Algebra I – MATH235

Course Notes by Dr. Eyal Goren

McGill University

Fall 2009

Last updated: September 16, 2013.

©Prof. Eyal Goren, Department of Mathematics and Statistics, McGill University.

Contents

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. There is a growing side to Algebra which is purely theoretic at present. It is important to realize though that some of the most abstract algebraic structures of the past, such as Galois fields (called nowadays just finite field), have become in modern days the cornerstone of applied algebra, for example in applications to coding theory and cryptography.

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). 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 recipes for 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]

MATH 235 is a first course in Algebra. Little is assumed in 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. A further challenge for most students is that the key ideas we learn in this course are very abstract, bordering on philosophy and art, but yet they are truly scientific in their precision. You should expect not to understand everything right away; you should expect to need time to reflect on the meaning of the new ideas and concepts we introduce. Here are some pointers as to how to cope with the challenges of this course:

- Read the class notes and the text book 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.
- Explain to your friends, and possibly to your family, the material of the course. Work together with your class mates on assignments, but write your own solutions by yourself in the end; try to understand different solutions to assignments and try and find flaws in your class mates 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.

So what is this course really about?

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

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 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 a seemingly esoteric beast initially, 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 publicity, to display all necklaces one can make using 10 diamonds and 10 rubies, perhaps under the (banal) slogan "to each her 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. 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. Turns out there are 4752 such different designs; that will require 47520 diamonds and 47520 rubies. Perhaps the idea should be reconsidered ;-)

Acknowledgments: Many thanks to all the students who have pointed out typos, in particular to Eric Beaudoin.

Part 1. **Some Language and Notation of Mathematics**

1. Sets

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 of 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).

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 $1, 3, 5$. 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 to "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*. The use of the letters $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ (and $\mathbb{R}, \mathbb{C}$ below) is standard. Other standard notation is

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

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 §5). 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, $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$ is saying $\mathbb{C}$ is the set whose elements are $a+bi$, where a and b are real numbers. For example, in the notation above, $3 \in S, 2 \notin S$, Jim's football $\in T$ but $\notin U^+$.

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 given in a logic course (usually for $\mathbb{N}$) or in an analysis course for $\mathbb{R}$. 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.

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). 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 don't 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 <u>is</u> a set but it has no elements. Admittedly, this 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

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

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

We shall also need arbitrary unions and intersections. Let I be a 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

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

Then

$$\cup_{0\leq x\leq 1} S_x$$

is the triangle 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 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 $A=B$.

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

diagrams for $A\cap B, A\cup B, A-B,\ A\cap B\cap C,\ A\cup B\cup C-(A\cap B\cap C)=(A-C)\cup(B-A)\cup(C-B)$

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

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

In particular,

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

Example 1.0.1. 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 1.0.2. 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).

1.1. We now wish to make a general statements relating some of these operations. Once a statement is important enough to highlight it, it falls under the heading of Lemma, Proposition or Theorem (or, more colloquially, Claim, 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.

Proposition 1.1.1. *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

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

2. Proofs: idea and technique

Proposition 1.1.1 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, the French word for 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 1.1.1 is by a standard technique.

2.1. **Proving equality by two inequalities.**

When one wants to show 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", $A \subset B$ and $B \subset A$. 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, test, or a research paper is:

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*.)

Lets 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). □

Here is another example.

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 negations. Logically the implication

$$A \Longrightarrow B$$

is equivalent

$$\neg B \Longrightarrow \neg A.$$

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 $A \Longrightarrow B$ 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 that either x or y is even. (Note that in mathematics "or" always means "and/or".)

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) 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 $n \geq n_0$.

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 $n_0 + 1, n_0 + 2, n_o + 3, \dots$. 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.

Include a picture

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

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

(The statement is P_n is $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ and 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

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

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,

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

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 $n \geq 1$. (iii) Consider the following scenario: there are n persons in a party and on the ring of the bell each is supposed to throw a pie at the person closest to him. (Well, things were getting wild...) Assume that all the distances 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, see if it is similar to other statements you know to hold true, or, when at lost, throw a coin. After deciding on your initial position, if you believe the statement 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 to provide *one* counterexample. The statement being false doesn't mean it's false for *every n*; it means it's false for *one n*. In the case in hand, if we take $n = 5$ we find that $4*5+1 = 21$ 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 deceptively simple, given how useful it is. It states the following:

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

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 residues of a, b, c, d, e, f upon division by 5. The residue is either $0, 1, 2, 3$ or 4 (these are the 5 pigeonholes) but we get from our numbers 6 residues (those are the pigeons). Therefore, there are two numbers among a, b, c, d, e, f with the same residue. Say, a and b have the same residue, 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. *In any group of $n \geq 2$ 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 $N+1$ numbers 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 proved. 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 have 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 know each other.)

3. Functions

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,

$$f : A \longrightarrow B,$$

is a rule assigning to *each* element of A a *unique* element of B. The set A is called the *source*, or the *domain*, of the function, and B the *target* 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 3.0.3. The simplest example is the *identity function*. Let A be any set and define

$$1_A : A \to A$$

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

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

for any $x \in A$.

Example 3.0.4. Let $A = \{1, 2, 3\}$, $B = \{1, 2\}$ and consider the following rules for $f : A \longrightarrow B$.

(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 3.0.5. Consider

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

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 values $f(0)$ 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 $a, f(a), f(f(a)), f(f(f(a))), \dots$). Since a mother is older from any of her children by at least, say one year (just not to worry about an exact number), it follows, if A and f are well defined, that people have existed forever. The various ways to resolve the paradox, namely to provide an explanation as to *why* this is not well-defined, are rather amazing and I leave it to you as an amusing exercise.

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 3.0.6. Let A be a set and $\Gamma \subset A$ 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, inverse image.** We introduce some attributes of functions. Let

$$f : A \to B$$

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 $f : A \to B$ 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) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Then f is neither surjective (a square is always non-negative) nor injective ($f(x) = f(-x)$).

We have $f^{-1}([1,4]) = [1,2] \cup [-2,-1]$ and $f^{-1}(0) = \{0\}$, $f^{-1}(-1) = \emptyset$.

(2) $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$, $f(x) = x^2$. Then f is surjective but not injective.

(3) $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $f(x) = x^2$. Then f is bijective.

3.2. **Composition of functions.** Let

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

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

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

The picture is

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[rr, bend left, "g\circ f"] & B \arrow[r, "g"] & C \end{tikzcd}$$

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 $g(b) = g(f(a)) = (g \circ f)(a) = c$. □

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

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

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 has revolutionized mathematics, 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

$$f : A \to B.$$

(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 $|A| = |B|$.*

Although it does not require any sophisticated mathematics, the proof will not be given in this course. It is an ingenious and intricate proof, but requires no more background than we already have. 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. We should conclude from that there is some bijective function h from A to B. It is not true that $h = f$ or g. One should somehow construct it 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" than to deduce it from f and g).

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. 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 4.0.3. 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 $|A| = |B|$ (though it is not hard to write a bijection from A to B either).

Example 4.0.4. 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 a n is 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 4.0.5. $|\mathbb{N}| = |\mathbb{Z}|$.

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 4.0.6. $|\mathbb{N}| = |\mathbb{N} \times \mathbb{N}|$.

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

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

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 4.0.7. $|\mathbb{Z}| = |\mathbb{Z} \times \mathbb{Z}|$.

