbb{F})$ has a non-trivial one sided ideal. The notion of one-sided ideals is studied in Higher Algebra I & II).

(4) **The ring of real quaternions** $\mathbb{H}$ (Hamilton's quaternions). Let i,j,k be formal symbols and
$$\mathbb{H}=\{a+bi+cj+dk : a,b,c,d\in\mathbb{R}\}.$$
 Addition on $\mathbb{H}$ is defined by
$$(a+bi+cj+dk)+(a'+b'i+c'j+d'k)=(a+a')+(b+b')i+(c+c')j+(d+d')k.$$
 Multiplication is determined by defining
$$i^2=j^2=-1,\quad ij=-ji=k,$$
 (and one extends this to a product rule by linearity).
 (a) Prove that the map
$$\mathbb{H}\to\left\{\begin{pmatrix} z_1 & z_2\\ -\overline{z_2} & \overline{z_1}\end{pmatrix} : z_1,z_2\in\mathbb{C}\right\},$$
 taking $a+bi+cj+dk$ to the matrix $\begin{pmatrix} a+bi & c+di\\ -c+di & a-bi\end{pmatrix}$ is bijective and satisfies: $f(x+y)=f(x)+f(y)$ and $f(i)^2=f(j)^2=-I_2, f(i)f(j)=f(k)=-f(j)f(i)$.
 (b) Use part (a) to conclude that $\mathbb{H}$ is indeed a ring, by proving it is a subring of $M_2(\mathbb{C})$.
 (c) Prove that $\mathbb{H}$ is a non-commutative division ring.

(5) In the following questions it's useful to remember that under our definitions a ring homomorphism takes 1 to 1.
 (a) Prove that there is no ring homomorphism $\mathbb{Z}/5\mathbb{Z}\to\mathbb{Z}$.
 (b) Prove that there is no ring homomorphism $\mathbb{Z}/5\mathbb{Z}\to\mathbb{Z}/7\mathbb{Z}$.
 (c) Prove that the rings $\mathbb{Z}/2\mathbb{Z}\times\mathbb{Z}/2\mathbb{Z}$ and $\mathbb{Z}/4\mathbb{Z}$ are not isomorphic.
 (d) Is there a ring homomorphism $\mathbb{Z}/4\mathbb{Z}\to\mathbb{Z}/2\mathbb{Z}\times\mathbb{Z}/2\mathbb{Z}$?
 (e) Is there a ring homomorphism $\mathbb{Z}/2\mathbb{Z}\times\mathbb{Z}/2\mathbb{Z}\to\mathbb{Z}/4\mathbb{Z}$?

(6) (a) Let R be a commutative ring and let $r_1,\dots,r_n$ be elements of R. We define $(r_1,\dots,r_n)$ to be the set
$$\{r_1a_1+\cdots+r_na_n : \forall i\ a_i\in R\}.$$
 Prove that $(r_1,\dots,r_n)$ is an ideal of R. We call it the ideal generated by $r_1,\dots,r_n$.

(b) Now apply that to the case where $R = \mathbb{Z}[x]$ (polynomials with integer coefficients). Let $(2, x)$ be the ideal generated by 2 and x.
(i) Prove that the ideal $(2, x)$ is not principal and conclude that $\mathbb{Z}[x]$ is not a principal ideal ring.
(ii) Find a homomorphism $f\colon R \to \mathbb{Z}/2\mathbb{Z}$ such that $(2, x) = \mathrm{Ker}(f)$.

(7) Prove that $\mathbb{C}[x, y]$ is not a principal ideal ring, for example, that the ideal (x, y) is not a principal ideal.

(8) Prove that no two of the following rings are isomorphic:
(a) $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (with addition and multiplication given coordinate by coordinate);
(b) $M_2(\mathbb{R})$;
(c) The ring $\mathbb{H}$ of real quaternions.

(9) Let $f\colon R \to S$ be a ring homomorphism.
(a) Let $J \lhd S$ be an ideal. Prove that $f^{-1}(J)$ (equal by definition to $\{r \in R : f(r) \in J\}$) is an ideal of R.
(b) Prove that if f is surjective and $I \lhd R$ is an ideal then $f(I)$ is an ideal (where $f(I) = \{f(i) : i \in I\}$).
(c) Show, by example, that if f is not surjective the assertion in (2) need not hold.

(10) Let $\mathbb{F}$ be a field and let

$$R = \left\{ \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} : a_{ij} \in \mathbb{F} \right\}.$$

Let

$$I = \{(a_{ij}) \in R : a_{11} = a_{22} = a_{33} = 0\}.$$

Prove that R is a subring of $M_3(\mathbb{F})$, I is an ideal of R and $R/I \cong \mathbb{F} \times \mathbb{F} \times \mathbb{F}$.

(11) Let d be an integer, which is not a square of another integer.
(a) Prove that d is not a square of a rational number.
(b) Let $\mathbb{Q}[\sqrt{d}] := \{a + b\sqrt{d} : a, b \in \mathbb{Q}\}$. Show that $\mathbb{Q}[\sqrt{d}]$ is a subring of $\mathbb{C}$ and is in fact a field.
(c) Prove that $\mathbb{Q}[\sqrt{d}] \cong \mathbb{Q}[x]/(x^2 - d)$.
(d) Prove that the fields $\mathbb{Q}[\sqrt{2}]$ and $\mathbb{Q}[\sqrt{3}]$ are not isomorphic.

(12) Prove a Chinese Remainder Theorem for polynomials:
Let $\mathbb{F}$ be a field and let $f(x), g(x)$ be two non-constant polynomials that are relatively prime, $\gcd(f, g) = 1$. Prove that

$$\mathbb{F}[x]/(fg) \cong \mathbb{F}[x]/(f) \times \mathbb{F}[x]/(g).$$

(Hint: mimic the proof of the Chinese Remainder Theorem for integers.)

(13) Let R and S be rings and let $I \lhd R$, $J \lhd S$ be ideals. Prove that

$$(R \times S)/(I \times J) \cong (R/I) \times (S/J).$$

(14) For each of the rings $\mathbb{Z}/60\mathbb{Z}$ and $\mathbb{F}[x]/(x^4 + 2x^3 + x^2)$ find all their ideals and identify all their homomorphic images. Suggestion: To find the ideals use exercise (9) to reduce the calculation to ideals of the ring $\mathbb{Z}$ or $\mathbb{F}[x]$ that contain the ideal (60), respectively $(x^4 + 2x^3 + x^2)$, but explain why this is valid.

(15) Using the Chinese remainder theorem, find the solutions (if any) to the following polynomial equations (for example, in (d), write the solutions as integers mod 30 and so on):
(a) $15x = 11$ in $\mathbb{Z}/18\mathbb{Z}$.
(b) $15x = 12$ in $\mathbb{Z}/63\mathbb{Z}$.
(c) $x^2 = 37$ in $\mathbb{Z}/63\mathbb{Z}$.
(d) $x^2 = 4$ in $\mathbb{Z}/30\mathbb{Z}$.

(16) Prove that $\mathbb{F}_5[x]/(x^2+2)$ is a field. Calculate the following expressions as polynomials of degree smaller than 2: $(x^2+x)*(x^3+1)-(x+1)$, $(x^2+3)^{-1}$ and $(x^2-1)/(x^2+3)$. Find **all** the roots of the polynomials t^2+2 and t^2+3 in this field.

(17) Prove that $\mathbb{F}_2[x]/(x^3+x+1)$ is a field. Calculate the following expressions as polynomials of degree smaller than 3: $(x^2+x)*(x^3+1)-(x+1)$, $(x^2+3)^{-1}$ and $(x^2-1)/(x^2+3)$. Find **all** the roots of the polynomial t^3+t^2+1 and t^3+1 in this field.

(18) Let $f(x)=x^2-2\in\mathbb{F}_{19}[x]$.
   (a) Prove that f is irreducible and deduce that $L:=\mathbb{F}_{19}[x]/(x^2-2)$ is a field.
   (b) Find the roots of the polynomial t^2-t+2 in the field L.
   (c) Prove that $h(x)=x^3-2$ is irreducible over $\mathbb{F}_{19}$. Prove that it is also irreducible in L.
   (d) Find $x^{19}-x$ in $\mathbb{F}_{19}[x]/(x^3-2)$ as being represented by a polynomial of degree at most 2 (hint: $x^{19}=(x^3)^6\cdot x$). Use this to rapidly calculate $\gcd(x^{19}-x,h(x))$ and conclude also in this way that $h(x)$ is irreducible over $\mathbb{F}_{19}$.

(19) Find all the solutions to the equation $x^3=38 \pmod{195}$.

Part 6. **Groups**

25. First definitions and examples

25.1. **Definitions and some formal consequences.**

Definition 25.1.1. A **group** G is a non-empty set with an operation

$$G \times G \to G, \quad (a,b) \mapsto ab,$$

such that the following axioms hold:

(1) $(ab)c = a(bc)$. (Associativity)
(2) There exists an element $e \in G$ such that $eg = ge$ for all $g \in G$. (Identity)
(3) For every $g \in G$ there exists an element $d \in G$ such that $dg = gd = e$. (Inverse)

Other common notations for e are 1 or 1_G.

Here are some formal consequences of the definition:

(1) $\boxed{e \textbf{ is unique}.}$ Say $\tilde{e}$ has the same property then $\tilde{e} = e\tilde{e}$, using the property of e, but also $e\tilde{e} = e$, using the property of $\tilde{e}$. Thus, $e = \tilde{e}$.
(2) $\boxed{d \textbf{ appearing in (3) is unique}}$ (therefore we shall call it "the inverse of g" and denote it by g^{-1}). Say $\tilde{d}$ also satisfies $\tilde{d}g = g\tilde{d} = e$. Then
$$\tilde{d} = \tilde{d}e = \tilde{d}(gd) = (\tilde{d}g)d = ed = d.$$
(3) $\boxed{\textbf{Cancelation: } ab = cb \Rightarrow a = c, \text{ and } ba = bc \Rightarrow a = c.}$ If $ab = cb$ then $(ab)b^{-1} = (cb)b^{-1}$ and so $a = a(bb^{-1}) = c(bb^{-1}) = c$.
(4) $\boxed{(ab)^{-1} = b^{-1}a^{-1}.}$ To show that we need to show that $b^{-1}a^{-1}$ "functions as the inverse of ab". We have $(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aea^{-1} = aa^{-1} = e$. Similarly, $(b^{-1}a^{-1})(ab) = b^{-1}(a^{-1}a)b = b^{-1}eb = b^{-1}b = e$.
(5) $\boxed{(a^{-1})^{-1} = a.}$ This is because $aa^{-1} = a^{-1}a = e$ also shows that a is the inverse of a^{-1}.
(6) Define $a^0 = e$, $a^n = a^{n-1}a$ for $n > 0$ and $a^n = (a^{-1})^{-n}$ for $n < 0$. Then we have
$$\boxed{a^m a^n = a^{m+n}, \quad (a^m)^n = a^{mn}.}$$

25.2. **Examples.**

Example 25.2.1. The **trivial group** G is a group with one element e and multiplication law $ee = e$.

Example 25.2.2. If R is a ring, then R with addition only is a group. It is a commutative group. The operation in this case is of course written $g + h$. In general a group is called **commutative** or **abelian** if for all $g, h \in G$ we have $gh = hg$. It is customary in such cases to write the operation in the group as $g + h$ and not as gh, but this is not a must. This example thus includes $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{F}, \mathbb{F}[\epsilon], M_2(\mathbb{F}), \mathbb{Z}/n\mathbb{Z}$, all with the addition operation.

Example 25.2.3. Let R be a ring. Recall that the **units** $R^\times$ of R are defined as

$$\{u \in R : \exists v \in R, uv = vu = 1\}.$$

This is a group. If $u_1, u_2 \in R$ with inverses v_1, v_2, respectively, then, as above, one checks that v_2v_1 is an inverse for u_1u_2 and so $R^\times$ is closed under the product operation. The associative law holds because it holds in R; 1_R serves as the identity. If R is not commutative there is no reason for $R^\times$ to be commutative, though in certain cases it may be.

Thus we get the examples of $\mathbb{Z}^\times = \{\pm 1\}, \mathbb{Q}^\times = \mathbb{Q} - \{0\}, \mathbb{R}^\times = \mathbb{R} - 0, \mathbb{C}^\times = \mathbb{C} - \{0\}$, and more generally, for a field $\mathbb{F}$, $\mathbb{F}^\times = \mathbb{F} - \{0\}$. We also have, $\mathrm{GL}_2(\mathbb{F}) = \{M \in M_2(\mathbb{F}) : \det(M) \neq 0\}, \mathbb{F}[\epsilon]^\times = \{a + b\epsilon : a \neq 0\}$.

Proposition 25.2.4. *Let $n > 1$ be an integer. The group $\mathbb{Z}/n\mathbb{Z}^\times$ (meaning, $(\mathbb{Z}/n\mathbb{Z})^\times$, the units of the ring $\mathbb{Z}/n\mathbb{Z}$) is precisely*

$$\{1 \leq a \leq n : (a,n) = 1\}.$$

Proof. If $\bar{a}$ is invertible then $ab = 1 \pmod{n}$ for some integer b; say $ab = 1 + kn$ for some $k \in \mathbb{Z}$. If $d|a, d|n$ then $d|1$. Therefore $(a,n) = 1$.

Conversely, suppose that $(a,n) = 1$ then for some u, v we have $1 = ua + vn$ and so $ua = 1 \pmod{n}$. □

One defines **Euler's φ function** on positive integers by

$$\varphi(n) = \begin{cases} 1 & n = 1 \\ |\mathbb{Z}/n\mathbb{Z}^\times| & n > 1. \end{cases}$$

One can prove that this is a **multiplicative function**, namely, if $(n,m) = 1$ then $\varphi(nm) = \varphi(n)\varphi(m)$. I invite you to try and prove this based on the Chinese Remainder Theorem.

Here are some examples:

n	$\mathbb{Z}/n\mathbb{Z}^\times$	$\varphi(n)$
2	$\{1\}$	1
3	$\{1,2\}$	2
4	$\{1,3\}$	2
5	$\{1,2,3,4\}$	4
6	$\{1,5\}$	2
7	$\{1,2,3,4,5,6\}$	6
8	$\{1,3,5,7\}$	4
9	$\{1,2,4,5,7,8\}$	6

Example 25.2.5. If G, H are groups then $G \times H$ is a group with the operation

$$(g_1, h_1)(g_2, h_2) = (g_1 g_2, h_1 h_2).$$

The identity is (e_G, e_H) and $(g,h)^{-1} = (g^{-1}, h^{-1})$.

Example 25.2.6. Consider a solid object and all the rotations that take it to itself. For example, if the object is a cube we may rotate it ninety degrees relative to an axis that goes from the middle point of a face to the middle point of the opposite face. Each such rotation, or succession of rotations, is thought of as an element of the group of symmetries of the solid.

With some effort one can calculate these groups. For example, for a tetrahedron we get a group of symmetries with 12 elements. For a cube we get a group of symmetries with 24 elements.

25.3. **Subgroups.** Let G be a group. A subset $H \subseteq G$ is called a **subgroup** if the following holds:

(1) $e_G \in H$;
(2) $a, b \in H \Rightarrow ab \in H$;
(3) $a \in H \Rightarrow a^{-1} \in H$.

Clearly then H is a group in its own right.

Example 25.3.1. The subset S^1 of $\mathbb{C}^\times$ consisting of all complex numbers of absolute value 1 is a subgroup. Indeed $1 \in S^1$. If $s_1, s_2 \in S^1$ then $|s_1 s_2| = |s_1|\ |s_2| = 1$ so $s_1 s_2 \in S^1$. If z is any non-zero complex number then $1 = |1| = |zz^{-1}| = |z| \cdot |z^{-1}|$ and so $|z^{-1}| = 1/|z|$. If $z \in S^1$ it therefore follows that $z^{-1} \in S^1$.