Proof. Let $h : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ be a bijection. A bijection $f : \mathbb{N} \to \mathbb{Z}$ induces a bijection $g = (f, f) : \mathbb{N} \times \mathbb{N} \to \mathbb{Z} \times \mathbb{Z}$ and the composition

$$\begin{tikzcd} \mathbb{Z} \arrow[r, "f^{-1}"] & \mathbb{N} \arrow[r, "h"] & \mathbb{N} \times \mathbb{N} \arrow[r, "g"] & \mathbb{Z} \times \mathbb{Z} \end{tikzcd} ,$$

is also a bijection. □

Exercise 4.0.8. 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:

Theorem 4.0.9. $|\mathbb{N}| \neq |\mathbb{R}|$.

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 $a_0, a_1, a_2, \dots$. Let us write the decimal expression of each number as

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

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

Now consider the number

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

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 $a_0, a_1, a_2, \dots$ cannot consist of *all* real numbers and so we arrive at a contradiction. □

Remark 4.0.10. 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. The discoveries were nothing short of shocking. They came at the time where humanity was fascinated with its own power, especially the power of the intellect. These results proved the inherent limitations of thought.

5. 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* 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

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

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 *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:

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

(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 prior interpretation in the model of the plane. 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 (or just negative, for that matter). 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 5.0.11. *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:*

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

Here a figure with $z = a + bi$ *in plane model,* $|z|$ *marked on the line connecting 0 to z,* $\bar{z}$ *plotted as well.*

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

We have $\overline{z} = a - bi$ and so $\overline{\overline{z}} = a + bi = z$. 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 5.0.12. 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 $1 + 2i + \frac{3-i}{1+5i} = \frac{12}{13} + \frac{18}{13}i$.

5.1. **The polar representation.** Considering $z = a + bi$ 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.

Here a figure

We have

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

Lemma 5.1.1. *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°. □

5.2. **The complex exponential function.** Let θ be any real number. Let $e^{i\theta}$ denote the unit vector whose angle is θ. That is, that complex number with 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 have proven 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

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

(Conversely, the identity $e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$ implies the formula for multiplication). 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. 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.

Example 5.2.1. 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 $z^n = R^ne^{in\Theta} = re^{i\theta}$. That is, if and only if

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

Thus, the solutions are exactly

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

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

figure of 3, 4 and 5-th roots of 1

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

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0,$$

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

$$f : \mathbb{C} \to \mathbb{C}, \quad z \mapsto f(z) = a_nz^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0.$$

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

Example 5.3.1. 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 5.2.1.)

Theorem 5.3.2. *(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 Honours Algebra 4 MATH371 one sees an algebraic proof; in Complex Variables and Transforms MATH 381 one sees an analytic proof. The theorem is attributed to Gauss who has proven it in 1799. There is some issues regarding the completeness of that proof; he later own 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 5.3.3. *Let $f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ 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 one (this is a convention: the empty product is equal to one, the empty sum is equal to zero.) 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

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

where $g(x)$ is a polynomial of degree $n-1$ and leading coefficient a_n and r is a complex number. Indeed, write $g(x) = b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$ 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.

Now, apply that for $z = z_n$. 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 $|z_1 z_2 \cdots z_a| = |z_1| \cdot |z_2| \cdots |z_a|$). 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. 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. But we may re-index the z_i so that $t_1 = z_1$. Dividing both sides by $x - z_1$ we then conclude that[2]

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

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

We remark that for $n = 1, 2$ the result is well known:

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

[2] 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 $f(x)/(x - z_1) = a_n \prod_{i=2}^{n}(x - z_i)$.

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

There are also formulas for the roots for polynomials of degree 3 and 4, but in degrees 5 and higher no such formulas exist. Not that they are merely not known; they cannot exist. This follows from Galois theory, taught in MATH371.

6. 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 reconsider previous examples. Later we shall enter a systematic development of the theory.

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

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

That is, it 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 6.0.4. 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. Not all authors insist on that, though.

Definition 6.0.5. Note that the multiplication 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 6.0.6. $\mathbb{Z}$ is a commutative ring. It is not a division ring and so it is not a field.

Example 6.0.7. The rational numbers $\mathbb{Q}$ form a field. The real numbers $\mathbb{R}$ form a field. In both those cases we are forced to assume the properties 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 6.0.8. *$\mathbb{C}$ is a field.*

Proof. Let $z_1 = a_1 + b_1 i, z_2 = a_2 + b_2 i, z_3 = a_3 + b_3 i$. 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. $(z_1 + z_2) + z_3 = [(a_1 + a_2) + (b_1 + b_2)i] + a_3 + b_3 i = [(a_1 + a_2) + a_3] + [(b_1 + b_2) + b_3]i = [a_1 + (a_2 + a_3)] + [b_1 + (b_2 + b_3)]i = z_1 + [(a_2 + a_3) + (b_2 + b_3)i] = z_1 + (z_2 + z_3)$.
3. Clearly $0 + z_1 = z_1$.
4. We have $(-a_1 - b_1 i) + (a_1 + b_1 i) = (-a_1 + a_1) + (-b_1 + b_1)i = 0 + 0i = 0$.
5. $(z_1 z_2)z_3 = [(a_1 + b_1 i)(a_2 + b_2 i)](a_3 + b_3 i) = ((a_1 a_2 - b_1 b_2) + (a_1 b_2 + b_1 a_2)i)(a_3 + b_3 i) = (a_1 a_2 - b_1 b_2)a_3 - (a_1 b_2 + b_1 a_2)b_3 + ((a_1 a_2 - b_1 b_2)b_3 + (a_1 b_2 + b_1 a_2)a_3)i = a_1 a_2 a_3 - b_1 b_2 a_3 - a_1 b_2 b_3 - b_1 a_2 b_3 + (a_1 a_2 b_3 - b_1 b_2 b_3 + a_1 b_2 a_3 + b_1 a_2 a_3)i$. One now develops the product $z_1(z_2 z_3)$ 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. $z_1(z_2 + z_3) = (a_1 + b_1 i)((a_2 + a_3) + (b_2 + b_3)i) = a_1(a_2 + a_3) - b_1(b_2 + b_3) + (b_1(a_2 + a_3) + a_1(b_2 + b_3)i) = (a_1 a_2 - b_1 b_2) + (b_1 a_2 + a_1 b_2)i + (a_1 a_3 - b_1 b_3) + (b_1 a_3 + a_1 b_3)i = (a_1 + b_1 i)(a_2 + b_2 i) + (a_1 + b_1 i)(a_3 + b_3 i) = z_1 z_2 + z_1 z_3$.

Before proving the next property, we check that $z_1 z_2 = z_2 z_1$. We have $z_1 z_2 = (a_1 + b_1 i)\ (a_2 + b_2 i) = a_1 a_2 - b_1 b_2 + (a_1 b_2 + b_1 a_2)i = a_2 a_1 - b_2 b_1 + (a_2 b_1 + b_2 a_1)i = (a_2 + b_2 i)(a_1 + b_1 i) = z_2 z_1$. In particular, $(z_1 + z_2)z_3 = z_3(z_1 + z_2) = z_3 z_1 + z_3 z_2 = z_1 z_3 + z_2 z_3$.

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 6.0.9. Here is an example 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 tiring 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.

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

6.1. **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 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 $(x+y)+(-x-y) = 0$. But, $-x+x = x+(-x) = 0$ and $(x+y)+(-x-y) = x+(-x)+y+(-y) = 0+0 = 0$.

Although this proof is simple and short, it's 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 $-(-x) = x$ is 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 6.1.1. 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 $a + b, -a, ab \in S$.

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 that of R.

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

Example 6.1.2. 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 6.1.3. 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 10.1.4 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

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

7. 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 $A_1, A_2, A_3, \dots$ 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 $\cup_{i \in \{1,2,3,\dots\}} A_i$.

(a) Let $N \geq 0$ 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 $B_n = \{(x, y) \in \mathbb{R}^2 : y^n \leq x\}$. What is $\cup_{n=1}^{\infty} B_n$? $\cap_{n=1}^{\infty} B_n$?

(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 r.h.s. we proved in class.

(8) Prove by induction that $2^n > n^2$ for $n \geq 5$. It is false for $n \geq 0$ (because it fails for $n = 2$); 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 guess an expression for the value of C_n. Prove your formula by induction.

(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 $f : A \to B$ 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.

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

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

Count them using diagonals as follows:

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

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 $f(0,n)$).

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

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

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

(17) Let A be a set. Then $|A| < |2^A|$, where 2^A is the set of all subsets of 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 $|A| < |2^A|$.

(18) (Planned exercise: guided proof of the Cantor-Bernstein theorem)

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

(20) 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.

(21) 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).

(22) 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) $z^2 = \bar{z}$. Also, in each of these cases, plot the answer on the complex plane.

(23) *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 $\mathbb{F}$.

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, but in this course we shall not discuss them beyond providing some margin notes.

8. Division

Theorem 8.0.4. *(Division with residue)*[3] *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.

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 $a > 0$ 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 $q' > q$. We then get

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

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. $\square$

[3] A *commutative ring* is called an integral domain if $xy = 0$ implies $x = 0$ or $y = 0$. An integral domain R is called a *Euclidean ring* if 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 $|r| < |x|$. (But in a general Euclidean ring q, r are not unique.) Thus, the theorem we are proving is that $\mathbb{Z}$ *is a Euclidean ring*.

9. GCD and the Euclidean algorithm

Definition 9.0.5. 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) $a|b, a|d \Rightarrow a|(b \pm d)$.

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 $a|(b \pm d)$. □

Remark 9.0.6. 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 9.0.7. Which are the numbers dividing 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 $a = \pm 1, \pm 5$.

Corollary 9.0.8. *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 $r = 0$. □

9.1. GCD.

Definition 9.1.1. 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 9.1.2. *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 9.1.3. *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 $a|uac + vbc = c$. □

9.2. **The Euclidean algorithm.** 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 9.2.1. *(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:*

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

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

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

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

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

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

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

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

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 $d|r_i, i = 0, 1, 2, \dots, n$. 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 $r_t, r_{t-1}, \dots, r_n$. Then, since $r_{n-1} = r_nq_{n+1} + r_{n+1}$ we also get $r_t|r_{n-1}$. Therefore, r_t divides $r_0, r_1, \dots, r_t$. Again, $b = r_0q_1 + r_1$ and so $r_t|b$ and then $a = bq_0 + r_0$ and so $r_t|a$. □

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,

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

and so $\gcd(113, 54) = 1$. We have $1 = 5 - 4 \cdot 1$, and we keep substituting for the residues we now have, expressions using 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

Definition 10.0.2. 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. A positive integer is prime if its only positive divisors are 1 and p.

The sieve of Eratosthenes: [4] 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, $\underline{4}$, 5, $\underline{6}$, 7, $\underline{8}$, 9, $\underline{10}$, 11, $\underline{12}$, 13, $\underline{14}$, 15, $\underline{16}$, 17, $\underline{18}$, 19, $\underline{20}$, 21, $\underline{22}$, 23, $\underline{24}$, 25, $\underline{26}$, 27, $\underline{28}$, 29, $\underline{30}$, 31, $\underline{32}$, 33, $\underline{34}$, 35, $\underline{36}$, 37, $\underline{38}$, 39, $\underline{40}$, 41, $\underline{42}$, 43, $\underline{44}$, 45, $\underline{46}$, 47, $\underline{48}$, 49, $\underline{50}$

[4] Eratosthenes of Cyrene, 276BC - 194BC, was a Greek mathematician who 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

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 an underline):

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

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

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 an underline):

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

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 below) and so any number less of equal to N, which is not prime, has a prime divisor smaller or equal to $\sqrt{N}$. Thus, any number left on our list is prime.

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

Theorem 10.0.3. *(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 10.0.4. *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, 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 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 10.0.5. *Let p be a positive 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 9.1.3), $p|b$.

Now suppose that p satisfies (ii). If $p = st$ then $p|st$ 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, and therefore $p = s$. So p has no proper divisors and hence is prime. □

Remark 10.0.6. 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 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. □

We next derive some consequences of the fundamental theorem of arithmetic. The theorem exhibits the prime numbers as the building blocks of the integers. In itself, it doesn't tell us if there are finitely or infinitely many of them.

Theorem 10.0.7. *(Euclid[5]) There are infinitely many prime numbers.*

[5] Euclid of Alexandria, 325BC - 265BC, was a Greek mathematician best known for his treatise on geometry: The Elements. This influenced the development of Western mathematics for more than 2000 years. For more see http://www-groups.dcs.st-and.ac.uk/%7Ehistory/Biographies/Euclid.html

Proof. Let $p_1, p_2, \dots, p_n$ 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 $q \in \{p_1, p_2, \dots, p_n\}$ 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. □

So! We know 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 the 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[6] at the age of 15 or 16 and proven by J. Hadamard and Ch. de la Vallée Poussin in 1896. The famous Riemann hypothesis, currently (Fall 2008) one of the Clay million dollar prizes, is initially a conjecture about the zeros of some complex valued function: the 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

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

– *How small can the gaps between consecutive primes be?* For example, we have $(3, 5)$, $(5, 7)$, $(11, 13)$, $(17, 19)$, ... are there infinitely many such pairs? The answer is believed to be yes but no one has proven it yet (Fall 2008). 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

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

This is a set of $N - 1$ consecutive integers non of which is prime.

– *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 $2n$, but this is a difficult result. (Do not use it in the exercises.) In 2013, Yitang Zhang proved that there are infinitely many pairs of primes (x, y) such that $|x - y| \leq 7 \cdot 10^7$. It is expected that his methods can be fined tuned to make the constant $7 \cdot 10^7$, but probably some radically new ideas are required to get it down to 2.

– *What about adding primes? Goldbach's conjecture* asserts that every even integer greater than 2 is the sum of two prime numbers $4 = 2 + 2$, $6 = 3 + 3$, $8 = 3+5$, $10 = 3+7$, $12 = 5 + 7$, $14 = 3 + 11$, $16 = 5+ 11$, ... It has been verified (Fall 2008) up to $n \leq 4 \times 10^{17}$, but no proof is currently known.

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

Proposition 10.1.1. *Let a, b be non-zero integers. Then $a|b$ if and only if $a = \epsilon p_1^{a_1} \cdots p_m^{a_m}$ 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 $a_i' \geq a_i$ for all $i = 1, \dots, m$.*

[6] 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. For more see http://www-groups.dcs.st-and.ac.uk/%7Ehistory/Biographies/Gauss.html

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 10.1.2. *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. The advantage of Proposition 10.1.4 is that it shows that a *specific* number is irrational. We will need the following result.

Proposition 10.1.3. *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 and 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, we can consider the following situation WLOG:

$$q = \epsilon p_1^{a_1}\dots p_m^{a_m} = \epsilon' p_1^{a'_1}\dots p_m^{a'_m}.$$

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. WLOG, 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

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

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 10.1.4. *$\sqrt{2}$ 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 $2|m^2$. If $m = 2k + 1$ then $m^2 = 4k^2 + 4k + 1$. Since $2|(4k^2 + 4k)$ it would follow that $2|1$ which is a contradiction. (In simple language, what we are saying is that m is odd implies that m^2 is odd.) Thus, m is even. Then $2|m$. 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 10.1.5. We can draw the conclusion that also $\alpha = \sqrt{1 + \sqrt{2}}$ is irrational. Indeed, if α were 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.