Let $n \geq 1$ be an integer. The subset μ_n of $\mathbb{C}^\times$ consisting of all complex numbers x such that $x^n = 1$ is a subgroup of $\mathbb{C}^\times$, and in fact of S^1, having n elements. It is called the **n-th roots of unity**. The proof is left as an exercise.

Example 25.3.2. Let $\mathbb{F}$ be a field and

$$H = \left\{ \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} : a \in \mathbb{F} \right\}.$$

Then H is a subgroup of $\mathrm{GL}_2(\mathbb{F})$.

Definition 25.3.3. Let G be a group. G is called **cyclic** if there is an element $g \in G$ such that any element of G is a power of g; that is, $G = \{g^n : n \in \mathbb{Z}\}$. The element g is then called a **generator** of G.

Example 25.3.4. Let G be any group. Let $g \in G$ and define

$$\langle g \rangle := \{g^n : n \in \mathbb{Z}\}.$$

This is a cyclic subgroup of G (it may be finite or infinite).

Example 25.3.5. The group $\mathbb{Z}$ is cyclic. As a generator we may take 1 (or -1).

Example 25.3.6. The group $(\mathbb{Z}/5\mathbb{Z})^\times = \{1,2,3,4\}$ is cyclic. The elements $2,3$ are generators. The group $\mathbb{Z}/8\mathbb{Z}^\times$ is not cyclic. One can check that the square of any element is 1.

26. Permutation groups and dihedral groups

26.1. **Permutation groups.**

Definition 26.1.1. A **permutation** of a set T is a bijective function $f\colon T \to T$. We shall denote the set of permutations of T by S_T (or Σ_T). If $T = \{1,2,\cdots,n\}$ then we shall denote S_T as S_n. It is called "the symmetric group on n letters".

Proposition 26.1.2. *For every non-empty set T, S_T is a group under composition of functions. The cardinality of S_n is $n!$.*

Proof. The product of two permutations f,g is their composition $f \circ g$; it is again a permutation. We have $[(f \circ g) \circ h](t) = (f \circ g)(h(t)) = f(g(h(t))) = f((g \circ h)(t)) = [f \circ (g \circ h)](t)$. Thus, as functions, we have $(f \circ g) \circ h = f \circ (g \circ h)$ and so the operation of composition of functions is associative.

The identity is just the identity function. The inverse of a permutation f is the inverse function f^{-1}, which satisfies $f \circ f^{-1} = f^{-1} \circ f = \mathrm{Id}_T$.

Finally, to define a permutation f on $\{1,2,\ldots,n\}$ we can choose the image of 1 arbitrarily (n choices), the image of 2 could be any element different from $f(1)$ ($n-1$ choices), the image of 3 can be any elements different from the images of 1 and 2 ($n-1$ choices), and so on. Altogether, we have $n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1 = n!$ choices. □

Example 26.1.3. (1) For $n = 1$, S_1 consist of a single element and so is the trivial group.

(2) For $n = 2$ we have two permutations.
   (a) Id. $\mathrm{Id}(1) = 1, \mathrm{Id}(2) = 2$.
   (b) σ. $\sigma(1) = 2, \sigma(2) = 1$.

   We may also represent these permutations in the form of tables:

$$\mathrm{Id} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \qquad \sigma = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

(3) For $n = 3$ we have 6 permutations. One of them is σ given by $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1$, or in table form

$$\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

The table form is a better notation and we list all elements of S_3 in that form.

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

The permutations in first line form a cyclic subgroup of S_3. It is the subgroup generated by $\begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}$.

(4) In general, let T_0 be a subset of T. We can define two subgroups of Σ_T. The first is $H = \{\sigma \in \Sigma_T : \sigma(t) = t, \forall t \in T_0\}$ and the second is $I = \{\sigma \in \Sigma_T : \sigma(T_0) = T_0\}$. Then $H \subseteq I \subseteq \Sigma_T$ are subgroups. H and I depend on the choice of T_0, but we do not reflect this in the notation. If $\Sigma_T = S_n$ and T_0 has m elements then H has $(n-m)!$ elements and I has $(n-m)!m!$ elements.

The groups S_n are not commutative for $n \geq 3$. For example:

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

Here is another example of multiplication, in S_5 this time:

$$\begin{pmatrix}1&2&3&4&5\\3&2&1&4&5\end{pmatrix}\begin{pmatrix}1&2&3&4&5\\5&2&4&3&1\end{pmatrix} = \begin{pmatrix}1&2&3&4&5\\5&2&4&1&3\end{pmatrix}.$$

26.2. **Cycles.** There is still more efficient notation for permutations in S_n. Fix $n \geq 1$. A **cycle** (in S_n) is an expression of the form

$$(a_1\ a_2 \cdots a_t),$$

where $a_i \in \{1, 2, \ldots, n\}$ are distinct elements. This expression is understood as the permutation σ given by

$$\sigma(a) = \begin{cases} a_{i+1} & a = a_i,\ i < n, \\ a_1 & a = a_n, \\ a & \text{else.} \end{cases}$$

Pictorially:

(and elements outside the cycle do not move).

For example, the permutation $\begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}$ is the cycle $(1\ 2\ 3)$ and the permutation $\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}$ is the cycle $(1\ 2)$. A cycle with two elements $(i\ j)$ (not necessarily consecutive) is called a **transposition**.

Definition 26.2.1. Let G be a group. The **order** of G, denoted $|G|$, or $\sharp G$, is the number of elements of G (written ∞ if not finite).

Let $g \in G$. The **order** of g is defined as $|\langle g \rangle|$, the order of the cyclic group generated by g. It is also denoted by $o(g)$, or $\mathrm{ord}(g)$.

Lemma 26.2.2. *Let $g \in G$, $o(g)$ is the minimal positive integer k such that $g^k = e$.*

Proof. Let k be the minimal integer such that $g^k = e$ (∞ if such doesn't exist).

Suppose first that $o(g)$ is finite, say equals r. Then the $r+1$ elements $\{e, g, g^2, \ldots, g^r\}$ cannot be distinct and so $g^i = g^j$ for some $0 \leq i < j \leq r$. It follows that $g^{j-i} = e$ and so $k \leq j - i \leq r$. In particular k is also finite. So r is finite implies k is finite and $k \leq r$.

Suppose now that k is finite. Let n be an integer and write $n = ak + b$ where $0 \leq b < k$. Then $g^n = (g^k)^a g^b = e^a g^b = g^b$. We conclude that $\langle g \rangle \subseteq \{e, g, \cdots, g^{k-1}\}$, and so k is finite implies that r is finite and $r \leq k$. □

Example 26.2.3. Let $(a_1\ a_2 \cdots a_t)$ be a cycle. Its order is t.

Two cycles σ, τ are called **disjoint** if they contain no common elements. In this case, clearly $\sigma\tau = \tau\sigma$. Moreover, $(\sigma\tau)^n = \sigma^n\tau^n$ and since σ^n and τ^n are disjoint, $(\sigma\tau)^n = \mathrm{Id}$ if and only if $\sigma^n = \mathrm{Id}$ and $\tau^n = \mathrm{Id}$. Thus $o(\sigma)|n, o(\tau)|n$ and we deduce that $o(\sigma\tau)$ (namely, the least n such that $(\sigma\tau)^n = \mathrm{Id}$) is $\mathrm{lcm}(o(\sigma), o(\tau))$. Arguing in the same way a little more generally we obtain:

Lemma 26.2.4. *Let $\sigma_1, \ldots, \sigma_n$ be disjoint permutations of orders $r_1, \ldots, r_n$, respectively. Then the order of the permutation $\sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_n$ is $lcm(r_1, r_2, \ldots, r_n)$.*

Combining this lemma with the following proposition allows us to calculate the order of every permutation very quickly.

Proposition 26.2.5. *Every permutation is a product of disjoint cycles.*

We shall not provide formal proof of this proposition, but illustrate it by examples.

Example 26.2.6. Consider the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix}$. To write it as a product of cycles we begin by $(1$ and check where 1 goes to. It goes to 3. So we write $(13$ and check where 3 goes to. It goes to 1 and so we have (13). The first number we didn't consider is 2. 2 goes to 4 and so we write $(13)(24$ and 4 goes to 5 and so we write $(13)(245$. Now, 5 goes to 2 and so we have $\sigma = (13)(245)$. The order of σ is $\mathrm{lcm}(2,3) = 6$.

Example 26.2.7. Consider the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 4 & 2 & 1 & 10 & 6 & 9 & 5 & 7 & 8 \end{pmatrix}$. It is written as a product of disjoint transposition as follows $(1324)(5\ 10\ 8)(79)$. To find this expression, we did the same procedure described above: We start with $(1$, continue with $(13$, because 1 goes to 3, and then with $(132$, because 3 goes to 2. Then we find that 2 goes to 4 which goes to 1 and we have found (1324). The first number not in this list is 5 which goes to 10 and so we have $(1324)(5\ 10$. Since 10 goes to 8 and 8 to 5 we get now $(1324)(5\ 10\ 8)$. The first number not in this list is 6 that goes to 6 and that gives $(1324)(5\ 10\ 8)(6)$. We then continue with 7. Since 7 goes to 9 which goes to 7 we have $(1324)(5\ 10\ 8)(6)(79)$. We have considered all numbers and so $\sigma = (1324)(5\ 10\ 8)(6)(79) = (1324)(5\ 10\ 8)(79)$. The order of σ is $\mathrm{lcm}(4,3,2) = 12$.

Example 26.2.8. Suppose we want to find a permutation of order 10 in S_7. We simply take $(12345)(67)$. If we want to find a permutation of order 10 in S_{10} we can take either $(12345)(67)$ or $(123456789\,10)$ (and all variants on this).

Finally we remark on the computation of σ^{-1} for a permutation σ. If σ is given in the form of a table, for example:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 4 & 2 & 1 & 10 & 6 & 9 & 5 & 7 & 8 \end{pmatrix},$$

then because $\sigma(i) = j \Leftrightarrow \sigma^{-1}(j) = i$, the table describing σ^{-1} is the same table but read from the bottom to the top. That is

$$\sigma^{-1} = \begin{pmatrix} 3 & 4 & 2 & 1 & 10 & 6 & 9 & 5 & 7 & 8 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}.$$

Only that we follow our convention and write the columns in the conventional order and so we get

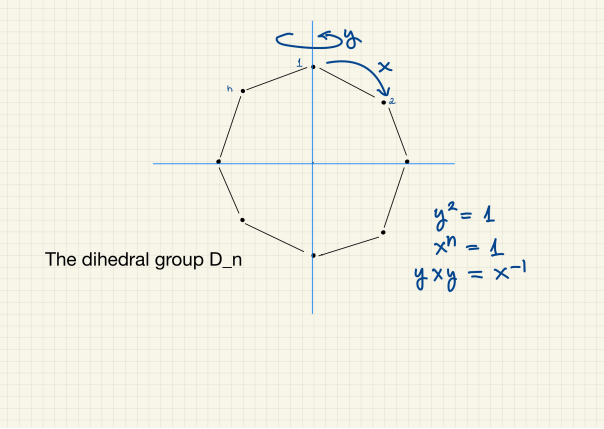
$$\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 4 & 3 & 1 & 2 & 8 & 6 & 9 & 10 & 7 & 5 \end{pmatrix}.$$

If σ is a cycle, say $\sigma = (i_1 i_2 \ldots i_{k-1} i_k)$, then σ^{-1} is easily seen to be $(i_k i_{k-1} \ldots i_2 i_1)$. So,

$$(i_1 i_2 \ldots i_{k-1} i_k)^{-1} = (i_k i_{k-1} \ldots i_2 i_1).$$

Now, if σ is a product of disjoint cycles, $\sigma = \sigma_1\sigma_2\dots\sigma_r$ then $\sigma^{-1} = \sigma_r^{-1}\dots\sigma_2^{-1}\sigma_1^{-1}$ (by a generalization of the rule $(ab)^{-1} = b^{-1}a^{-1}$), but, since those cycles are disjoint they commute, and so we can also write this as $\sigma^{-1} = \sigma_1^{-1}\sigma_2^{-1}\dots\sigma_r^{-1}$. (This last manipulation is wrong if the cycles are not disjoint!) Thus, for example, the inverse of $\sigma = (1324)(5\ 10\ 8)(79)$ is $(4231)(8\ 10\ 5)(97)$, which we can also write, if we wish, as $(1423)(5\ 8\ 10)(79)$.

26.3. **The Dihedral group.** Consider a regular polygon with n sides, $n \geq 3$, in the plane, symmetric around the origin $(0,0)$. The **dihedral group** D_n is defined as the symmetries of the polygon. Let us number the vertices of the polygon by $1,2,\dots,n$ in the clockwise direction and say the first vertex 1 lies on the y-axis. One sees that every symmetry must permutes the vertices and in fact either maintains or reverses their order (we allow flip-over of the polygon). In fact, if σ is a symmetry then $\sigma(1) = j$ and $\sigma(2) = j+1$ or $j-1$ (where we understand $n+1$ as 1 and $1-1$ as n) and σ is uniquely determined by these conditions.



For example, the permutation x given by the cycle $(1\ 2\ 3\ \cdots\ n)$ is an element of the dihedral group rotating the polygon by angle $360^\circ/n$ in the clockwise direction (so if t is a point on the boundary of the polygon such that the line from $(0,0)$ to t forms an angle θ with the x-axis, then $y(t)$ is the point forming an angle $\theta - 360^\circ/n$).

Another symmetry, y, is reflection through the y-axis. The symmetry y is given as permutation by the product $(2\ n),(3\ n-1)\cdots(n/2\ \ 2+n/2)$ if n is even and $(2\ n),(3\ \ n-1)\cdots((n+1)/2\ 1+(n+1)/2)$ if n is odd. In terms of angles, y changes an angle θ to $180^\circ - \theta$.

Theorem 26.3.1. *The elements of the dihedral group are*

$$D_n = \{e, x, \dots, x^{n-1}, y, xy, x^2y, \dots, x^{n-1}y\},$$

and the relations $x^n = y^2 = 1$ and $xyxy = 1$ hold. In particular, D_n has $2n$ elements.

Proof. It is enough to show that any vertex $j \in \{1,2,\dots,n\}$ there is a unique element of the set

$$\{e, x, \dots, x^{n-1}, y, xy, x^2y, \dots, x^{n-1}y\}$$

that takes 1 to j and 2 to $j+1$ and there is a unique element taking 1 to j and 2 to $j-1$. This shows both that every element of D_n is in the list and that all elements of the list are different.

We calculate that

$$x^a(1) = a+1, \quad x^a(2) = a+2,$$

and

$$x^ay(1) = a+1, \quad x^ay(2) = x^a(n) = a.$$

This proves our claims.

The relations $x^n = y^2 = 1$ are evident. We check that $xyxy = 1$, by checking that $xyxy(j) = j$ for $j = 1,2$. We have $xyxy(1) = xyx(1) = xy(2) = x(n) = 1$ and $xyxy(2) = xyx(n) = xy(1) = x(1) = 2$. □

The nature of the symmetries $1, x, \dots, x^{n-1}$ is clear: x^j rotates clockwise by angle $j \cdot 360^\circ/n$.