Although the second proof of Proposition 10.1.4 is more elementary, the first proof is better. It is much easier to generalize and indeed in a similar way, one can show $\sqrt{3}$, $\sqrt[5]{17}$ etc. are irrational.

It is also known that e is irrational (not too hard) and π is irrational (hard). But it is still an open question (Fall 2008) 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 $a|c$.

(4) If $r \in \mathbb{Z}$ and r is a nonzero 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, $n \geq 1$ 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.

Part 3. **Congruences and modular arithmetic**

12. 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 have for functions, where there is precisely one t. 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. A relation is called *transitive* if $x \sim y$ and $y \sim z$ implies $x \sim z$. A relation is called a *partial order* if it is transitive and in addition we don't have both $x \sim y$ and $y \sim x$. We then use the notation $x < y$ for $x \sim y$. We then require $x < y, y < z \Rightarrow x < z$ and we do not allow both $x < y$ and $y < x$. There may very well be x, y for which neither $x < y$ nor $y < x$ holds. A *linear order* (or a *simple order*) is a partial order such that for every x, y we have either $x < y$ or $y < x$.

Another important class of relations, even more important for this course, 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. Implicit in the word "identify" is that x is identified with x (well, obviously!), 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 have an equivalence relation.

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 12.0.6. *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 $S = \cup_{i \in I} U_i$ 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 $x, z \in U_j$, meaning $x \sim z$. The equivalence classes are clearly the U_i. □

We introduce the following terminology: we say that a set $\{x_i : i \in I\}$, 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 some $i \in I$ and no two equivalence classes $[x_i], [x_j]$ for $i, j \in I$ are the same, unless $x_i = x_j$.

13. Congruence relations

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 13.0.7. *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 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 13.0.8. *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* $\bar{i}$ *is* $\overline{-i} = \overline{n - i}$.

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

Example 13.0.9. 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 $\bar{4} - \bar{5} = \bar{4} + \bar{8} = \bar{1}2$, 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. 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, that doesn't make sense. However, if you'd like you can think about multiplication as repeated addition $5 \cdot 3 = 5 + 5 + 5$. 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 try and solve the equation $4x + 2 = 7$ in $\mathbb{Z}/13\mathbb{Z}$. Now, and from now on, we are just writing $4, 2, 7$ etc. for $\bar{4}, \bar{2}, \bar{7}$. So we need to solve $4x = 5$. We are now looking for a residue class r modulo 13 so that $4r \equiv 1 \pmod{13}$. We guess that $r = 10$ and check: $4 \cdot 10 = 40 \equiv 1 \pmod{13}$. 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 $x = 11$. We remark that in general such an r need not exists if the modulos n is not a prime, and that in that case one may need to go back to the original equation and verify that indeed the solution to the reduced equation solves the original equation. These issues will be discussed later. In particular, as we shall see, there are efficient ways to find the solution to $ax = 1 \pmod{n}$, while there is no way to do that directly for equations of the form $ax = b \pmod{n}$.

Example 13.0.10. As another example, we give the addition and multiplication table of $\mathbb{Z}/5\mathbb{Z}$.

$+$	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

,

$\cdot$	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) 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 $\{0, 1, 2, \dots, n-1\}$. We defined $\bar{i} + \bar{j} = \overline{i+j}$. 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) $\bar{i} + \bar{j} = \overline{i+j} = \overline{j+i} = \bar{j} + \bar{i}$.

(2) $(\bar{i} + \bar{j}) + \bar{k} = \overline{i+j} + \bar{k} = \overline{(i+j)+k}$. Note that at this point we used the simplification that we can use any representative of the equivalence class to carry out the operations. Had we insisted on always using the representative in the set $\{0, 1, 2, \dots, n-1\}$ we would usually need 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) $(\bar{i} \cdot \bar{j})\bar{k} = \overline{ij} \cdot \bar{k} = \overline{(ij)k} = \overline{i(jk)} = \bar{i} \cdot \overline{jk} = \bar{i}(\bar{j} \cdot \bar{k})$.

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

(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 axioms hold follows from the fact they hold for R.

Theorem 13.0.11. *$\mathbb{Z}/n\mathbb{Z}$ 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 13.0.12. *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 $(n, a) = 1$. Consider the list of elements

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

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

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. In fact, 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. Finite fields, such as $\mathbb{F}_p$, play an important role in coding and cryptography as well as in pure mathematics.

13.1. **Fermat's little theorem.**

Theorem 13.1.1. *(Fermat[7]) 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 delegated to the assignments.

Lemma 13.1.2. *We have $p| \binom{p}{i}$ for every $1 \leq i \leq p-1$.*

[7] 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. He is also important in the foundations of the calculus. For more, see http://www-groups.dcs.st-and.ac.uk/%7Ehistory/Biographies/Fermat.html

Lemma 13.1.3. *Let R be a commutative ring and $x, y \in R$. Interpret $\binom{n}{i}$ as adding $\binom{n}{i}$ times the element 1. Then the binomial formula holds in R:*

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

Proof. (Of Fermat's little theorem) We prove that 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 $\mathbb{F}_p$, $y(1+a) \equiv 1$ and we get $y(1+a)^p \equiv 1$, that is $(1+a)^{p-1} \equiv 1$. □

Example 13.1.4. 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. There are infinitely many such numbers. The first being 561, 1105, 1729, 2465, 2821, 6601, 8911, 10585, 15841, 29341, ...[8] Primality testing programs first test divisibility by small primes available to the program as precomputed data and then 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. 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 to decide, without any doubt, if an integer is prime. Such an algorithm was discovered by Agrawal, Kayal and Saxena in 2002. 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. For more, see
http://www.ams.org/notices/200305/fea-bornemann.pdf

[8] For more see http://mathworld.wolfram.com/CarmichaelNumber.html

13.2. **Solving equations in $\mathbb{Z}/n\mathbb{Z}$.** There is no general method to solving polynomials equations in $\mathbb{Z}/n\mathbb{Z}$. We just present some selected topics.

13.2.1. *Linear equations.* We want to consider the equation $ax+b$ in $\mathbb{Z}/n\mathbb{Z}$. Let us assume that $(a,n)=1$. 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 $ua=1 \pmod{n}$.

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

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 $ax+b\equiv \pmod{n}$. 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

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

13.2.2. *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 $\mathbb{Z}/5\mathbb{Z}$). One the other hand, x^2+x+1 can be solved in $\mathbb{Z}/7\mathbb{Z}$. The solutions are $4(-1\pm\sqrt{-3})=-4\pm4\sqrt{4}=-4\pm 8=-4\pm 1=\{2,4\}$.

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.

13.3. **Public key cryptography; the RSA method.** [9]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 $ed \equiv 1 \pmod{k}$.
$\boxed{\text{X publishes for anyone to see the data } e, n.}$ 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 be kept, and kept secret! This is called the *private key*.