Proposition 26.3.2. *Let $0 \leq j < n$. The element x^jy is a reflection through the line forming an angle $90^\circ - j \cdot 360^\circ/2n$ with the x-axis.*

Proof. The symmetry x^jy is not trivial. If it fixes an angle θ it must be the reflection through the line with that angle. Note that x^jy first sends the angle θ to $180^\circ - \theta$ and then adds $-j \cdot 360^\circ/n$ so the equation is $\theta = 180^\circ - \theta - j \cdot 360^\circ/n \pmod{360}$. That is $\theta = 90^\circ - j \cdot 360^\circ/2n$. □

Example 26.3.3. We revisit Example 25.2.6. The technique is rather similar to studying the dihedral group. The symmetries of solids we considered there preserve orientation (we don't allow flip-overs through a fourth dimension). If we number the vertices of a tetrahedron by $1,2,3,4$ and the vertices of the cube by $1,2,\ldots,8$, we may view the group of symmetries, call them A and B, respectively, as subgroups of S_4 and S_8, respectively.

If we fix an edge on the tetrahedron, or the cube, then a symmetry is completely determined by its effect on this edge. The edge, as a moment's reflection shows, can go to any other edge on the solid and in two ways. Since a tetrahedron has 6 edges and a cube has 12, we conclude that the group A has 12 elements and the group B has 24. If we wish we can enumerate the elements of the groups A, B, by listing the permutations they induce.

27. The theorem of Lagrange

27.1. **Cosets.** Let $H < G$ be a subgroup of G. A **left coset** of H in G is a subset of the form

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

for some $g \in G$. The set gH is called **the left coset of** g; g is called a **representative** of the coset gH.

Example 27.1.1. Consider the subgroup H of S_3 given by $\{1,(123),(132)\}$. Here are some cosets: $H = 1H = (123)H = (132)H$, $(12)H = (13)H = (23)H = \{(12),(23),(13)\}$. We leave the verification to the reader.

Lemma 27.1.2. *Let H be a subgroup of G.*

(1) Two left cosets are either equal or disjoint.
(2) Let g_1H, g_2H be two left cosets. The following are equivalent: (i) $g_1H = g_2H$; (ii) $g_1 \in g_2H$; (iii) $g_2^{-1}g_1 \in H$.

Proof. Suppose that $g_1H \cap g_2H \neq \emptyset$, so for some h_1, h_2 we have $g_1h_1 = g_2h_2$. We prove that $g_1H \subseteq g_2H$. By symmetry we also have $g_2H \subseteq g_1H$ and so $g_1H = g_2H$.

Let $h \in H$. Then $g_1h = ((g_2h_2)h_1^{-1})h = g_2(h_2h_1^{-1}h) \in g_2H$.

We now prove the equivalence of the assertions (i) - (iii). Suppose (i) holds. Then $g_1 = g_1e \in g_2H$ and (ii) holds. Suppose (ii) holds; say $g_1 = g_2h$. Then $g_2^{-1}g_1 = h \in H$ and (iii) holds. Suppose that (iii) holds; $g_2^{-1}g_1 = h$ for some $h \in H$. Then $g_1 = g_2h$ and so $g_1H \cap g_2H \neq \emptyset$. By what we have proved in the first part, $g_1H = g_2H$. □

Remark 27.1.3. The Lemma and its proof should be compared with Lemma 21.0.3. In fact, since R is an abelian group and an ideal I is a subgroup, that lemma is special case of the lemma above.

Corollary 27.1.4. *G is a disjoint union of cosets of H. Let $\{g_i : i \in I\}$ be a set of elements of G such that each coset has the form g_iH for a unique g_i. That is, $G = \coprod_{i\in I} g_iH$. Then the $\{g_i : i \in I\}$ are called a complete set of representatives.*

In the same manner one defines a **right coset** of H in G to be a subset of the form $Hg = \{hg : h \in H\}$ and Lemma 27.1.2 holds for right cosets with the obvious modifications. Two right cosets are either equal or disjoint and the following are equivalent: (i) $Hg_1 = Hg_2$; (ii) $g_1 \in Hg_2$; (iii) $g_1g_2^{-1} \in H$. Thus, the Corollary holds true for right cosets as well.

We remark that the intersection of a left coset and a right coset may be non-empty, yet not a coset itself. For example, take $H = \{1,(12)\}$ in S_3. We have the following table.

g	gH	Hg
1	$\{1,(12)\}$	$\{1,(12)\}$
(12)	$\{(12),1\}$	$\{(12),1\}$
(13)	$\{(13),(123)\}$	$\{(13),(132)\}$
(23)	$\{(23),(132)\}$	$\{(23),(123)\}$
(123)	$\{(123),(13)\}$	$\{(123),(23)\}$
(132)	$\{(132),(23)\}$	$\{(132),(13)\}$

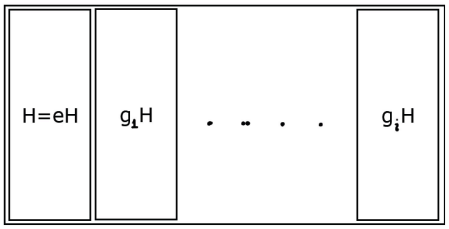
The table demonstrates that indeed any two left (resp. right) cosets are either equal or disjoint, but the intersection of a left coset with a right coset may be non-empty and properly contained in both.

27.2. **Lagrange's theorem.**

Theorem 27.2.1. *Let G be a finite group and H a subgroup of G. Then,*

$$|H| \text{ divides } |G|.$$

Moreover, let $\{g_i : i \in I\}$ be a complete set of representatives for the cosets of H, then $|I| = \frac{|G|}{|H|}$. In particular, the cardinality of I does not depend on the choice of a complete set of representatives. It is called the **index** *of H in G.*



Proof. We have,

$$G = \coprod_{i\in I} g_i H.$$

Let $a, b \in G$. We claim that the function

$$f : aH \to bH, \quad x \mapsto ba^{-1}x,$$

is a well defined bijection. First, $x = ah$ for some h and so $ba^{-1}x = bh \in bH$ and so the map is well defined. It is surjective, because given an element $y \in bH$, say $y = bh$ it is the image of ah. The map is also injective: if $ba^{-1}x_1 = ba^{-1}x_2$ then multiplying both sides by ab^{-1} we get $x_1 = x_2$.

We conclude that each coset g_iH has the same number of elements, which is exactly the number of elements in $H = eH$. We get therefore that

$$|G| = |H| \cdot |I|.$$

This completes the proof. □

Here are some applications of Lagrange's theorem:

(1) Let G be a finite group of prime order p. Then G is cyclic; in fact, every element of G that is not the identity generates G.

Indeed, let $g \neq e$. Then $H = \langle g \rangle$ is a non-trivial subgroup. So $|H| > 1$ and divides p. It follows that $|H| = |G|$ and so that $\langle g \rangle = G$.

(2) In a similar vein, we conclude that a group of order 6 say, cannot have elements of order 4, or 5, or of any order not dividing 6. This follows immediately from Lagrange's theorem, keeping in mind that $\mathrm{ord}(g) = |\langle g \rangle|$.

27.3. **Orders of elements of $\mathbb{Z}/n\mathbb{Z}$.** Let $n > 1$ be an integer and consider the group $\mathbb{Z}/n\mathbb{Z}$, where the group operation is addition. We want to find the order d of an element $\bar{a}$ of this group, and to determine how many elements are there of order d. We may assume that $\bar{a}$ is the congruence class of an integer a such that $1 \leq a \leq n$.

First, to say that $d = \mathrm{ord}(\bar{a})$, means that in the list

$$\bar{a},\ \ 2\bar{a} := \bar{a} + \bar{a}, \ldots,\ \ d\bar{a} := \underbrace{\bar{a} + \bar{a} + \cdots + \bar{a}}_{d-\text{times}},$$

the first time $\bar{0}$ appears is for $d\bar{a}$. Passing to integers that means that d is the minimal positive integer such that

$$n | da.$$

That means that among all multiples of a, we are looking for the first multiple da that is also divisible by n. This multiple is $\mathrm{lcm}(a,n)$, and by Exercise 8 on page 42,

$$da = \mathrm{lcm}(a,n) = \frac{an}{\gcd(a,n)},$$

implying that

$$d = \frac{n}{\gcd(a,n)}.$$

Conclusion: The congruence class $\bar{a}$ in $\mathbb{Z}/n\mathbb{Z}$ has order n if and only if $\gcd(a,n) = 1$ and, thus, the number of element of order n in $\mathbb{Z}/n\mathbb{Z}$ is

$$\varphi(n) = \sharp\{1 \leq a \leq n : \gcd(a,n) = 1\}.$$

(Here φ is **Euler's totient function** defined by this formula.)

Proposition 27.3.1. *Let $n > 1$ be an integer. Let $d|n$ be a positive integer. The number of elements of order d in the group $\mathbb{Z}/n\mathbb{Z}$ is $\varphi(d)$.*

Remark 27.3.2. By Lagrange's Theorem the condition $d|n$ is necessary for the existence of elements of order d. The converse does not hold for a general group. For example, S_3 is a group of order 6, $6|6$, but S_3 does not have an element of order 6. The Proposition states that the converse does hold for cyclic groups and even tells us how many elements there are of every order.

Proof. Write $n = dk$. Let

$$H_d := \{\bar{0}, \bar{k}, 2\bar{k}, \ldots, (d-1)\bar{k}\} = \langle \bar{k} \rangle.$$

This is the cyclic group of order d generated by k. As it is a cyclic group, it "behaves" exactly like $\mathbb{Z}/d\mathbb{Z}$ and there are therefore $\varphi(d)$ elements of order d in H.[21] It therefore suffices to prove that every element of order d belongs to H_d.

Let $\bar{a}$ be an element of order d and assume without loss of generality that $0 \leq a \leq n-1$. Divide the a by k (in $\mathbb{Z}$) with a residue to get $a = qk + r$, $\ 0 \leq r < k$. Then $dr = da - qdk \equiv 0 \pmod{n}$. But, $0 \leq dr < dk = n$ and so we must have $dr = 0$ and thus $r = 0$. Namely, $a = qk$ and therefore $\bar{a} \in H_d$. □

27.3.1. *Euler's totient function.* We discuss some properties of Euler's totient function; a function that appears a lot in number theory.

Note that $\{1 \leq a \leq n : \gcd(a,n) = 1\}$ are precisely the units of the ring $\mathbb{Z}/n\mathbb{Z}$. Indeed, if $\bar{a}$ is a unit, there is some $\bar{b}$ such that $\bar{a}\bar{b} = \bar{1}$, meaning that $n|(ab-1)$. Then, if $k|n$ and $k|a$ we find that $k|1$ and so $\gcd(a,n) = 1$. Conversely, if $\gcd(a,n) = 1$ then for some b, c we have $ab + nc = 1$ and taken mod n we find $\bar{a}\bar{b} = \bar{1}$. So, $\bar{a}$ is a unit. We conclude

- $$\varphi(n) = \sharp(\mathbb{Z}/n\mathbb{Z})^{\times}.$$

[21] To be more precise, and to use terminology explained a bit later, there is an isomorphism of group $H_d \cong \mathbb{Z}/d\mathbb{Z}$ taking the element $\bar{k} \in H_d$ to 1 in $\mathbb{Z}/d\mathbb{Z}$.

Let us now apply the Chinese Remainder Theorem. Assume that m, n are positive integers with $\gcd(m,n) = 1$. There is an isomorphism

$$f\colon \mathbb{Z}/mn\mathbb{Z} \xrightarrow{\cong} \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z},$$

and this is an isomorphism of *rings*. Thus, there is an induced isomorphism on the unit groups:

$$(\mathbb{Z}/mn\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times.$$

This gives:

- φ is a multiplicative function:

$$\varphi(mn) = \varphi(m)\varphi(n), \quad \text{if } (m,n) = 1.$$

It also follows directly from the definition that

- For a prime p and positive integer n:

$$\varphi(p^n) = p^n - p^{n-1} = p^n\left(1 - \frac{1}{p}\right).$$

Combining all that we proved, we find:

- For every positive integer n

$$\varphi(n) = n \cdot \prod_{p|n}\left(1 - \frac{1}{p}\right).$$

An additional interesting property is the following: every element of $\mathbb{Z}/n\mathbb{Z}$ has *some* order d; that order divides n, and in fact for such an integer d there are $\varphi(d)$ elements of order d. We therefore find:

-

$$n = \sum_{d|n} \varphi(d).$$

28. Homomorphisms and isomorphisms

28.1. **homomorphisms of groups.**

Definition 28.1.1. Let G, H be groups and

$$f\colon G \to H,$$

a function. The function f is called a **group homomorphism**, if

$$f(g_1g_2) = f(g_1)f(g_2), \quad \forall g_1, g_2 \in G.$$

In that case, we define the **kernel** of f as:

$$\mathrm{Ker}(f) = \{g \in G : f(g) = e_H\}.$$

Lemma 28.1.2. *Let $f\colon G \to H$ be a group homomorphism. Then:*

(1) $f(e_G) = e_H$;
(2) $f(g^{-1}) = f(g)^{-1}$;
(3) The image of f is a subgroup of H.

Proof. We have $f(e_G) = f(e_Ge_G) = f(e_G)f(e_G)$. Multiplying (in H) both sides by $f(e_G)^{-1}$ we find $e_H = f(e_G)$. Now, $e_H = f(e_G) = f(gg^{-1}) = f(g)f(g^{-1})$, which shows that $f(g^{-1}) = f(g)^{-1}$.

Finally, we show that $\mathrm{Im}(f)$ is a subgroup of H. Note that $e_H = f(e_G) \in \mathrm{Im}(f)$. If $h_1, h_2 \in \mathrm{Im}(f)$, say $h_i = f(g_i)$ then $h_1h_2 = f(g_1g_2)$ and $h_1^{-1} = f(g_1^{-1})$. This shows that $h_1h_2, h_1^{-1} \in \mathrm{Im}(f)$. □

Proposition 28.1.3. *Let $f\colon G \to H$ be a group homomorphism. $\mathrm{Ker}(f)$ is a subgroup of G. The homomorphism f is injective if and only if $\mathrm{Ker}(f) = \{e_G\}$.*

Proof. First, we proved that $f(e_G) = e_H$ and so $e_G \in \mathbf{Ker}(f)$. Next, if $g_1, g_2 \in \mathbf{Ker}(f)$ then $f(g_1) = f(g_2) = e_H$ and so $f(g_1g_2) = f(g_1)f(g_2) = e_He_H = e_H$. Therefore, $g_1g_2 \in \mathbf{Ker}(f)$. Finally, if $g \in \mathbf{Ker}(f)$ then $f(g^{-1}) = f(g)^{-1} = e_H^{-1} = e_H$ and so $g^{-1} \in \mathbf{Ker}(f)$ as well.

Suppose f is injective. Then, since $f(e_g) = e_H$, e_G is the only element mapping to e_H and so $\mathbf{Ker}(f) = \{e_G\}$. Conversely, suppose $\mathbf{Ker}(f) = \{e_G\}$ and $f(g_1) = f(g_2)$. Then $e_H = f(g_1)^{-1}f(g_2) = f(g_1^{-1})f(g_2) = f(g_1^{-1}g_2)$. That means that $g_1^{-1}g_2 \in \mathbf{Ker}(f)$ and so $g_1^{-1}g_2 = e_G$. That is, $g_1 = g_2$. □

28.2. **Isomorphism.**

Definition 28.2.1. A group homomorphism $f\colon G \to H$ is called an **isomorphism** if it is bijective. We use the notation $G \cong H$ to say that there is some isomorphism $f\colon G \to H$.

As in the case of rings, one verifies that if f is an isomorphism, the inverse function $g = f^{-1}$ is automatically a homomorphism and so an isomorphism as well. Also, one easily checks that a composition of group homomorphisms is a group homomorphism. It follows that being isomorphic is an equivalence relation on groups. Cf. §22.1.

Example 28.2.2. Let n be a positive integer. Any two cyclic groups of order n are isomorphic.