Y, wishing to send a secret, writes it as a number b modulo n, which is also relatively prime to n, and sends b^e (mod 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 13.3.1. *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
```

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

```
b^e = 33265563
33265563^10001 = 3 Mod n.
```

14. Exercises

(1) 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.

(2) Consider the relation on $\mathbb{Z}$ defined by $a \sim b$ if $a|b$. Is it reflexive? symmetric? transitive?

(3) 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 $N = n_0 + 10n_1 + 10^2 n_2 + \cdots + 10^k n_k$. 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 $n_0 + n_1 + \cdots + n_k$ is divisible by 3. (Hint: show that in fact N and $n_0 + n_1 + \cdots + n_k$ 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 $n_0 - n_1 + n_2 - \cdots \pm n_k$ is divisible by 11. The same Hint applies here. 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 $M - 2n_0$ 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.

(4) 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 mulitplication 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.

(5) 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 $a \cdot b^{-1}$.

(6) (a) Solve that equation $x^2 + x = 0$ in $\mathbb{Z}/5\mathbb{Z}$.

   (b) Solve that equation $x^2 + x = 0$ in $\mathbb{Z}/6\mathbb{Z}$.

   (c) Solve that equation $x^2 + x = 0$ in $\mathbb{Z}/p\mathbb{Z}$, where p is prime.

(7) Solve each of 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 $\mathbb{Z}/63\mathbb{Z}$.

(8) (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 $\mathbb{Z}/7\mathbb{Z}$.

(9) 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 before computing).

Part 4. Polynomials and their arithmetic

Some of our main achievements in studying arithmetic $\mathbb{Z}$ 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 we shall see that these constructions, notions and theorems hold true also for rings of polynomials over a field. One particular application will be, once we develop sufficient theory, the construction of finite fields with p^n elements, where p is a prime power and $n \geq 1$ an integer. One should note that 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, exist at all, let alone how to calculate in them. All this we'll be able to do.

15. 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. 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. Note that we allow some, or even all, coefficients to be zero. The *zero polynomial* 0 is the choice $n = 0$ and $a_0 = 0$. We define addition as (assume $n \geq m$)

$$(a_n x^n + \cdots + a_1 x + a_0) + (b_m x^m + \cdots + b_1 x + b_0) = a_n x^n + \cdots + (a_m + b_m) x^m + \cdots + (a_1 + b_1) x + (a_0 + b_0).$$

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 15.0.2. Take $R = \mathbb{Z}$ then

$$(2x^2 + x - 2) + (x^3 + x - 1) = x^3 + 2x^2 + 2x - 3, \qquad (2x^2 + x - 2)(x^3 + x - 1) = 2x^5 + x^4 - x^2 - 3x + 2.$$

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.

Proposition 15.0.3. *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.

Proposition 15.0.4. *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,*

$$\deg(f(x)g(x)) = \deg(f(x)) + \deg(g(x)).$$

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 $b_m \neq 0$. Then $f(x)g(x) = a_n b_m x^{n+m} + (a_n b_{m-1} + a_{n-1} b_m) x^{n+m-1} + \cdots$. Since R is an integral domain $a_n b_m \neq 0$ and so $f(x)g(x) \neq 0$ and $\deg(f(x)g(x)) = n + m$. □

16. 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 16.0.5. *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 $r_1(x) = r(x) - r_n g_m^{-1} x^{n-m} g(x) = (r_{n-1} - r_n g_m^{-1} g_{m-1}) x^{n-1} + \cdots$ has degree smaller then $r(x)$. On the other hand, $r_1(x) = r(x) - r_n g_m^{-1} x^{n-m} g(x) = f(x) - q(x)g(x) - r_n g_m^{-1} x^{n-m} g(x) = f(x) - (q(x) + r_n g_m^{-1} x^{n-m}) g(x)$ 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 $\deg(q(x) - q_1(x)) + \deg(g(x)) \geq \deg(g(x))$ and we get a contradiction. □

17. 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 $\mathbb{F}$.

17.1. **Some remarks about divisibility in a commutative ring T.** The definitions we made in § 9 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.

17.2. **GCD of polynomials.**

Definition 17.2.1. 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 the monic polynomial of largest degree dividing both $f(x)$ and $g(x)$. (We shall see below that there is a unique such polynomial.)

Theorem 17.2.2. *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

$$S = \{a(x) : a(x) = u(x)f(x) + v(x)g(x) \text{ for some } u(x), v(x) \in \mathbb{F}[x], a(x) \text{ monic}\}.$$

It is a non-empty set because if $f(x) = bx^n + l.o.t.$[10] 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 $r(x) = f(x) - q(x)(u(x)f(x) + v(x)g(x)) = (1 - q(x)u(x))f(x) - q(x)v(x)g(x)$ and so, if we let $r_1(x)$ be $r(x)$ divided by its leading coefficients, 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've shown the gcd is unique. □

[10] l.o.t. = lower order terms.

17.3. **The Euclidean algorithm for polynomials.**

Theorem 17.3.1. *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 16.0.5. The process must terminate because the degrees decrease.

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 so $a(x)|(g(x) - q_1(x)r_0(x)) = r_1(x)$, 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 17.3.2. (1) $f(x) = x^2 + 1, g(x) = x^2 + 2ix - 1$, complex polynomials. We have

$$f(x) = 1 \cdot (x^2 + 2ix - 1) + (-2ix + 2)$$

$$(x^2 + 2ix - 1) = (\frac{1}{-2i}x - \frac{1}{2})(-2ix + 2).$$

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,

$$f(x) = x \cdot (x^2 + 1) + (x + 1)$$

$$(x^2 + 1) = (x - 1) \cdot (x + 1) + 2$$

$$x + 1 = (2x + 2) \cdot 2.$$

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:

$$\begin{aligned}\frac{-15}{16}(x+1) &= g(x) - (\frac{1}{4}x - \frac{1}{16})(4x^2 + 5x + 1)\\ &= g(x) - (\frac{1}{4}x - \frac{1}{16})(f(x) - g(x))\\ &= -(\frac{1}{4}x - \frac{1}{16}) \cdot f(x) + (\frac{15}{16} + \frac{1}{4}x)\end{aligned}$$

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 $(f(x), g(x)) = x^2 + 2x + 1 = (x + 1)^2$.

17.4. **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 17.4.1. *This relation is an equivalence relation. Related polynomials are called* associates.

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 $ab \neq 0$. □

A non-constant polynomial f is called *irreducible* if $g|f$ implies that $g \sim 1$ or $g \sim f$. Note that 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.

Proposition 17.4.2. *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 $h|f$, the situation is 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 17.4.3. Here are some comments on irreducible polynomials.

(1) Every linear polynomial is irreducible.

(2) If $f = ax^2 + bx + c$ is reducible then $f = (\alpha x + \beta)(\gamma x + \delta)$, where $\alpha, \beta, \gamma, \delta \in \mathbb{F}$ and $\alpha \neq 0, \gamma \neq 0$. It follows then that f has a root in $\mathbb{F}$, for example $x = -\alpha^{-1}\beta$.

Conversely, suppose that f has a root $\alpha \in \mathbb{F}$ then, as we shall see shortly (Theorem 17.5.1), $f = (x - \alpha)g(x)$ 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, as we have seen in the assignments, we know 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 17.4.4. *(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 $r_1, \dots, r_g$ such that*

$$f = af_1^{r_1} \cdots f_g^{r_g}.$$

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 $f(x) = a_nx^n + \cdots + a_0 = a_n \cdot (a_n^{-1}f(x))$ 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

$$f_1(x) = c_1a_1(x)\cdots a_m(x), \qquad f_2(x) = c_2b_1(x)\cdots b_n(x),$$

with $c_i \in \mathbb{F}$ and a_i, b_j monic irreducible polynomials, not necessarily distinct. It follows that

$$f(x) = (c_1c_2)a_1(x)\cdots a_m(x)b_1(x)\cdots b_n(x),$$

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

$$f(x) = c_1a_1(x)\cdots a_m(x) = c_2b_1(x)\cdots b_n(x),$$

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 we have proved uniqueness for all polynomials of degree $\leq d$ and $\deg(f) = d+1 \geq 1$. Since $a_1(x)|c_2b_1(x)\cdots b_n(x)$ 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

$$c_1a_2(x)\cdots a_m(x) = c_2b_2(x)\cdots b_n(x).$$

Induction gives that $m = n$ and, after re-numbering the b_i, $a_i(x) = b_i(x), i = 2, 3, \dots, n$. □

Example 17.4.5. 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 $2 \neq 0$ 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 17.5.1). 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 try and find 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 5.3.3) and this is precisely its unique factorization.

We can now deduce from Theorem 17.4.4 the analogues of Proposition 10.1.1 and Corollary 10.1.2. The proofs are the same.

Proposition 17.4.6. *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 $i=1,\dots,m$.*

Corollary 17.4.7. *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*

$$\gcd(f,g)=p_1^{\min(a_1,b_1)}\dots p_m^{\min(a_m,b_m)}.$$

17.5. **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 17.5.1. *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:

$$f=A(x-a_1)^{s_1}\cdots(x-a_m)^{s_m}f_1(x)^{r_1}\dots f_n(x)^{r_n},$$

where the f_i are irreducible polynomials of degree larger than 1, and the r_i,s_i are positive. Note that if $f(a)=0$ then, since $f_i(a)\neq 0$ (else $f_i(x)=(x-a)g_i(x)$), 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 17.5.2 (The Fundamental Theorem of Algebra)**.** *The field of complex numbers is algebraically closed.*

It is a fact (proven in MATH370) 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 ones $x - a, a \in \mathbb{F}$. 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. This is in general impossible to decide, but we have some partial answers in special cases.

Proposition 17.5.3. *Let $f(x) = a_n x^n + \cdots + a_1 x + a_0$ 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

$$a_n s^n + a_{n-1}s^{n-1}t + \cdots + a_1 s t^{n-1} + a_0 t^n = 0.$$

Since s divides $a_n s^n + a_{n-1}s^{n-1}t + \cdots + a_1 s t^{n-1}$, it follows that $s|a_0 t^n$. Then, since $(s, t) = 1$, we get that $s|a_0$. Similarly, t divides $a_{n-1}s^{n-1}t + \cdots + a_1 s t^{n-1} + a_0 t^n$, so t divides $a_n s^n$. Now $(s, t) = 1$ implies that $t|a_n$. □

Example 17.5.4. *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 $s = \pm 1, \pm 3, t = \pm 1, \pm 2$. We have the possibilities $\pm 1, \pm 1/2, \pm 3, \pm 3/2$. 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 $\pm 3/2$ 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 $\pm 1, \pm 5$ and one verifies those are not roots. Thus, the polynomial is irreducible.

Proposition 17.5.5. *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 $f(x) = a_n x^n + \cdots + a_1 x + a_0$, $a_i \in \mathbb{R}$, $a_n > 0$. 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, $-N \leq a \leq N$ such that $f(a) = 0$. □

17.6. **Eisenstein's criterion.**

Theorem 17.6.1. *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 divisibly 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. □

17.7. **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 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 17.7.1. *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 $h(x) = \prod_{i=1,\dots,n}(x - a_i)$ for some distinct elements $a_1, \dots, a_n$ 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 numbers, 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 $\approx 2^{512}$. Even with a computer doing 10^{10} operations per second, checking all these possibilities will take more than 10^{134} years!

Proposition 17.7.1 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.

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.

18. 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 $\mathbb{F}_7[x]$?

(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 $\overline{g(x)} = \overline{a_n} x^n + \cdots + \overline{a_1} x + \overline{a_0}$. 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.

Part 5. **Rings**

19. Some basic definitions and examples

Recall our definition of a ring

Definition 19.0.2. 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 19.0.3. $\mathbb{Z}$ 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) 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 the ring axioms hold; this is left as an exercise.

Example 19.0.4. 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 $(a_{ij})_{i,j=1}^{n}$. 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

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

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. Indeed, for $n = 2$, for example, 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 19.0.5. 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 19.0.6. 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.

Definition 19.0.7. 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 19.0.8. 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 the assignments.

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 $\mathbb{R} \times \mathbb{R}$, because the identity element of $\mathbb{R} \times \mathbb{R}$, which is $(1, 1)$, does not belong to the set $\{(r, 0) : r \in \mathbb{R}\}$.

Example 19.0.9. Let $n \neq \pm 1$ be a square free integer (i.e., 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 $p^2 \nmid n$. Then $nt^2 = s^2$ and so $p|s^2$. 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:

$$(a + b\sqrt{n}) + (c + d\sqrt{n}) = a + c + (b + d)\sqrt{n}, \quad (a + b\sqrt{n})(c + d\sqrt{n}) = ac + bdn + (ad + bc)\sqrt{n}.$$

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.

20. Ideals

Definition 20.0.10. 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 20.0.11. Note that if $a \in I$ then $-a = -1 \cdot a \in I$.

We shall use the notation $I \triangleleft R$ to indicate that I is an ideal of R.

Example 20.0.12. $I = \{0\}$ and $I = R$ are always ideals. They are called the *trivial ideals*.

Example 20.0.13. Suppose that R is a division ring (e.g., a field) and $I \triangleleft R$ 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 20.0.14. 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 $s(ra_1) = (sr)a_1 = (rs)a_1 = r(sa_1) \in R$ (using commutativity!), $(ra_1)s = r(a_1 s) \in R$.

Definition 20.0.15. Let R be a commutative ring. If every ideal of R is principal, one calls R a *principal ideal ring.*

Theorem 20.0.16. *$\mathbb{Z}$ is a principal ideal ring. In fact, the list*

$$(0), (1), (2), (3), (4), \dots$$

is a complete list of the ideals of $\mathbb{Z}$. (Note that another notation is $0, 1\mathbb{Z}, 2\mathbb{Z}, 3\mathbb{Z}, 4\mathbb{Z}, \dots$.)

Proof. We already know these are 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. The only possibility is that $r = 0$ and so $b \in (i)$. Thus, $I = (i)$. □

Theorem 20.0.17. *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 small 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, $I \subseteq (f(x))$.

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, $(f(x)) = (g(x))$. □

Example 20.0.18. Let $\mathbb{F}$ be a field. One can 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 20.0.19. 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.

Example 20.0.20. 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 20.0.21. Let us consider the ring $\mathbb{F}[x]$ and in it the set

$$S = \{f(x) : f(x) = a_0 + a_2 x^2 + a_3 x^3 + \dots\},$$

of polynomials with no x term. Note that $0 \in S$ and $s_1, s_2 \in S \Rightarrow s_1 + s_2 \in S$. However, S is not an ideal. We have $1 \in S$ but $x = x \cdot 1 \notin S$.

Example 20.0.22. 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 20.0.23. 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 $I_1 + I_2 + \cdots + I_n$ is an ideal. A particular case is the following: Let R be a commutative ring and $I_i = r_iR$ a principal ideal. Then

$$r_1R + r_2R + \cdots + r_nR$$