Indeed, suppose that $G = \langle g\rangle, H = \langle h\rangle$ are cyclic groups of order n. Define, for any integer a,

$$f(g^a) = h^a.$$

This is well defined; if $g^a = g^b$ then $g^{a-b} = e_G$ and so $n|(a-b)$. Thus, $a = b + kn$ and $f(g^a) = h^a = h^b(h^n)^k = h^b = f(g^b)$. Obviously f is a surjective homomorphism; f is also injective, because $f(g^a) = h^a = e_H$ implies that $n|a$ and so $g^a = e_G$.

In particular, we conclude that any cyclic group of order n is isomorphic to the group $\mathbb{Z}/n\mathbb{Z}$ (with the group operation being addition).

Example 28.2.3. Let p be a prime number then any two groups of order p are isomorphic. Indeed, we have seen that such groups are necessarily cyclic.

Theorem 28.2.4. (Cayley) *Let G be a finite group of order n then G is isomorphic to a subgroup of S_n.*

Proof. Let $g \in G$ and let

$$\sigma_g : G \to G, \quad \sigma_g(a) = ga.$$

We claim that σ_g is a permutation. It is injective, because $\sigma_g(a) = \sigma_g(b) \Rightarrow ga = gb \Rightarrow a = b$. It is surjective, because for any $b \in G$, $\sigma_g(g^{-1}b) = b$.

Identifying the permutations of G with S_n (just call the elements of G, $g_1, g_2, g_3, \dots$ and then their permutations match the permutation of the indices $1, 2, 3, \dots$), we got a map

$$G \to S_n, \quad g \mapsto \sigma_g.$$

This map is a homomorphism of groups: $\sigma_{gh}(a) = ghg = \sigma_g(\sigma_h(a))$. That is, $\sigma_{gh} = \sigma_g \circ \sigma_h$. This homomorphism is injective: if σ_g is the identity permutation then, in particular, $\sigma_g(e) = e$ and that implies $ge = e$, that is $g = e$. We get that G is isomorphic to its image, which is a subgroup of S_n, under this homomorphism. □

Remark 28.2.5. We were somewhat informal about identifying the permutations of G with S_n. A more rigorous approach is the following.

Lemma 28.2.6. *Let T, Z be sets and $f\colon T \to Z$ a bijection. The group of permutations of T and Z are isomorphic.*

Proof. Let $\sigma \in S_T$, a permutation of T. Then $f \circ \sigma \circ f^{-1}$ is a function from Z to itself, and being a composition of bijections is a bijection itself. We shall write more simply $f\sigma f^{-1}$ for $f \circ \sigma \circ f^{-1}$. We therefore got a function

$$S_T \to S_Z, \qquad \sigma \mapsto f\sigma f^{-1}.$$

We claim that $\sigma \mapsto f\sigma f^{-1}$ is a homomorphism. Indeed, given $\sigma_1, \sigma_2 \in S_T$ we have

$$f\sigma_1\sigma_2 f^{-1} = (f\sigma_1 f^{-1})(f\sigma_2 f^{-1}).$$

Moreover, it is easy to write an inverse to this homomorphism,

$$S_Z \to S_T, \qquad \tau \mapsto f^{-1}\tau f.$$

Therefore, we found a bijective homomorphism $S_T \to S_Z$, which shows those two permutation groups are isomorphic. □

29. Group actions on sets

29.1. **Basic definitions.** Let G be a group and let S be a non-empty set. We say that G **acts on** S if we are given a function

$$G \times S \to S, \quad (g,s) \longmapsto g \star s,$$

such that;

(i) $e \star s = s$ for all $s \in S$;

(ii) $(g_1 g_2) \star s = g_1 \star (g_2 \star s)$ for all $g_1, g_2 \in G$ and $s \in S$.

Given an action of G on S we can define the following sets. Let $s \in S$. Define the **orbit** of s

$$\mathrm{Orb}(s) = \{g \star s : g \in G\}.$$

Note that $\mathrm{Orb}(s)$ is a subset of S, equal to all the images of s under the action of the group G. We also define the **stabilizer** of s to be

$$\mathrm{Stab}(s) = \{g \in G : g \star s = s\}.$$

Note that $\mathrm{Stab}(s)$ is a subset of G. In fact, it is a subgroup, as Lemma 29.2.1 states.

29.2. **Basic properties.**

Lemma 29.2.1. *(1) Let $s_1, s_2 \in S$. We say that s_1 is related to s_2, i.e., $s_1 \sim s_2$, if there exists $g \in G$ such that $g \star s_1 = s_2$. This is an equivalence relation. The equivalence class of s_1 is its orbit $\mathrm{Orb}(s_1)$.*

(2) Let $s \in S$. The set $\mathrm{Stab}(s)$ is a subgroup of G.

(3) Suppose that both G and S have finitely many elements. Then

$$|\mathrm{Orb}(s)| = \frac{|G|}{|\mathrm{Stab}(s)|}.$$

Proof. (1) We need to show that this relation is reflexive, symmetric and transitive. First, we have $e \star s = s$ and hence $s \sim s$, meaning the relation is reflexive. Second, if $s_1 \sim s_2$ then for a suitable $g \in G$ we have $g \star s_1 = s_2$. Therefore, $g^{-1} \star (g \star s_1) = g^{-1} \star s_2$ and $(g^{-1}g) \star s_1 = g^{-1} \star s_2$. It follows that, $e \star s_1 = g^{-1} \star s_2$ and so, $s_1 = g^{-1} \star s_2$, which implies that $s_2 \sim s_1$.

It remains to show the relation is transitive. If $s_1 \sim s_2$ and $s_2 \sim s_3$ then for suitable $g_1, g_2 \in G$ we have $g_1 \star s_1 = s_2$ and $g_2 \star s_2 = s_3$. Therefore, $(g_2 g_1) \star s_1 = g_2 \star (g_1 \star s_1) = g_2 \star s_2 = s_3$, and hence $s_1 \sim s_3$.

Moreover, by its very definition, the equivalence class of an element s_1 of S is all the elements of the form $g \star s_1$ for some $g \in G$, namely, $\mathrm{Orb}(s_1)$.

(2) Let $H = \mathrm{Stab}(s)$. We have to show that: (i) $e \in H$, (2) if $g_1, g_2 \in H$ then $g_1 g_2 \in H$, and (iii) if $g \in H$ then $g^{-1} \in H$.

First, by the definition of group action, we have $e \star s = s$. Therefore, $e \in H$. Next, suppose that $g_1, g_2 \in H$, i.e., $g_1 \star s = s$ and $g_2 \star s = s$. Then, $(g_1 g_2) \star s = g_1 \star (g_2 \star s) = g_1 \star s = s$, which proves that $g_1 g_2 \in H$. Finally, if $g \in H$ then $g \star s = s$ and so $g^{-1} \star (g \star s) = g^{-1} \star s$. That is, $(g^{-1}g) \star s = g^{-1} \star s$ and so $e \star s = g^{-1} \star s$, or $s = g^{-1} \star s$, and therefore $g^{-1} \in H$.

(3) We claim that there exists a bijection between the left cosets of $H = \mathrm{Stab}(s)$ and the orbit of s. If we show that, then by Lagrange's theorem,

$$|\mathrm{Orb}(s)| = \text{no. of left cosets of } H = \text{index of } H = |G|/|H|.$$

Define a function

$$\{\text{left cosets of } H\} \xrightarrow{\phi} \mathrm{Orb}(s),$$

by

$$\phi(gH) = g \star s.$$

We claim that ϕ is a well-defined bijection. First
Well-defined: Suppose that $g_1H = g_2H$. We need to show that the rule ϕ gives the same result whether we take the representative g_1 or the representative g_2 to the coset, that is, we need to show $g_1 \star s = g_2 \star s$. Note that $g_1^{-1}g_2 \in H$, i.e., $(g_1^{-1}g_2) \star s = s$. We get $g_1 \star s = g_1 \star ((g_1^{-1}g_2) \star s) = (g_1(g_1^{-1}g_2)) \star s = g_2 \star s$.
ϕ is surjective: Let $t \in \mathrm{Orb}(s)$ then $t = g \star s$ for some $g \in G$. Thus, $\phi(gH) = g \star s = t$, and we get that ϕ is surjective.
ϕ is injective: Suppose that $\phi(g_1H) = \phi(g_2H)$. We need to show that $g_1H = g_2H$. Indeed, $\phi(g_1H) = \phi(g_2H)$ implies $g_1 \star s = g_2 \star s$ and so that $g_2^{-1} \star (g_1 \star s) = g_2^{-1} \star (g_2 \star s)$; that is, $(g_2^{-1}g_1) \star s = e \star s = s$. Therefore, $g_2^{-1}g_1 \in \mathrm{Stab}(s) = H$ and hence $g_1H = g_2H$. □

Corollary 29.2.2. *The set S is a disjoint union of orbits.*

Proof. The orbits are the equivalence classes of the equivalence relation $\sim$ defined in Lemma 29.2.1. Any equivalence relation partitions the set into disjoint equivalence classes. □

29.3. **Some examples.**

Example 29.3.1. Let $G = S_n$, the symmetric group on $T = \{1, 2, \ldots, n\}$. The group G acts on T in a natural way:

$$\sigma \star i := \sigma(i).$$

The stabilizer of i is the permutations fixing i. These permutations can be identified with permutations of the set $T - \{i\}$, because $\sigma \in \mathrm{Stab}(i)$ induces a permutation of $T - \{i\}$ and a permutation of $T - \{i\}$ can be extended to a permutation of T sending i to itself. Thus, for every i, $\mathrm{Stab}(i) \cong S_{n-1}$. The orbit of i is T; for every j the transposition (ij) shows that $j \in \mathrm{Orb}(i)$.

Example 29.3.2. Let G be the group of real numbers $\mathbb{R}$. The group operation is addition. Let S be the sphere in $\mathbb{R}^3$ of radius 1 about the origin. The group $\mathbb{R}$ acts by rotating around the z-axis. An element $r \in \mathbb{R}$ rotates by r radians. For every point $s \in S$, different from the poles, the stabilizer is $2\pi\mathbb{Z}$. For the poles the stabilizer is $\mathbb{R}$. The orbit of every point is the altitude line on which it lies.

Example 29.3.3. Let G be a group and H a subgroup of G. Then H acts on G by

$$H \times G \to G, \quad (h, g) \mapsto hg.$$

Here H plays the role of the group and G the role of the set in the definition. This is indeed a group action: $e_H g = g$ for all $g \in G$, because by definition $e_H = e_G$. Also, $h_1(h_2)g = (h_1h_2)g$ is nothing but the associative law.

The orbit of $g \in G$ is

$$\mathrm{Orb}(g) = \{hg : h \in H\} = Hg.$$

That is, the orbits are the right cosets of H. We have that G is a disjoint union of orbits, namely, a disjoint union of cosets. The stabilizer of any element $g \in G$ is $\{e\}$. The formula we have proven, $|\mathrm{Orb(g)}| = |H|/|\mathrm{Stab}(g)|$, gives us $|Hg| = |H|$ for any $g \in G$, and we see that we have another point of view on Lagrange's theorem.

Example 29.3.4. We consider a roulette with n sectors and write $n = i_1 + \cdots + i_k$, for some positive (and fixed) integers $i_1, \ldots, i_k$. We suppose we have different colors $c_1, \ldots, c_k$ and we color i_1 sectors of the roulette by the color c_1, i_2 sectors by the color c_2 and so on. The sectors can be chosen as we wish and so there are many possibilities. We get a set S of colored roulettes. Now, we turn the roulette r steps clockwise, say, and we get another colored roulette, usually with different coloring. Nonetheless, it is natural to view the two coloring as the same, since "they only depend on your point of view". We may formalize this by saying that the group $\mathbb{Z}/n\mathbb{Z}$ acts on S; r acts on a colored roulette by turning it r steps clockwise, and by saying that we are interested in the number of orbits for this action.

Example 29.3.5. Let G be the dihedral group D_8. Recall that G is the group of symmetries of a regular octagon in the plane.

$$G = \{e, x, x^2, \ldots, x^7, y, xy, x^2y, \ldots, x^7y\},$$

where x is rotation clockwise by angle $2\pi/8$ and y is reflection through the y-axis. We have the relations

$$y^2 = x^8 = e, \quad xyxy = e.$$

We let S be the set of colorings of the octagon (= necklaces laid on the table) having 4 red vertices (rubies) and 4 green vertices (sapphires). The group G acts on S by its action on the octagon.

For example, the coloring s_0, consisting of alternating green and red, is certainly preserved under y and under x^2. Therefore, the stabilizer of s_0 contains at least the set of eight elements

$$\{e, x^2, x^4, x^6, y, x^2y, x^4y, x^6y\}. \tag{4}$$

Remember that the stabilizer is a subgroup and, by Lagrange's theorem, of order dividing $16 = |G|$. On the other hand, $\mathrm{Stab}(s_0) \neq G$ because $x \notin \mathrm{Stab}(s_0)$. It follows that the stabilizer has exactly 8 elements and is equal to the set in (4).

Let H be the stabilizer of s_0. According to Lemma 29.2.1 the orbit of s_0 is in bijection with the left cosets of $H = \{e, x^2, x^4, x^6, y, x^2y, x^4y, x^6y\}$. By Lagrange's theorem there are two cosets. For example, H and xH are distinct cosets. The proof of Lemma 29.2.1 tells us how to find the orbit: it is the set $\{s_0, xs_0\}$, which is of course quite clear if you think about it.

30. The Cauchy-Frobenius Formula

Theorem 30.0.1. (CFF)[22] *Let G be a finite group acting on a finite set S. Let N be the number of orbits of G in S. Define*[23]

$$I(g) = |\{s \in S : g \star s = s\}|$$

(the number of elements of S fixed by the action of g). Then [24]

$$N = \frac{1}{|G|} \sum_{g \in G} I(g).$$

Remark 30.0.2. Note that $I(g)$ is the number of fixed points for the action of g on S. Thus, the CFF can be interpreted as saying that the number of orbits is the average number of fixed points (though this does not make the assertion more obvious).

Proof. We define a function

$$T : G \times S \to \{0, 1\}, \quad T(g, s) = \begin{cases} 1 & g \star s = s, \\ 0 & g \star s \neq s. \end{cases}$$

[22] This is also sometimes called Burnside's formula.

[23] Another notation for $I(g)$ is $Fix(g)$.

[24] The sum appearing in the formula means just that: If you write $G = \{g_1, \ldots, g_n\}$ then $\sum_{g \in G} I(g)$ is $\sum_{i=1}^{n} I(g_i) = I(g_1) + I(g_2) + \cdots + I(g_n)$. The double summation $\sum_{g \in G} \sum_{s \in S} T(g, s)$ appearing in the proof means that if we write $S = \{s_1, \ldots, s_m\}$ then the double sum is $T(g_1, s_1) + T(g_1, s_2) + \cdots + T(g_1, s_m) + T(g_2, s_1) + T(g_2, s_2) + \cdots + T(g_2, s_m) + \cdots + T(g_n, s_1) + T(g_n, s_2) + \cdots + T(g_n, s_m)$.