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.

21. Homomorphisms

Let R, S be rings. A function $f : R \to S$ 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) $f(r_1r_2) = f(r_1)f(r_2)$.

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 $0_S = f(0_R)$.
- 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 $f(r_1r_2) = f(r_1)f(r_2)$, 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 $i : S \to R, i(s) = s$, is a ring homomorphism. Note that this explains why in the definition of a subring we insisted on $1_R \in S$.

Proposition 21.0.24. *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 $0_S, 1_S \in \text{Im}(f)$. Let now $s_1, s_2 \in \text{Im}(f)$, 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 $s_1 \pm s_2 \in \text{Im}(f)$. Similarly, $s_1s_2 = f(r_1r_2)$ and so $s_1s_2 \in \text{Im}(f)$. □

Definition 21.0.25. Let $f : R \to S$ be a homomorphism of rings. The *kernel* of f, $\text{Ker}(f)$, is defined as follows:

$$\text{Ker}(f) = \{r \in R : f(r) = 0\}.$$

Proposition 21.0.26. $\text{Ker}(f)$ *is an ideal of R. The map f is injective if and only if* $\text{Ker}(f) = \{0\}$.

Proof. First, since $f(0_R) = 0_S$ we have $0_R \in \text{Ker}(f)$. Suppose that $r_1, r_2 \in \text{Ker}(f)$ 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 $r_1 + r_2 \in \text{Ker}(f)$.

Now suppose that $r_1 \in \text{Ker}(f)$ and $r \in R$ is any element. We need to show that $rr_1, r_1r \in \text{Ker}(f)$. We calculate $f(rr_1) = f(r)f(r_1) = f(r)0_S = 0_S$, so $rr_1 \in \text{Ker}(f)$. Similarly for r_1r.

We have so far proven that $\text{Ker}(f)$ 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, $\text{Ker}(f) = \{0_R\}$.

Suppose conversely that $\text{Ker}(f) = \{0_R\}$. If $f(r_1) = f(r_2)$ then $0_S = f(r_1) - f(r_2) = f(r_1 - r_2)$ and so $r_1 - r_2 \in \text{Ker}(f)$. Since $\text{Ker}(f) = \{0_R\}$, 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 21.0.27. Let $n \geq 1$ be an integer. Define a function,

$$f : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z},$$

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) $f(ab) = \overline{ab} = \bar{a}\,\bar{b} = f(a)f(b)$.

The kernel of f is $\{a : \bar{a} \equiv 0 \pmod{n}\} = \{a : n|a\} = (n)$.

Example 21.0.28. Let R, S be any rings and define

$$f : R \times S \to R, \qquad f((r, s)) = r.$$

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) $f((r_1, s_1)(r_2, s_2)) = f((r_1 r_2, s_1 s_2)) = r_1 r_2 = f((r_1, s_1))f((r_2, s_2))$.

The kernel of f is $\{(r, s) : r = 0\} = \{(0, s) : s \in S\} = \{0\} \times S$.

Example 21.0.29. Let $\mathbb{F}$ be a field and $\mathbb{F}[\epsilon]$ the ring of dual numbers. Define

$$f : \mathbb{F}[\epsilon] \to \mathbb{F}, \qquad f(a + b\epsilon) = a.$$

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 $\text{Ker}(f) \subseteq (\epsilon)$. On the other hand $(c + d\epsilon)\epsilon = c\epsilon$ and that shows $(\epsilon) \subseteq \text{Ker}(f)$.

Example 21.0.30. Let $\mathbb{F}$ be a field. Let $a \in \mathbb{F}$ be a fixed element. Define

$$\alpha : \mathbb{F}[x] \to \mathbb{F}, \qquad \alpha(g(x)) = g(a).$$

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 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 17.5.1 says that $g(x) = (x - a)g_1(x)$ for some polynomial $g_1(x)$ and so $g(x) \in (x - a)$.

Example 21.0.31. Let A be the set of all continuous functions $f : [0, 1] \to \mathbb{R}$. Define the sum (resp. 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

$$\varphi : A \to \mathbb{R}, \qquad \varphi(f) = f(a).$$

Then φ is a ring homomorphism whose kernel are all the functions vanishing at the point a.

21.1. **Units.** 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, $R^\times = R - \{0\}$.

Lemma 21.1.1. *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 21.1.2. We have $\mathbb{Z}^\times = \{\pm 1\}$. We have $\mathbb{Q}^\times = \mathbb{Q} - \{0\}$.

Example 21.1.3. 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 21.1.4. Let $n \neq \pm 1$ be a square free integer. We have

$$\mathbb{Z}[\sqrt{n}]^\times = \{a + b\sqrt{n} : a^2 - b^2 n = \pm 1\}.$$

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

$$(a + b\sqrt{n})(a - b\sqrt{n}) = a^2 - b^2 n = \pm 1.$$

Example 21.1.5. Let $\mathbb{F}$ be a field. The units of the ring $M_2(\mathbb{F})$ are the matrices

$$\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc \neq 0 \right\}.$$

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 $y = z$.

22. 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 degree 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 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 22.0.6. 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 22.0.7. 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.

Lemma 22.0.8. *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) $a\in b+I$; (3) $a-b\in I$.*

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 22.0.9. *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 $a+I$. 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 22.0.8 above, $a-a'\in I, b-b'\in I$ and so $(a+b)-(a'+b')\in I$ and so, 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 it follows that $ab-a'b'\in I$. 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 $\overset{!}{=}$.

1. $\bar a+\bar b=\overline{a+b}\overset{!}{=}\overline{b+a}=\bar b+\bar a$.
2. $\bar a+(\bar b+\bar c)=\bar a+\overline{b+c}=\overline{a+(b+c)}\overset{!}{=}\overline{(a+b)+c}=\overline{a+b}+\bar c=(\bar a+\bar b)+\bar c$.
3. We have $\bar 0+\bar a=\overline{0+a}\overset{!}{=}\bar a$. (We remark that $\bar 0=I$.)
4. We have $\bar a+\overline{-a}=\overline{a+(-a)}\overset{!}{=}\bar 0$.

(5) $\bar{a}(\bar{b}\,\bar{c}) = \bar{a}\,\overline{bc} = \overline{a(bc)} \overset{!}{=} \overline{(ab)c} = \overline{ab}\,\bar{c} = (\bar{a}\,\bar{b})\,\bar{c}$.

(6) We have $\bar{a}\,\bar{1} = \overline{a\,1} \overset{!}{=} \bar{a}$ and $\bar{1}\,\bar{a} = \overline{1\,a} \overset{!}{=} \bar{a}$.

(7) $(\bar{a}+\bar{b})\bar{c} = \overline{a+b}\;\bar{c} = \overline{(a+b)c} \overset{!}{=} \overline{ac+bc} = \overline{ac}+\overline{bc} = \bar{a}\,\bar{c}+\bar{b}\,\bar{c}$. Also, $\bar{c}(\bar{a}+\bar{b}) = \bar{c}\;\overline{a+b} = \overline{c(a+b)} \overset{!}{=} \overline{ca+cb} = \overline{ca}+\overline{cb} = \bar{c}\,\bar{a}+\bar{c}\,\bar{b}$.

□

Proposition 22.0.10. *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 $a+I = 0+I$. By the lemma above this is the set of elements a such that $a-0 \in I$, namely, the kernel is precisely I. □

Example 22.0.11. 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 and the operations on $\mathbb{Z}/(n)$ are then just the operations we defined on congruence classes.

Thus, the quotient rings of $\mathbb{Z}$ are (either $\mathbb{Z}$ or) the familiar rings of congruences. 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 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.

22.1. **The quotient ring $\mathbb{F}[x]/(f(x))$.** Let $\mathbb{F}$ be a field , $f(x) \in \mathbb{F}[x]$ a non-constant polynomial. 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 13.0.7.

Lemma 22.1.1. *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 $h(x)+(f(x)) = g(x)+(f(x))$ is to say that $h(x) = q(x)f(x)+g(x)$. The requirement that $\deg(g) < \deg(f)$ is exactly to say that the expression

$$h(x) = q(x)f(x) + g(x)$$

is dividing h by f with residue. We know this is always possible and in a unique fashion. □

Theorem 22.1.2. *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 of 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 $1 = 1 - 0 \notin (f)$ (if it did, 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 $g(x) \notin (f(x))$, or $f(x) \nmid g(x)$, and so that $\gcd(f, g) = 1$ (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 $\bar{v}\bar{g} = \bar{1}$, which is the identity of the quotient ring.

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 $a_0, a_1, \dots, a_{n-1}$ 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 $a_0, a_1, \dots, a_{n-1}$. □