Note that for a fixed $g \in G$ we have

$$I(g) = \sum_{s \in S} T(g,s),$$

and that for a fixed $s \in S$ we have

$$|\mathrm{Stab}(s)| = \sum_{g \in G} T(g,s).$$

Let us fix representatives $s_1, \ldots, s_N$ for the N disjoint orbits of G in S. Now,

$$\begin{aligned}
\sum_{g \in G} I(g) &= \sum_{g \in G} \left(\sum_{s \in S} T(g,s) \right) = \sum_{s \in S} \left(\sum_{g \in G} T(g,s) \right) \\
&= \sum_{s \in S} |\mathrm{Stab}(s)| = \sum_{s \in S} \frac{|G|}{|\mathrm{Orb}(s)|} \\
&= \sum_{i=1}^{N} \sum_{s \in \mathrm{Orb}(s_i)} \frac{|G|}{|\mathrm{Orb}(s)|} = \sum_{i=1}^{N} \sum_{s \in \mathrm{Orb}(s_i)} \frac{|G|}{|\mathrm{Orb}(s_i)|} \\
&= \sum_{i=1}^{N} \frac{|G|}{|\mathrm{Orb}(s_i)|} \cdot |\mathrm{Orb}(s_i)| = \sum_{i=1}^{N} |G| \\
&= N \cdot |G|.
\end{aligned}$$

□

Remark 30.0.3. If N, the number of orbits, is equal 1 we say that G acts **transitively** on S. It means exactly that: For every $s_1, s_2 \in S$ there exists $g \in G$ such that $g \star s_1 = s_2$. Note that if G and S are finite then if G acts transitively then the number of elements in S divides the number of elements in G,

$$|S| \ \mid \ |G|,$$

because, if $S = \mathrm{Orb}(s)$ then $|S| = |G|/|\mathrm{Stab}(s)|$.

Corollary 30.0.4. *Let G be a finite group acting transitively on a finite set S. Suppose that $|S| > 1$. Then there exists $g \in G$ without fixed points.*

Proof. By contradiction. Suppose that every $g \in G$ has a fixed point in S. That is, suppose that for every $g \in G$ we have

$$I(g) \geq 1.$$

Since $I(e) = |S| > 1$ we have that

$$\sum_{g \in G} I(g) > |G|.$$

By Cauchy-Frobenius formula, the number of orbits N is greater than 1. Contradiction. □

30.1. **Some applications to Combinatorics.**

Example 30.1.1. How many roulettes with 11 wedges painted 2 blue, 2 green and 7 red are there when we allow rotations?

Let S be the set of painted roulettes. Let us enumerate the sectors of a roulette by the numbers $1, \ldots, 11$. The set S is a set of $\binom{11}{2}\binom{9}{2} = 1980$ elements (choose which 2 are blue, and then choose out of the nine left which 2 are green).

Let G be the group $\mathbb{Z}/11\mathbb{Z}$. It acts on S by rotations. The element 1 rotates a painted roulette by angle $2\pi/11$ anti-clockwise. The element n rotates a painted roulette by angle $2n\pi/11$ anti-clockwise. We are interested in N, the number of orbits for this action. We use **CFF**.

The identity element always fixes the whole set. Thus $I(0) = 1980$. We claim that if $1 \leq i \leq 10$ then i doesn't fix any element of S. We use the following fact that we have proved before: Let G be a finite group of prime order p. Let $g \neq e$ be an element of G. Then $\langle g \rangle = G$.

Suppose that $1 \leq i \leq 10$ and i fixes s. Then so does $\langle i \rangle = \mathbb{Z}/11\mathbb{Z}$ (the stabilizer is a subgroup). But any coloring fixed under rotation by 1 must be single colored! Contradiction.

Applying **CFF** we get

$$N = \frac{1}{11}\sum_{n=0}^{10} I(n) = \frac{1}{11}\cdot 1980 = 180.$$

Example 30.1.2. How many roulettes with 12 wedges painted 2 blue, 2 green and 8 red are there when we allow rotations?

Let S be the set of painted roulettes. Let us enumerate the sectors of a roulette by the numbers $1,\dots,12$. The set S is a set of $\binom{12}{2}\binom{10}{2} = 2970$ elements (choose which 2 are blue, and then choose out of the ten left which 2 are green).

Let G be the group $\mathbb{Z}/12\mathbb{Z}$. It acts on S by rotations. The element 1 rotates a painted roulette by angle $2\pi/12$ anti-clockwise. The element n rotates a painted roulette by angle $2n\pi/12$ anti-clockwise. We are interested in N, the number of orbits for this action. We use **CFF**.

The identity element always fixes the whole set. Thus $I(0) = 2970$. We claim that if $1 \leq i \leq 11$ and $i \neq 6$ then i doesn't fix any element of S. Indeed, suppose that i fixes a painted roulette. Say in that roulette the r-th sector is blue. Then so must be the $i+r$ sector (because the r-th sector goes under the action of i to the $r+i$-th sector). Therefore so must be the $r+2i$ sector. But there are only 2 blue sectors! The only possibility is that the $r+2i$ sector is the same as the r sector, namely, $i = 6$.

If i is equal to 6 and we enumerate the sectors of a roulette by the numbers $1,\dots,12$ we may write i as the permutation

$$(1\ 7)(2\ 8)(3\ 9)(4\ 10)(5\ 11)(6\ 12).$$

In any coloring fixed by $i = 6$ the colors of the pairs $(1\ 7), (2\ 8), (3\ 9), (4\ 10), (5\ 11)$ and $(6\ 12)$ must be the same. We may choose one pair for blue, one pair for green. The rest would be red. Thus there are $30 = 6\cdot 5$ possible choices. We summarize:

element g	$I(g)$
0	2970
$i \neq 6$	0
$i = 6$	30

Applying **CFF** we get that there are

$$N = \frac{1}{12}(2970 + 30) = 250$$

different roulettes.

Example 30.1.3. In this example S is the set of necklaces made of four rubies and four sapphires laid on the table. We ask how many necklaces there are when we allow rotations and flipping-over. We may talk of S as the colorings of a regular octagon, four vertices are green and four are red. The group $G = D_8$ acts on S and we are interested in the number of orbits for the group G. The results are the following

element g	$I(g)$
e	70
y, y^3, y^5, y^7	0
y^2, y^6	2
y^4	6
xy^i for $i = 0,\dots,7$	6

We explain how the entries in the table are obtained:

The identity always fixes the whole set S. The number of elements in S is $\binom{8}{4} = 70$ (choosing which 4 would be green).

The element y cannot fix any coloring, because any coloring fixed by y must have all sections of the same color (because $y = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$). If y^r fixes a coloring s_0 so does $(y^r)^r = y^{(r^2)}$ because the stabilizer is a subgroup. Apply that for $r = 3, 5, 7$ to see that if y^r fixes a coloring so does y , which is impossible. [25]

Now, y^2, written as a permutation, is $(1\ 3\ 5\ 7)(2\ 4\ 6\ 8)$. We see that if, say 1 is green so are $3, 5, 7$ and the rest must be red. That is, all the freedom we have is to choose whether the cycle $(1\ 3\ 5\ 7)$ is green or red. This gives us two colorings fixed by y^2. The same rationale applies to $y^6 = (8\ 6\ 4\ \ 2)(7\ 5\ 3\ 1)$.

Consider now y^4. It may written in permutation notation as $(1\ 5)(2\ 6)(3\ 7)(4\ 8)$. In any coloring fixed by y^4 each of the cycles $(1\ 5)(2\ 6)(3\ 7)$ and $(4\ 8)$ must be single colored. There are thus $\binom{4}{2} = 6$ possibilities (Choosing which 2 out of the four cycles would be green).

It remains to deal with the elements xy^i. We recall that these are all reflections. There are two kinds of reflections. One may be written using permutation notation as

$$(i_1\ i_2)(i_3\ i_4)(i_5\ i_6)$$

(with the other two vertices being fixed. For example $x = (2\ 8)(3\ 7)(4\ 6)$ is of this form). The other kind is of the form

$$(i_1\ i_2)(i_3\ i_4)(i_5\ i_6)(i_7\ i_8).$$

(For example $xy = (1\ 8)(2\ 7)(3\ 6)(4\ 5)$ is of this sort). Whichever the case, one uses similar reasoning to deduce that there are 6 colorings preserved by a reflection.

One needs only apply **CFF** to get that there are

$$N = \frac{1}{16}(70 + 2 \cdot 2 + 6 + 8 \cdot 6) = 8$$

distinct necklaces.

Example 30.1.4. Consider a tetrahedron with faces marked $1, 2, 3, 4$. It takes a little thinking, but one can see that each symmetry is determined by its action on the faces. As we have done previously, consider symmetries of the tetrahedron that preserve orientation. We noted before (Example 26.3.3) that there are 12 such symmetries, but here we want a more explicit description. For example, (123) is such a symmetry; it rotates the tetrahedron relative to the plane on which the face 4 lies. On the other hand, (12) is not such a symmetry. We conclude that the symmetries that preserve orientation are a subgroup of S_4 that is not equal to S_4. Let us call this subgroup A_4 (later on, we shall define the groups A_n in general and our notation is consistent). Clearly A_4 contains $\{1, (123), (132), (234), (243), (134), (143), (124), (142)\}$ and since it is closed under multiplication (being a subgroup) also $(132)(134) = (12)(34)$ and similarly, $(13)(24)$ and $(14)(23)$, are elements of A_4. We have already identified 12 elements in A_4 and since the order of A_4 divides the order of S_4, which is 24, A_4 is in fact equal to

$$\{1, (123), (132), (234), (243), (134), (143), (124), (142), (12)(34), (13)(24), (14)(23)\}.$$

Let us count how many colorings of the faces of the tetrahedron are there using 4 distinct colors, each once. The number of coloring is $4! = 24$ (choose for each face its color). No symmetry but the identity preserves a coloring and so by CFF we get that the number of colorings up to A_4 identifications is 2.

Suppose now we want to color with 2 colors, say red and blue, painting two faces red and two faces blue. The total number of colorings are $\binom{4}{2} = 6$. In this case, a three cycle cannot fix a coloring, while each permutation of the type $(ab)(cd)$ fixes exactly two colorings (choose if the faces a, b are both red or both blue). Therefore, the number of colorings up to symmetries is

$$N = \frac{1}{12}(6 + 3 \times 2) = 1.$$

[25] $y^{(3^2)} = g^9 = g$ because $y^8 = e$, etc.

31. Cauchy's theorem: a wonderful proof

One application of group actions is to provide a simple proof of an important theorem in the theory of finite groups. Every other proof I know is very complicated.

Theorem 31.0.1. (Cauchy) *Let G be a finite group of order n and let p be a prime dividing n. Then G has an element of order p.*

Proof. Let S be the set consisting of p-tuples $(g_1,\dots,g_p)$ of elements of G, considered up to cyclic permutations. Thus if T is the set of p-tuples $(g_1,\dots,g_p)$ of elements of G, S is the set of orbits for the action of $\mathbb{Z}/p\mathbb{Z}$ on T by cyclic shifts. One may therefore apply **CFF** and get

$$|S| = \frac{n^p - n}{p} + n.$$

If $n\mid |S|$, then n divides $(n^p-n)/p$ and so p divides $n^{p-1}-1$. But p divides n and we get a contradiction. Thus, $n \not| |S|$.

Now define an action of G on S. Given $g \in G$ and $(g_1,\dots,g_p) \in S$ we define

$$g(g_1,\dots,g_p) = (gg_1,\dots,gg_p).$$

This is well defined. Since the order of G is n, since $n \not| |S|$, and since S is a disjoint union of orbits of G, there must be an orbit $\mathrm{Orb}(s)$ whose size is not n. However, the size of an orbit is $|G|/|\mathrm{Stab}(s)|$, and we conclude that there must an element $(g_1,\dots,g_p)$ in S with a non-trivial stabilizer. This means that for some $g \in G$, such that $g \neq e$, we have

$$(gg_1,\dots,gg_p) \text{ is equal to } (g_1,\dots,g_p) \text{ up to a cyclic shift.}$$

This means that for some i we have

$$(gg_1,\dots,gg_p) = (g_{i+1},g_{i+2},g_{i+3},\dots,g_p,g_1,g_2,\dots,g_i).$$

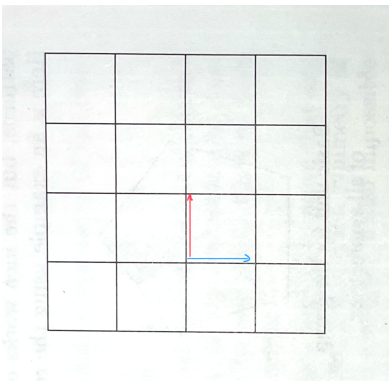
Therefore, $gg_1 = g_{i+1}$, $g^2g_1 = gg_{i+1} = g_{2i+1},\dots$, $g^pg_1 = \cdots = g_{pi+1} = g_1$ (we always read the indices mod p). That is, there exists $g \neq e$ with

$$g^p = e.$$

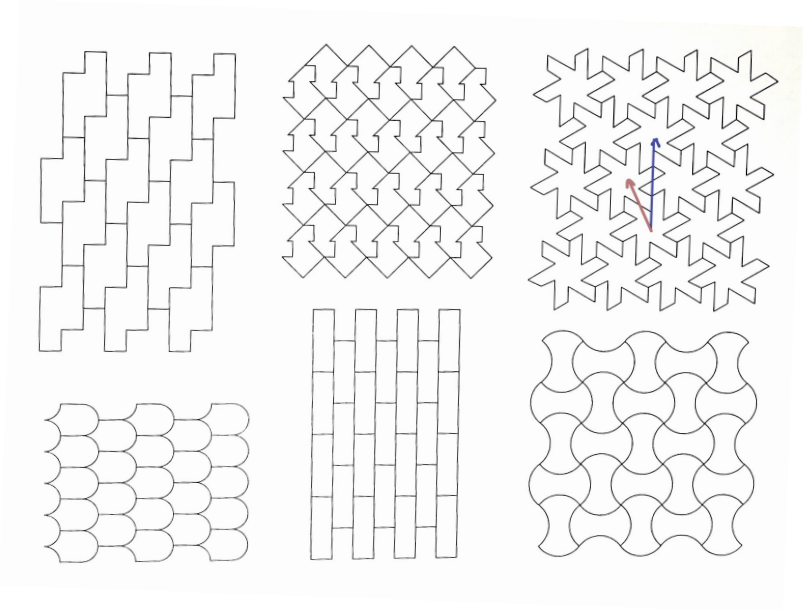
Since the order of g divides p and p is prime, the order of g must be p. □

32. Wallpaper groups

A **wallpaper pattern** is a planar pattern with translation symmetries in two different directions. For example, the square paper pattern, has both a horizontal translation and vertical translation:



Here are a few more patterns:



The symmetry group Γ of a wall paper pattern is called a **wallpaper group**. Because of its definition Γ always contain a subgroup isomorphic to $\mathbb{Z}^2$. If we let $\mathbb{T}$ be all the translation maps in Γ, one can show that $\mathbb{T}$ is isomorphic to $\mathbb{Z}^2$.

Typically, Γ is larger than $\mathbb{T}$. The subgroup of Γ consisting of permutations preserving a given base point can be thought of as rigid linear maps of $\mathbb{R}^2$ that take 0 to itself and the pattern to itself. This is a finite group and we denote it by Γ_0. For example, for the square paper pattern $\Gamma_0 \cong D_4$, but for an infinite honeycomb design it is D_6. The group Γ_0 is heavily restricted by geometric considerations and is in fact a subgroup of D_n for $n = 4, 6$. It not hard to show that, conversely, any such subgroup arises for some wallpaper pattern.

In 1891 E. Fedorov proved that here are only 17 distinct wallpaper groups. A key point is to note that (using concepts introduced in the next section) $\mathbb{T} \lhd \Gamma$ and $\Gamma/\mathbb{T} \supseteq \Gamma_0$ (but not always equal). Moreover, $\Gamma/\mathbb{T}$ is itself a finite subgroup of orthogonal transformation of the plane. An approach to classifying wallpaper groups based on these observations was carried out by P. Morandi and can be found here. Other good resources are the book The symmetry of things by Conway, Burgiel and Goodman-Stauss, and the book Groups and Symmetry by Armstrong.

33. The first isomorphism theorem for groups

33.1. **Normal subgroups.**

Definition 33.1.1. Let G be a group and H a subgroup of G. H is called a **normal** subgroup if for every $g \in G$ we have

$$gH = Hg.$$

Note that $gH = Hg$ if and only if $gHg^{-1} = H$, where $gHg^{-1} = \{ghg^{-1} : h \in H\}$. Thus, we could also define a normal subgroup H to be a subgroup such that $gHg^{-1} = H$ for all $g \in G$, equivalently, $\forall g \in G, \forall h \in H, ghg^{-1} \in H$.

Lemma 33.1.2. *Let H be a subgroup of a group G. Then H is normal if and only if*

$$gHg^{-1} \subset H, \quad \forall g \in G.$$

Proof. Clearly if H is normal, $gHg^{-1} \subset H, \forall g \in G$. Suppose then that $gHg^{-1} \subset H, \forall g \in G$. Given $g \in G$ we have then $gHg^{-1} \subset H$ and also $g^{-1}H(g^{-1})^{-1} \subset H$. The last inclusion is just $g^{-1}Hg \subset H$, which is equivalent to $H \subset gHg^{-1}$. We conclude that $gHg^{-1} = H$. □

Our main example of a normal subgroup is the kernel of a homomorphism.

Proposition 33.1.3. *Let $f : G \to H$ be a group homomorphism. Then $\mathrm{Ker}(f)$ is a normal subgroup of G.*

Proof. We proved already that $\mathrm{Ker}(f)$ is a subgroup. Let $g \in G, h \in \mathrm{Ker}(f)$; we need to show that $ghg^{-1} \in \mathrm{Ker}(f)$, that is $f(ghg^{-1}) = e_H$. We calculate $f(ghg^{-1}) = f(g)f(h)f(g^{-1}) = f(g)e_H f(g^{-1}) = f(g)f(g^{-1}) = f(g)f(g)^{-1} = e_H$. □

Example 33.1.4. For any group G, $\{e_G\}$ and G are normal subgroups. If G is a commutative group, any subgroup of G is a normal subgroup.

33.2. **Quotient groups.** Similar to the construction of a quotient ring, we construct quotient groups.

Let G be a group and H a normal subgroup of G. We let the **quotient group** G mod H, denoted G/H, be the collection of left cosets of H. We define multiplication by

$$(aH)(bH) = abH.$$

We claim that this is well defined, namely, if $aH = a_1H, bH = b_1H$ then $abH = a_1b_1H$. Indeed, we have $a = a_1h$ for some $h \in H$ and $b = b_1h'$ for some $h' \in H$. Also, $hb_1 \in Hb_1 = b_1H$ and so $hb_1 = b_1h''$ for some $h'' \in H$. Then, $abH = a_1hb_1h'H = a_1b_1h''h'H = a_1b_1H$ (if $t \in H$ then $tH = H$).

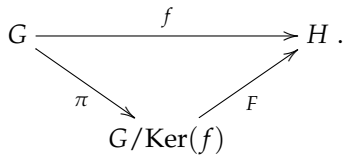
We now verify the group axioms. We use the notation $\bar{a}$ for aH. Then the group law is

$$\bar{a}\ \bar{b} = \overline{ab}.$$

We have $(\bar{a}\ \bar{b})\bar{c} = \overline{ab}\ \bar{c} = \overline{(ab)c} = \overline{a(bc)} = \bar{a}\ \overline{bc} = \bar{a}(\bar{b}\ \bar{c})$. Thus, this is an associative operation. We have $\bar{a}\ \bar{e}_G = \overline{ae_G} = \bar{a}$ and $\bar{e}_G\ \bar{a} = \overline{e_Ga} = \bar{a}$. So there is an identity element $e_{G/H}$ and it is equal to $\bar{e}_G = H$. Finally, $\bar{a}\ \overline{a^{-1}} = \overline{aa^{-1}} = \bar{e}_G = e_{G/H}$ and $\overline{a^{-1}}\ \bar{a} = \overline{a^{-1}a} = \bar{e}_G = e_{G/H}$. Thus, every $\bar{a}$ is invertible and its inverse is $\overline{a^{-1}}$ (that is, $(aH)^{-1} = a^{-1}H$).

33.3. **The first isomorphism theorem.**

Theorem 33.3.1. *Let $f : G \to H$ be a surjective group homomorphism. The canonical map $\pi : G \to G/\mathrm{Ker}(f)$ is a homomorphism with kernel $\mathrm{Ker}(f)$. There is an isomorphism $F : G/\mathrm{Ker}(f) \to H$, such that the following diagram commutes:*



Proof. First we check that $\pi : G \to G/\mathrm{Ker}(f)$ is a homomorphism, where $\pi(a) = \bar{a} = a\mathrm{Ker}(f)$. Indeed, this is just the formula $\overline{ab} = \bar{a}\ \bar{b}$. The kernel is $\{a \in G : a\mathrm{Ker}(f) = \mathrm{Ker}(f)\} = \mathrm{Ker}(f)$.

Let us define

$$F\colon G/\mathrm{Ker}(f) \to H, \qquad F(\bar{a}) = f(a).$$

This is well defined: if $\bar{a} = \bar{b}$ then $b^{-1}a \in \mathrm{Ker}(f)$, so $f(b) = f(b)f(b^{-1}a) = f(b(b^{-1}a)) = f(a)$. Clearly $F \circ \pi = f$.

F is a homomorphism: $F(\bar{a}\ \bar{b}) = F(\overline{ab}) = f(ab) = f(a)f(b) = F(\bar{a})F(\bar{b})$. Furthermore, F is surjective, since given $h \in H$ we may find $a \in G$ such that $f(a) = h$ and so $F(\bar{a}) = h$. Finally, F is injective, because $F(\bar{a}) = f(a) = e_H$ means that $a \in \mathrm{Ker}(f)$ so $\bar{a} = e_{G/H}$. □

Example 33.3.2. Let $\mathbb{F}$ be a field. Recall the group of matrices $\mathrm{GL}_2(\mathbb{F})$,

$$\mathrm{GL}_2(\mathbb{F}) = \left\{ M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{F}, \det(M) = ad - bc \neq 0 \right\}.$$

We have also noted that the determinant is multiplicative

$$\det(MN) = \det(M)\det(N).$$

We may now view this fact as saying that the function

$$\det : \mathrm{GL}_2(\mathbb{F}) \to \mathbb{F}^\times,$$

is a group homomorphism. It is a surjective group homomorphism, because given any $a \in \mathbb{F}^\times$ the matrix $\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}$ has determinant a. The kernel is called $\mathrm{SL}_2(\mathbb{F})$, it is equal to the matrices with determinant 1. It is a normal subgroup of $\mathrm{GL}_2(\mathbb{F})$ and by the first isomorphism theorem $\mathrm{GL}_2(\mathbb{F})/\mathrm{SL}_2(\mathbb{F}) \cong \mathbb{F}^\times$.

Example 33.3.3. The homomorphic images of S_3. We wish to identify all the homomorphic images of S_3. If $f : S_3 \to G$ is a group homomorphism then $\mathrm{Ker}(f)$ is a normal subgroup of S_3. We begin therefore by finding all normal subgroups of S_3.

We know that every nontrivial subgroup of S_3 contains a subgroup of the form $\langle (ij) \rangle$ for some transposition (ij) or the subgroup $A_3 := \langle (123) \rangle$. That there are no other subgroups follows from the following observation: if $H \subset K \subset G$ are groups and G is finite, then $|G|/|K|$ divides $|G|/|H|$, because the quotient is $|K|/|H|$. In our situation, for a non-trivial subgroup H we have $|S_3|/|H|$ is either 2 or 3 and those are prime. It follows that either $|K| = |G|$ or $|K| = |H|$ and so that either $K = G$ or $K = H$.

The subgroups of order 2 are not normal. For example, $(13)(12)(13)^{-1} = (13)(12)(13) = (23)$, which shows that $\{1, (12)\}$ is not normal, etc. On the other hand, the subgroup $A_3 := \{1, (123), (132)\}$ is normal. This follows from it being of index 2 (see assignments); another argument appears below. Since S_3/A_3 has order 2, it must be isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

We conclude that there are three options:

(1) $\mathrm{Ker}(f) = \{1\}$. In this case, S_3 is isomorphic to its image.
(2) $\mathrm{Ker}(f) = S_3$. In this case $S_3/\mathrm{Ker}(f) = S_3/S_3 \cong \{1\}$ is the trivial group.
(3) $\mathrm{Ker}(f) = A_3$. In this case S_3/A_3 is a group of 2 elements, obviously cyclic. Thus $S_3/A_3 \cong \mathbb{Z}/2\mathbb{Z}$

33.4. **Groups of low order.**

33.4.1. *Groups of order 1.* There is a unique group of order 1, up to isomorphism. It consists of its identity element alone. There is only one way to define a homomorphism between two groups of order 1 and it is an isomorphism.

33.4.2. *Groups of order 2, 3, 5, 7.* Recall that we proved that every group G of prime order is cyclic, and, in fact, any non-trivial element is a generator. This implies that any subgroup of G different from $\{e_G\}$ is equal to G. We also proved that any two cyclic groups having the same order are isomorphic. We therefore conclude:

Corollary 33.4.1. *Every group G of prime order p is isomorphic to $\mathbb{Z}/p\mathbb{Z}$; it has no subgroups apart from the trivial subgroups $\{e_g\}, G$.*

In particular, this corollary applies to groups of order $2, 3, 5, 7$.

33.4.3. *Groups of order 4.* Let G be a group of order 4.
First case: G is cyclic.

In this case we have $G \cong \mathbb{Z}/4\mathbb{Z}$. Its subgroups are $\{0\}, \mathbb{Z}/4\mathbb{Z}$ and $H = \langle 2 \rangle = \{0, 2\}$. There are no other subgroups because if a subgroup J contains an element g it contains the cyclic subgroup generated by g. In our case, the elements 1 and 3 are generators, so any subgroup not equal to G is contained in $\{0, 2\}$.

Since G is abelian, H is normal. G/H has order $|G|/|H| = 4/2 = 2$ and so $G/H \cong \mathbb{Z}/2\mathbb{Z}$.

Second case. G is not cyclic.
Claim: **Every element of G different from e_G has order 2.**
Proof: we have $\mathrm{ord}(g) = |\langle g \rangle|$ and it divides $|G|$. So, in our case, $\mathrm{ord}(g) = 1, 2$ or 4. If $\mathrm{ord}(g) = 4$, we get that G is cyclic and if $\mathrm{ord}(g) = 1$ then $g = e_G$. Thus, we must have $\mathrm{ord}(g) = 2$.

Claim: **Let G be a group in which every element different from the identity has order 2. Then G is commutative.**
Proof: Note first that if $a \in G$ has order 2 (or is the identity) then $aa = e_G$ and so $a^{-1} = a$. Now, we need to show that for every $a, b \in G$ we have $ab = ba$. But this is equivalent to $ab = b^{-1}a^{-1}$. Multiply both sides by ab and we see that we need to prove that $abab = e_G$. But, $abab = (ab)^2$ and so is equal to e_G, by assumption.

One example of a group of order 4 satisfying all these properties is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. We claim that $G \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Pick two distinct elements g_1, g_2 of G, not equal to the identity. Define a map

$$f\colon \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \to G, \qquad f(a,b) = g_1^a g_2^b.$$

This is well defined: if $(a,b) = (a',b')$ then $a = a' + 2c, b = b' + 2d$ and we get $f(a,b) = g_1^a g_2^b = g_1^{a'}(g_1^2)^c g_2^{b'}(g_2^2)^d = g_1^{a'} g_2^{b'} = f(a',b')$. The map is also a homomorphism: $f((a_1,b_1) + (a_2,b_2)) = f(a_1 + a_2, b_1 + b_2) = g_1^{a_1+a_2} g_2^{b_1+b_2} = g_1^{a_1} g_1^{a_2} g_2^{b_1} g_2^{b_2}$. Because G is commutative we can rewrite this as $f(a_1 + a_2, b_1 + b_2) = g_1^{a_1} g_2^{b_1} g_1^{a_2} g_2^{b_2} = f(a_1,b_1) \cdot f(a_2,b_2)$.

The image of f is a subgroup with at least 3 elements, namely, e_G, g_1, g_2. By Lagrange the image then must be G. It follows that f is surjective and so, because both source and target have four elements, is also injective.

The non-trivial subgroups of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ are all cyclic. They are $\{(0,0),(0,1)\}, \{(0,0),(1,0)\}$ and $\{(0,0),(1,1)\}$. Since the group is commutative they are all normal and the quotient in every case has order 2, hence isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

33.4.4. *Groups of order 6.* We know three candidates already $\mathbb{Z}/6\mathbb{Z}, \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ and S_3. Now, in fact, $\mathbb{Z}/6\mathbb{Z} \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ (CRT). And since S_3 is not commutative it is not isomorphic to $\mathbb{Z}/6\mathbb{Z}$. In fact, every group of order 6 is isomorphic to either $\mathbb{Z}/6\mathbb{Z}$ or S_3. We don't prove it here.

The subgroups of $\mathbb{Z}/6\mathbb{Z}$: Let n be a positive integer. We have a surjective group homomorphism $\pi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$. Similar to the situation with rings one can show that this gives a bijection between subgroups H of $\mathbb{Z}$ that contain $n\mathbb{Z}$ and subgroups K of $\mathbb{Z}/n\mathbb{Z}$. The bijection is given by

$$H \mapsto \pi(H), \quad K \mapsto \pi^{-1}(K).$$

The subgroups of $\mathbb{Z}$ are all cyclic, having the form $n\mathbb{Z}$ for some n (same proof as for ideals, really). We thus conclude that the subgroups of $\mathbb{Z}/n\mathbb{Z}$ are cyclic and generated by the elements m such that $m|n$. Thus, for $n = 6$ we find the cyclic subgroups generated by $1, 2, 3, 6$. Those are the subgroups $\mathbb{Z}/6\mathbb{Z}, \{0,2,4\}, \{0,3\}, \{0\}$. They are all normal and the quotients are isomorphic respectively to $\{0\}, \mathbb{Z}/2\mathbb{Z}, \mathbb{Z}/3\mathbb{Z}, \mathbb{Z}/6\mathbb{Z}$.

The subgroups of S_3: Those were classified above.

33.5. **Odds and evens.** Let $n \geq 2$ be an integer. One can show that there is a way to assign a sign, ± 1, to any permutation in S_n such that the following properties hold:

- $\mathrm{sgn}(\sigma\tau) = \mathrm{sgn}(\sigma) \cdot \mathrm{sgn}(\tau)$.
- $\mathrm{sgn}((ij)) = -1$ for $i \neq j$.

We do not prove that here, but we shall prove that next term in MATH 251. Note that since any permutation is a product of transpositions, the two properties together determine the sign of any permutation. Here are some examples: $\mathrm{sgn}((12)) = -1, \mathrm{sgn}((123)) = \mathrm{sgn}((13)(12)) = \mathrm{sgn}((13)) \cdot \mathrm{sgn}((12)) = 1, \mathrm{sgn}((1234)) = \mathrm{sgn}((14)(13)(12)) = -1^3 = -1$.

The property $\mathrm{sgn}(\sigma\tau) = \mathrm{sgn}(\sigma) \cdot \mathrm{sgn}(\tau)$ could be phrased as saying that the function

$$\mathrm{sgn} : S_n \longrightarrow \{\pm 1\}$$

is a surjective group homomorphism. We define A_n as the kernel of the homomorphism $\mathbf{sgn}$. It is called the **alternating group** on n letters and its elements are called **even permutations**. The elements of $S_n \setminus A_n$ are called **odd permutations**. The group A_n is a normal subgroup of S_n, being a kernel of a homomorphism. Its cardinality is $n!/2$. Here are some examples:

- $A_2 = \{1\}$;
- $A_3 = \{1, (123), (132)\}$;
- $A_4 = \{1, (12)(34), (13)(24), (14)(23), (123), (132), (234), (243), (124), (142), (134), (143)\}$. (Easy to check those are distinct 12 even permutations, so the list must be equal to A_4).

33.6. **Odds and Ends.**

Example 33.6.1. We prove that **in $\mathbb{Z}/p\mathbb{Z}$ any element is a sum of two squares**.

Clearly this holds for $p = 2$, so we assume $p > 2$. To begin with, $\mathbb{Z}/p\mathbb{Z}^\times = \{1, \ldots, p-1\}$ is a group under multiplication; it has $p-1$ elements. Consider the homomorphism:

$$sq : \mathbb{Z}/p\mathbb{Z}^\times \to \mathbb{Z}/p\mathbb{Z}^\times, \qquad sq(x) = x^2.$$

Let H be its image, a subgroup of $\mathbb{Z}/p\mathbb{Z}^\times$. The kernel of sq is the solutions to $x^2 = 1$, which are precisely ± 1. Note that $1 \neq -1$. It follows that $H \cong \mathbb{Z}/p\mathbb{Z}^\times/\{\pm 1\}$ is a group with $(p-1)/2$ elements consisting precisely of the non-zero congruence classes that are squares. Let $H^* = H \cup \{0\}$; it is a subset of $\mathbb{Z}/p\mathbb{Z}$ with $(p+1)/2$ elements consisting of all squares.

Let $a \in \mathbb{Z}/p\mathbb{Z}$ then the two sets H^* and $a - H^* := \{a - h : h \in H^*\}$ have size $(p+1)/2$ and so must intersect (because $\mathbb{Z}/p\mathbb{Z}$ has $p < 2 \cdot \frac{p+1}{2}$ elements). That is, there are two squares x^2, y^2 such that $a - x^2 = y^2$ and so $a = x^2 + y^2$.

We next tie together the notions of homomorphism and group action.

Lemma 33.6.2. *Let G be a group and T a non-empty set. To give an action of G on T is equivalent to giving a homomorphism $\rho : G \to S_T$.*

Proof. Suppose that we are given an action of G on S. Pick an element $g \in G$. We claim that the function

$$T \to T, \qquad t \mapsto g \star t,$$

is a permutation of T. Indeed, if $gt_1 = gt_2$ then $g^{-1}(gt_1) = g^{-1}(gt_2)$, so $(g^{-1}g)t_1 = (g^{-1}g)t_2$; that is, $et_1 = et_2$ and so $t_1 = t_2$. Also, given $t \in T$ we have $g(g^{-1}t) = (gg^{-1})t = et = t$, showing surjectivity. Let us denote then this function by $\rho(g)$, $\rho(g)t = gt$. We have a function

$$\rho : G \to S_T.$$

We claim that this function is a homomorphism. We need to show that $\rho(g_1g_2) = \rho(g_1) \circ \rho(g_1)$, i.e., that for every $t \in T$ we have $\rho(g_1g_2)(t) = (\rho(g_1) \circ \rho(g_2))(t)$. Indeed, $\rho(g_1g_2)t = (g_1g_2)t = g_1(g_2)t = g_1(\rho(g_2)t) = \rho(g_1)(\rho(g_2)(t)) = (\rho(g_1) \circ \rho(g_2))(t)$.

Conversely, suppose that

$$\rho : G \to S_T$$

is a group homomorphism. Define an action of G on S by

$$g \star t := \rho(g)(t).$$

We claim this is a group action. Since ρ is a homomorphism we have $\rho(e) = \mathrm{Id}_T$ and so $e * t = \rho(e)(t) = \mathrm{Id}_T(t) = t$. Now, $g_1 \star (g_2 \star t) = \rho(g_1)(\rho(g_2)(t)) = (\rho(g_1) \circ \rho(g_2))(t) = \rho(g_1g_2)(t) = (g_1g_2) \star t$. □

34. Exercises

(1) Write that following permutations as a product of disjoint cycles in S_9 and find their order:
   (a) $\sigma\tau^2\sigma$, where $\sigma = (1234)(68)$ and $\tau = (123)(398)(45)$.
   (b) $\sigma\tau\sigma\tau$, $\sigma = (123)$, $\tau = (345)(17)$.
   (c) $\sigma^{-1}\tau\sigma$, $\sigma = (123456789)$, $\tau = (12)(345)(6789)$.
   (d) $\sigma^{-1}\tau\sigma$, $\sigma = (123456789)$, $\tau = (12345)(6789)$.

(2) Which of the following are subgroups of S_4?
   (a) $\{1, (12)(34), (13)(24), (14)(23)\}$.
   (b) $\{1, (1234), (13), (24), (13)(24)\}$.
   (c) $\{1, (423), (432), (42), (43), (23)\}$.
   (d) $\{1, (123), (231), (124), (142)\}$.

(3) (a) Let Q_8 be the set of eight elements $\{\pm 1, \pm i, \pm j, \pm k\}$ in the quaternion ring $\mathbb{H}$ (so $ij = k = -ji$ etc.). Show that Q_8 is a group.
   (b) For each of the groups S_3, D_4, Q_8 do the following:
      (i) Write their multiplication table;
      (ii) Find the order of each element of the group;
      (iii) Find all the subgroups. Which of them are cyclic?

(4) (a) Find the order of the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 1 & 5 & 6 & 2 & 7 & 8 & 4 \end{pmatrix}$ and write it as a product of cycles.
   (b) Find a permutation in S_{12} of order 60. Is there a permutation of larger order in S_{12}?

(5) Let H_1, H_2 be subgroups of a group G. Prove that $H_1 \cap H_2$ is a subgroup of G.

(6) Let G be a group and let H_1, H_2 be subgroups of G. Prove that if $H_1 \cup H_2$ is a subgroup then either $H_1 \subseteq H_2$ or $H_2 \subseteq H_1$.

(7) Let $\mathbb{F}$ be a field. Prove that

$$\mathrm{SL}_2(\mathbb{F}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{F}, ad - bc = 1 \right\}$$

is a group. If $\mathbb{F}$ has q elements, how many elements are in the group $\mathrm{SL}_2(\mathbb{F})$? (You may use the fact that $\mathrm{GL}_2(\mathbb{F})$ is a group, being the units of the ring $M_2(\mathbb{F})$.)

(8) Let $n \geq 2$ be an integer. Let S_n be the group of permutations on n elements $\{1, 2, ..., n\}$. A permutation σ is called a transposition if $\sigma = (i\ j)$ for some $i \neq j$, namely, σ exchanges i and j and leaves the rest of the elements in their places. Prove that every element of S_n is a product of transpositions.
Hint: Reduce to the case of cycles.

(9) (a) Let $n \geq 1$ integer. Show that the order of $a \in \mathbb{Z}/n\mathbb{Z}$, viewed as a group of order n with respect to addition, is $\frac{n}{\gcd(a,n)}$.
   (b) Let $n \geq 3$. Find the order of every element of the dihedral group D_n.

(10) Let $n > 1$ be an integer, relatively prime to 10. Consider the decimal expansion of $1/n$. It is periodic, as we have proven in a previous assignment. Prove that the length of the period is precisely the order of 10 in the group $\mathbb{Z}/n\mathbb{Z}^\times$ (the group of congruence classes relatively prime to n, under multiplication).

For example: $1/3 = 0.33333\ldots$ has period 1 and the order of $10 \pmod 3 = 1 \pmod 3$, with respect to multiplication, is 1. We have $1/7 = 0.142857142857142857142857142857\cdots$, which has period 6; the order of $10 \pmod 7 = 3 \pmod 7$ is 6 as we may check: $3, 3^2 = 9 \equiv 2, 3^3 = 6, 3^4 = 18 \equiv 4, 3^5 = 12 \equiv 5, 3^6 = 15 \equiv 1$.

Hint: how does one calculates the decimal expansion in practice?

(11) (a) Prove that $\mathbb{Z}_2 \times \mathbb{Z}_3 \cong \mathbb{Z}_6$.
(b) Prove that $\mathbb{Z}_6 \ncong S_3$, though both groups have 6 elements.
(c) Prove that the following groups of order 8 are not isomorphic: $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, $\mathbb{Z}_8$, D_4, Q. (Hint: use group properties such as "commutative" , "exists an element of order k", "number of elements of order k",...)

Remark: One can prove that every group of order 6 is isomorphic to $\mathbb{Z}_6$ or S_3. One can prove that every group of order 8 is isomorphic to one in our list.

(12) Let G be a finite group acting on a finite set S. Prove that if $\langle a \rangle = \langle b \rangle$ for two elements $a, b \in G$ then $I(a) = I(b)$, where for any element $g \in G$, $I(g) = |\{s \in S : g \star s = s\}|$. Another common notation for $I(g)$ is $Fix(g)$.

(13) Let p be a prime number. Let G be a finite group of p^r elements. Let S be a finite set having N elements and assume that $(p, N) = 1$. Assume that G acts of S. Prove that G has a fixed point in S. Namely, there exists $s \in S$ such that $g \star s = s$ for every $g \in G$.

(14) Find how many necklaces with 4 Rubies, 5 Sapphires and 3 Diamonds are there, up to the usual identifications.

(15) Find all the homomorphic images of the groups D_4, Q. (Guidance: If G is one of these groups, show that this amounts to classifying the normal subgroups of G and for each such normal subgroup H find an isomorphism of G/H with a group known to us.)

(16) Let R be the ring of matrices

$$\left\{ \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} : a_{ij} \in \mathbb{Z}_2 \right\}.$$

Note that R has $2^6 = 64$ elements. Let G be the group of units of R. In this case it consists of the matrices such that $a_{11} = a_{22} = a_{33} = 1$. Therefore G has 8 elements. Prove that $G \cong D_4$.

(17) Let H be a subgroup of index 2 of a group G. Prove that H is normal in G.

Part 7. Appendices

Appendix A: The Cantor-Bernstein Theorem

In this appendix we prove the Cantor-Bernstein theorem (Theorem 4.0.2).

Theorem 34.0.1. *Let X, Y, be two sets such that $|X| \leq |Y|$ and $|Y| \leq |X|$. Then, $|X| = |Y|$.*

Proof. We divide the proof into three parts. First, given a set X, denote by $\mathscr{P}(X)$ the set whose elements are the subsets of X. The first step is the following "fixed point" lemma.

Lemma 34.0.2. *Let X be a set and let $\varphi\colon \mathscr{P}(X) \to \mathscr{P}(X)$ be a function with the property that for $A_1, A_2 \in \mathscr{P}(X)$, $A_1 \subseteq A_2 \Longrightarrow \varphi(A_1) \subseteq \varphi(A_2)$. We will refer to this as monotonicity. Then, there is $A_0 \in \mathscr{P}(X)$ such that*

$$\varphi(A_0) = A_0.$$

Proof. [Lemma] Consider the set

$$D = \{A \in \mathscr{P}(X) : A \subseteq \varphi(A)\}.$$

Let

$$A_0 = \cup_{A\in D} A.$$

First, monotonicity implies

$$\varphi(A_0) \supseteq \cup_{A\in D}\varphi(A) \supseteq \cup_{A\in D} A = A_0.$$

From this we also conclude that $\varphi(\varphi(A_0)) \supseteq \varphi(A_0)$. Thus, $\varphi(A_0) \in D$ and consequently, from the definition of A_0, $\varphi(A_0) \subseteq A_0$. It follows that $A_0 = \varphi(A_0)$. □

Now, by assumption, there are injective functions

$$f\colon X \to Y, \qquad g\colon Y \to X.$$

Lemma 34.0.3. *There is a subset A_0 of X such that g^{-1} gives a bijection between $X - A_0$ and $Y - f(A_0)$.*

Proof. [Lemma] Define a function $\varphi\colon \mathscr{P}(X) \to \mathscr{P}(X)$ by

$$\varphi(A) = X - g(Y - f(A)).$$

We claim that φ is monotone: If $A_1 \subseteq A_2$ then $f(A_1) \subseteq f(A_2)$ and so $Y - f(A_1) \supseteq Y - f(A_2)$ and the same after applying g. Thus, $\varphi(A_1) \subseteq \varphi(A_2)$. We can thus apply the previous lemma and find a subset A_0 of X such that

$$A_0 = \varphi(A_0) = X - g(Y - f(A_0)).$$

That means that $X - A_0 = g(Y - f(A_0))$ and so that $X - A_0$ is contained in the image of g and g^{-1} gives a bijection $X - A_0 \to Y - f(A_0)$. □

We now proceed to the last part of the proof where we construction a bijection

$$h\colon X \to Y.$$

We define

$$h(x) = \begin{cases} f(x) & x \in A_0, \\ g^{-1}(x) & x \in X - A_0. \end{cases}$$

Note that h is injective when restricted to A_0 and when restricted to $X - A_0$. At the same time $h(A_0) = f(A_0)$ is disjoint from $h(X - A_0) = Y - f(A_0)$. Thus, h is injective. On the other hand, as $Y = f(A_0) \cup (Y - f(A_0))$, h is also surjective. □

Appendix B: The irrationality of e

In this section we prove that e is irrational. Every rational number has the property that its decimal expansion is eventually periodic and so one idea could be to prove that e doesn't have this property. The problem though is that we know very little about the expansion of e. For example, we do not know if every digit appears infinitely often in its expansion. Instead, we shall use a different method. It is based on the fact that e has an expression as an infinite sum that "converges too fast".

We make use of the formula, or, in certain approach, the definition, of e as an infinite sum:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \cdots \approx 2.7182818\ldots.$$

Suppose that e is a rational number then $e = a/N$, where a and N are positive integers, $N > 1$ and so $N! \cdot e$ is surely an integer, as is $N! \cdot e - N! \sum_{n=0}^{N} \frac{1}{n!}$. Therefore,

$$N! \cdot \sum_{n=N+1}^{\infty} = \frac{1}{N+1} + \frac{1}{(N+1)(N+2)} + \frac{1}{(N+1)(N+2)(N+3)} + \ldots$$

is an integer. This number is clearly positive. On the other hand, it is smaller than

$$\frac{1}{N+1} + \frac{1}{(N+1)^2} + \frac{1}{(N+1)^3} + \cdots = \frac{1}{N} < 1.$$

As there is no positive integer small than 1, we have arrived at a contradiction and so e is irrational.

Appendix C: Euclidean rings

Euclidean rings are a class of rings where division with residue is possible. They unite together our two examples of such rings: the integers $\mathbb{Z}$ and rings of polynomials over a field $\mathbb{F}[x]$. Moreover, there are other interesting examples of Euclidean rings and in such rings many of the results we had obtained for $\mathbb{Z}$ and $\mathbb{F}[x]$ will hold, but we shall only demonstrate a few of those.

Let R be a commutative ring which is an integral domain. That means that R has no zero divisors; that is for two elements $x, y \in R$, $xy = 0$ implies $x = 0$ or $y = 0$. Suppose that there is a function

$$|\cdot| : R - \{0\} \to \mathbb{N},$$

such that for all x, y in R with $y \neq 0$ there are elements $q, r \in R$ such that $x = qy + r$ and either $r = 0$ or $|r| < |y|$. (In a general Euclidean ring q, r need not be unique.) Then R is called a **Euclidean ring**.

Let R be a Euclidean ring and a, b non-zero elements of R. We may perform the Euclidean algorithm for a, b in R as follows:

$$\begin{aligned}
a &= bq_0 + r_0, && |r_0| < |b|,\\
b &= r_0 q_1 + r_1, && |r_1| < |r_0|,\\
r_0 &= r_1 q_2 + r_2, && |r_2| < |r_1|,\\
&\vdots
\end{aligned}$$

For some t we must first get that $r_{t+1} = 0$. That is,

$$\begin{aligned}
r_{t-2} &= r_{t-1} q_t + r_t, && |r_t| < |r_{t-1}|\\
r_{t-1} &= r_t q_{t+1}.
\end{aligned}$$

As said, this process of repeated division with residue much stop as the sizes $|r_i|$ of the residues are decreasing natural numbers. Then r_t is the gcd of a and b. That is, r_t divides both a and b and any element of R dividing both a and b divides r. Note though that r_t is not uniquely determined by these properties as for every unit u of R also ur_t would have these properties (also, in each step of the devision with residue we have choose q_i and that effects the residues as well). In the case of $\mathbb{Z}$ we could make the gcd unique by demanding that it is positive, and in the case of $\mathbb{F}[x]$ we could make it unique by requiring it to be a monic polynomial, but

for a general ring R there are no such normalizations. And so, we should really say "a gcd" of a,b and not "the gcd" of a,b. As in the case of $\mathbb{Z}$ or $\mathbb{F}[x]$ we can prove:

Proposition 34.0.4. *R is a principal ideal domain. Let $a,b \in R$ not both zero. Let r be a generator of the ideal of R generated by a and b, $\langle r\rangle = \langle a,b\rangle$. Then r is a gcd of a and b.*

Proof. Let us show first that R is a principal ideal domain. Let I be a non-zero ideal of R and consider the non-empty set of natural numbers

$$\{|a| : a \in I - \{0\}\}.$$

This set has a minimal element, necessarily of the form $|a_0|$ for some element $a_0 \in I, a_0 \neq 0$. Clearly $\langle a_0\rangle \subseteq I$. Let $x \in I$ and write $x = qa_0 + r$ where either $r = 0$ or $|r| < |a_0|$. As $r = x - qa_0 \in I$, we cannot have $|r| < |a_0|$ and so we must have $r = 0$. That is, $x \in \langle a_0\rangle$ and we got $I \subseteq \langle a_0\rangle$. Therefore, $I = \langle a_0\rangle$ and is principal.

Assume that b is not zero and let r be a gcd of a,b obtain by the Euclidean algorithm. It follows from the algorithm that $r = xa + yb$ for some $x,y \in R$ and so $r \in \langle a,b\rangle$ and consequently $\langle r\rangle \subseteq \langle a,b\rangle$. On the other hand, as $r|a$ we have $a \in \langle r\rangle$ and similarly $b \in \langle r\rangle$ and it follows that $\langle a,b\rangle \subseteq \langle r\rangle$. That is, $\langle r\rangle = \langle a,b\rangle$, where r is a gcd of a and b. Remark that the generators of the ideal $\langle r\rangle$ are precisely the gcd's of a,b (they are the elements $\{ur : u \in R^\times\}$). □

The next step is to show that for $r \in R$, $r \neq 0, r \notin R^\times$, the following properties are equivalent: (i) if $r|ab$ then $r|a$ or $r|b$; (ii) if $r = ab$ then either a or b are units. Such an element is called a prime element of R. A proof very similar to those for $\mathbb{Z}$ or $\mathbb{F}[x]$ gives us unique factorization:

Theorem 34.0.5. *Let R be a Euclidean ring and $a \neq 0$ an element of R. Then, there are non-associated prime elements $p_1,\dots,p_t$ of R and positive integers $a_1,\dots,a_t$ and a unit u such that*

$$a = up_1^{a_1}\dots p_t^{a_t}.$$

Moreover, the factorization is unique in the sense that if $a = vq_1^{b_1}\cdots q_s^{b_s}$, with v a unit, q_i are non-associated prime elements and b_i positive integers, then possibly after renaming the q_i we have $s = t$, $p_i \sim q_i$ for all i and $a_i = b_i$ for all i.

Here is an interesting example of a Euclidean ring, called the ring of **Gaussian integers** . It is the ring $\mathbb{Z}[i] = \{x + yi : x,y \in \mathbb{Z}\}$, where

$$|x + yi| = x^2 + y^2.$$

Given $a,b \in \mathbb{Z}[i]$ such that $b \neq 0$, write using division in $\mathbb{C}$, $a/b = s + ti$, where s,t are real numbers (in fact they come out rational numbers) and let $q = s_0 + t_0 i$, where s_0 is the integer nearest to s and t_0 the integer nearest to t. There is a unique $r \in \mathbb{Z}[i]$ such that the equation

$$a = qb + r$$

holds; that is, define $r = a - qb$. One then verifies that

$$|r| < |b|,$$

and so we get that $\mathbb{Z}[i]$ is Euclidean. It is interesting that if you try to mimic this argument of $R = \mathbb{Z}[\sqrt{-5}]$ and $|x + y\sqrt{-5}| = x^2 + 5y^2$ the proof doesn't work. We know that because we have seen that R is not a principal ideal domain (see page 67). It is interesting to analyze the exact point where the proof fails, but we leave that to the interested reader.

Appendix D: The complex exponential function

Consider the series

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z}{2!} + \frac{z^3}{3!} + \dots$$

Lemma 34.0.6. *For every complex number z the series converges to a complex number that we shall call e^z.*

Proof. Left as an exercise. The key is to show that for every for every real number $\epsilon > 0$ and $z \in \mathbb{C}$ there is an integer $n \geq 0$ such that for all integers $N \geq 0$,

$$|\sum_{k=n}^{n+N} z^k/k!\,| < \epsilon.$$

This implies that the real part and imaginary part of this sum are both less then ϵ and this, in turn, implies that the series $\mathrm{Re}(\sum_{k=1}^{N} z^k/k!)$, when N varies, and the series $\mathrm{Im}(\sum_{k=1}^{N} z^k/k!)$, when N varies, are Cauchy series of real numbers that thus converge. □

Theorem 34.0.7. *The complex exponential function e^z satisfies*

$$e^{z_1+z_2} = e^{z_1}e^{z_2}.$$

Proof. As formal power series we have

$$\begin{aligned} e^{z_1+z_2} &= \sum_{n=0}^{\infty} \frac{(z_1+z_2)^n}{n!} \\ &= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \frac{\binom{n}{j} z_1^j z_2^{n-j}}{n!} \\ &= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \frac{z_1^j}{j!} \frac{z_2^{n-j}}{(n-j)!} \\ &= (\sum_{n=0}^{\infty} \frac{z_1^n}{n!})(\sum_{n=0}^{\infty} \frac{z_2^n}{n!}). \end{aligned}$$

To justify that as equality of complex numbers we should work with finite sums $\sum_{n=0}^{2M} \frac{(z_1+z_2)^n}{n!}$ where we get equalities up to the last step where we don't quite get $(\sum_{n=0}^{2M} \frac{z_1^n}{n!})(\sum_{n=0}^{2M} \frac{z_2^n}{n!})$, but a sub-sum of this containing all terms appearing in $(\sum_{n=0}^{M} \frac{z_1^n}{n!})(\sum_{n=0}^{M} \frac{z_2^n}{n!})$. However, it is easy to show that the two sums differ by a quantity that goes to zero as M goes to infinity. We leave the details to the reader. □

Lemma 34.0.8 (Euler's formula)**.** *Let θ be a real number then*

$$e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

Remark 34.0.9. In the text, where we discussed complex numbers, we have defined $e^{i\theta}$ this way as a quick and dirty method to get a function e^z with good properties. We now show that this formula follows from the more systematic approach of defining e^z via a power series, valid for either real or complex number z.

Proof. We have

$$\begin{aligned} e^{i\theta} &= \sum_{n=0}^{\infty} \frac{(i\theta)^n}{n!} \\ &= \sum_{n=0}^{\infty} i^{2n} \frac{\theta^{2n}}{(2n)!} + i\sum_{n=0}^{\infty} i^{2n} \frac{\theta^{2n+1}}{(2n+1)!} \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{\theta^{2n}}{(2n)!} + i\sum_{n=0}^{\infty} (-1)^n \frac{\theta^{2n+1}}{(2n+1)!} \\ &= \cos(\theta) + i\sin(\theta). \end{aligned}$$

We have assumed here familiarity with the Taylor series expansion for $\sin$ and $\cos$. □

Corollary 34.0.10 (de Moivre's formula)**.** *Let θ be a real number.*

$$(\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta).$$

Proof. The left hand side is equal to $(e^{i\theta})^n$, while the right hand side is equal to $e^{in\theta}$. The theorem implies these are equal. □

Example 34.0.11. By expanding we find trigonometric formulas, and this is the easiest way to know them, without memorizing them!

$$\cos(2\theta) + i\sin(2\theta) = (\cos(\theta) + i\sin(\theta))^2 = \cos(\theta)^2 - \sin(\theta)^2 + i2\cos(\theta)\sin(\theta).$$

By separating real and imaginary parts, we find:

$$\cos(2\theta) = \cos(\theta)^2 - \sin(\theta)^2, \quad \sin(2\theta) = 2\cos(\theta)\sin(\theta).$$

Example 34.0.12. As $e^{i\theta} \cdot e^{-i\theta} = 1$ we find that $(\cos(\theta) + i\sin(\theta))(\cos(-\theta) + i\sin(-\theta)) = (\cos(\theta) + i\sin(\theta))(\cos(\theta) - i\sin(\theta))$ ($\cos$ is an even function and $\sin$ is an odd function; this is well known but is also evident from their Taylor expansions). This gives the identity

$$\cos(\theta)^2 + \sin(\theta)^2 = 1.$$

Index

(r), 66
$(r_1, r_2, \ldots, r_n)$, 67
D_n, 90
$Fix(g)$, 98
I_n, 64
$M_n(\mathbb{F})$, 64
R/I, 72
$R[x]$, 52
S_n, 17, 87
$\mathbb{C}$, 3
$\mathbb{F}[\epsilon]$, 64
$\mathbb{F}_p$, 46
$\mathrm{Im}(z)$, 19
$\mathbb{N}$, 3
$\mathbb{N}^+$, 19
$\mathbb{Q}$, 3
$\mathbb{R}$, 3
$\mathrm{Re}(z)$, 19
$\Rightarrow$, 6
$\mathbb{Z}$, 3
$\mathbb{Z}/n\mathbb{Z}$, 44
$\mathbb{Z}/n\mathbb{Z}^\times$, 85
$\bar{a}$, 73
$\bar{z}$, 20
$\cap$, 4
$\cap_{i\in I}$, 4
$\cong$, 76, 95
$\cup$, 4
$\cup_{i\in I}$, 4
$\emptyset$, 4
$\equiv \pmod{n}$, 44
$\exists$, 11
$\forall$, 4, 11
$\gcd(a,b)$, 34
$\in$, 3
$\langle g \rangle$, 87
$\langle r_1, r_2, \ldots, r_n \rangle$, 67
$\mapsto$, 23
μ_n, 86
$\neg$, 7
$\lhd$, 65
$\notin$, 3
$\setminus$, 4
$\sim$, 16
$\square$, 6
$\subset$, 4
$\subseteq$, 4
$\mathrm{Orb}(s)$, 96
$\mathrm{Stab}(s)$, 96
sgn, 107
$\times$, 5
φ, 93
$\{\ \}$, 3
$a|b$, 34
e, 85
$e^{i\theta}$, 22
$o(g)$, 88
43

Archimedes Cattle problem, 70
associate
  in a ring, 81
  polynomials, 56

binomial
  coefficient, 46, 51
  theorem, 46, 51
Burnside's formula, 98

Cantor's diagonal argument, 16
cardinality, 13
Carmichael number, 47
Cauchy-Frobenius formula, 98
Cayley's theorem, 95
Chinese Remainder Theorem, 77, 94
complete set of representatives, 18, 44, 91
complex conjugate, 20
congruence, 44
Continuum hypothesis, 16
coset, 72, 91
cycle, 88
  disjoint, 89

de Moivre's formula, 114
Dedekind, 4
degree, 52
determinant, 71
direct product
  ring, 64
division, 34
  with residue, 33
division with residue, 53

Eisenstein's criterion, 60
equivalence
  class, 18
  relation, 17
Euclid, 35, 38
Euclidean algorithm, 35, 55
Euler's constant, 41
Euler's formula, 114
Euler's totient function, 93

Fermat, 46
  little theorem, 46, 51
fibre, 12
field, 26
  algebraically closed, 59
First isomorphism theorem, 76, 104
function, 10
  bijective, 12
  composition, 12
  graph, 11
  identity, 11
  image, 11
  injective (one-one), 12
  inverse, 13
  multiplicative, 94
  source, domain, 11
  sujective (onto), 12
  target, codomain, 11
Fundamental Theorem of Algebra, 23
Fundamental Theorem of Arithmetic, 37

Gauss, 23, 38
Gaussian integers, 113
generator, 87
Goldach's conjecture, 39
graph, 10

finite, 10
simple, 10
greatest common divisor, 59
greatest common divisor (gcd), 34, 39, 54, 59
group, 85
n-th roots of unity, 86
abelian, 85
action, 96
transitive, 99
cyclic, 87
dihedral, 90
homomorphism, 94
isomorhism, 95
order, 88
quotient, 104
symmetric, 87
trivial, 85
units, 85
wallpaper, 103

Hamilton, 82
homomorphism, 68, 94
isomorphism, 76
kernel, 68
specialization/evaluation, 69

ideal, 65
generated, 82
maximal, 80
non-principal, 67
prime, 80
principal, 66
sum, 67
trivial, 65
index, 92
induction, 7
integral domain, 53
isomorphism, 76, 95
ring, 76

kernel, 68, 94

Lagrange's theorem, 92

necklace, 98
number
complex, 3, 19
integer, 3
irrational, 40
natural, 3
prime, 36
rational, 19
real, 3

operation, 25
orbit, 96
order
element, 88
group, 88
linear, simple, total, 17
partial, 17

Peano, 4
Pell equation, 70
permutation, 17, 87
even, 107
odd, 107
pigeonhole principle, 9
polar representation, 21
polynomial
associate, 56
complex, 23
constant, 52
degree, 52
irreducible, 57
monic, 52
rational, 23
real, 23
zero, 23, 52
pre-image, 12
prime
number, 36
prime element, 81
Prime Number Theorem, 38
principal ideal domain, 81
principal ideal ring, 66
proof
by contradiction, 7
contrapositive, 7
induction, 7
pigeonhole, 9
prove or disprove, 9

quaternions, 82
quotient, 33
quotient group, 104
quotient ring, 72

relation, 16
congruence, 44
equivalence, 17
complete set of representatives, 18
reflexive, 17
symmetric, 17
transitive, 17
residue, 33
ring, 26, 63
commutative, 26
direct product, 64
division, 26, 63
dual numbers, 64
Euclidean, 53, 112
homomorphism, 68
integral domain, 53, 112
isomorphism, 76
matrices, 63, 64
polynomial, 52
quaternion, 82
quotient, 72
skew-field, 63
subring, 27, 65
root, 23
of unity, 22
roulette, 98
RSA, 48

set, 3
contain, 4
countable, 14
difference, 4
disjoint union, 18
equal, 4
intersection, 4
product, 5

union, 4
sieve of Eratosthenes, 36
sign (of a permutation), 107
skew-field, 63
stabilizer, 96
subgroup, 86
normal, 104
subring, 27

Theorem
Eisenstein's Criterion, 60
Euclidean algorithm, 35
Fundamental Theorem of Algebra, 23, 59
Fundamental Theorem of Arithmetic, 37
Prime Number, 38
transposition, 88
Twin Prime conjecture, 39

unique factorization
integers, 37
polynomials, 57
rationals, 40
unit, 66, 70

wallpaper
group, 103
pattern, 103
Wilson's theorem, 51, 62

zero, 23
zero divisor, 45