Example 22.1.3. 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 22.1.4. 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 370 we learn a systematic theory for deciding which polynomials stay irreducible and which do not.

Example 22.1.5. 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 22.1.6. 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.

22.2. **Every polynomial has a root in a bigger field.**

Theorem 22.2.1. *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 so as to say that $L \supset \mathbb{F}$.

Now, suppose that $f(x) = a_n x^n + \dots a_1 x + a_0$. To say that f has a root in L is to say that for some element $\ell \in L$ we have

$$a_n \ell^n + \dots a_1 \ell + a_0 = 0.$$

We check that this hold for the element $\ell = \bar{x}$. Indeed,

$$a_n \bar{x}^n + \dots a_1 \bar{x} + a_0 = \overline{a_n x^n + \dots a_1 x + a_0} = \overline{f(x)} = 0_L.$$

□

Example 22.2.2. According to this result, -1 has a square root in the field $\mathbb{R}[x]/(x^2 + 1)$. One can show that $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$.

22.3. **Roots of polynomials over $\mathbb{Z}/p\mathbb{Z}$.**

We can now continue our discussion, begun in § 17.7, 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, $p = \sum a_i 2^i$, $a_i \in \{0, 1\}$, $x^p = \prod_{\{i : a_i \neq 0\}} x^{2^i}$ 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 field $\mathbb{L} = \mathbb{F}_{17}[x]/(x^3 + x + 1)$. 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. This 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.

23. The First Isomorphism Theorem

23.1. **Isomorphism of rings.**

Definition 23.1.1. Let R, S be rings. A ring homomorphism $f\colon R \to S$ is called an *isomorphism* if f is bijective.

Lemma 23.1.2. *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 $f(g(s_1)g(s_2)) = f(g(s_1))f(g(s_2)) = s_1 s_2 = f(g(s_1))f(g(s_2))$. □

Definition 23.1.3. Let R, S be rings. We say that R and S are isomorphic if there is a ring isomorphism $R \to S$.

Lemma 23.1.4. *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 $R \cong S$.

23.2. **The First Isomorphism Theorem.**

Theorem 23.2.1. *Let $f\colon R \to S$ be a surjective homomorphism of rings. Let $I = \ker(f)$ then there is an isomorphism $F\colon R/I \to S$, such that the following diagram commutes*

$$\begin{tikzcd}
R \arrow[rr, two heads, "f"] \arrow[dr, "\pi"'] & & S , \\
& R/I \arrow[ur, "F"'] &
\end{tikzcd}$$

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:

- $F(1_{R/I}) = F(\bar{1}_R) = f(1_R) = 1_S$;
- $F(\bar{g} + \bar{h}) = F(\overline{g + h}) = f(g + h) = f(g) + f(h) = F(\bar{g}) + F(\bar{h})$;
- $F(\bar{g}\,\bar{h}) = F(\overline{gh}) = f(gh) = f(g) \cdot f(h) = F(\bar{g}) \cdot F(\bar{h})$.

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 23.2.2. 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 23.2.3. 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

$$\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}.$$

23.3. **The Chinese Remainder Theorem.**

Theorem 23.3.1. *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

$$f : \mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}, \quad f(a) = (a \pmod m, a \pmod n).$$

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 23.3.2. *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

$$\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}.$$

23.3.1. *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 23.3.3. *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 23.3.4. *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 23.3.5. *Let $m_1, \dots, m_k$ be relatively prime non-zero integers (that is $(m_i, m_j) = 1$ for $i \neq j$). Then there is an isomorphism*

$$\mathbb{Z}/m_1m_2\dots m_k\mathbb{Z} \cong \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_k\mathbb{Z},$$

given by

$$a \pmod{m_1m_2\dots m_k} \mapsto (a \pmod{m_1}, a \pmod{m_2}, \dots, a \pmod{m_k}).$$

The theorem is not hard to prove by induction on k. The main case, $k = 2$, is the one we proved above. Now, let $g(x) = a_nx^n + \dots a_1x + a_0$ be a polynomial with integer coefficients. Let S be the solutions of g in $\mathbb{Z}/m_1m_2 \dots m_k\mathbb{Z}$ 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

$$a \pmod{m_1m_2 \dots m_k} \mapsto (a \pmod{m_1}, a \pmod{m_2}, \dots, a \pmod{m_k}).$$

Indeed, $g(a) \pmod{m_1m_2 \dots m_k}$ is mapped to $(g(a) \pmod{m_1}, g(a) \pmod{m_2}, \dots, g(a) \pmod{m_k})$ and $g(a) \equiv 0 \pmod{m_1m_2 \dots m_k}$ 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 23.3.6. *Let $m_1, \dots, m_k$ 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 23.3.7. 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 23.3.8. 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

$$\mathbb{Z}/m_1m_2 \dots m_k\mathbb{Z} \to \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_k\mathbb{Z}.$$

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 23.3.9. *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).

24. Prime and maximal ideals

Let R be a commutative ring and I an ideal of R.

Definition 24.0.10. 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 $I \subsetneqq J \subsetneqq R$. The ideal I is called *prime* if $ab \in I$ implies $a \in I$ or $b \in I$.

Theorem 24.0.11. *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 $\bar{b} = \bar{0}$. 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 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 24.0.12. Suppose that R is a principal ideal commutative ring, namely, a commutative ring in which every ideal is principal. An ideal $I = (f)$ is maximal if and only if $f \neq 0$ and $g|f$ implies $g \sim 1$ or $g \sim f$. An element f with such properties is called irreducible and so I is maximal if and only if f is irreducible.

An ideal $I = (f)$ is prime if and only if $f|ab$ implies $f|a$ or $f|b$. An element f with such property is called prime and so I is prime if and only if f is prime.

The general implication "I maximal implies I prime" tells us that every irreducible element is also prime. In the rings $\mathbb{Z}$ and $\mathbb{F}[x]$ the converse is true - a prime element is irreducible - but that is not true in general.

25. 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

$$I \cap J = \{r : r \in I, r \in J\}$$

(the intersection of the sets). 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 usually studied in MATH570, MATH571).

(4) **The ring of real quaternions $\mathbb{H}$.** 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 extend 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 any commutative ring and let $a_1, \dots, a_n$ be elements of R. We define $(a_1, \dots, a_n)$ to be the set
$$\{r_1a_1 + \cdots + r_na_n : \forall i\ r_i \in R\}.$$
Prove that $(a_1, \dots, a_n)$ is an ideal of R. We call it the ideal generated by $a_1, \dots, a_n$.

(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 $f : R \to \mathbb{Z}_2$ 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 $f : R \to S$ 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 a rational number.
(b) Let $\mathbb{Q}[\sqrt{d}] := \{a + b\sqrt{d} : a, b \in \mathbb{Q}\}$ 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 of 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 (4), 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**

26. First definitions and examples

26.1. **Definitions and some formal consequences.**

Definition 26.1.1. 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)

Here are some formal consequences of the definition:

(1) $\boxed{e \textit{ is unique.}}$ 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) $\boxed{d \textit{ appearing in (3) is unique}}$ (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) $\boxed{\textit{Cancelation: } ab = cb \Rightarrow a = c, \text{ and } ba = bc \Rightarrow a = c.}$ 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

$$\boxed{a^m a^n = a^{m+n}, \quad (a^m)^n = a^{mn}.}$$

26.2. **Examples.**

Example 26.2.1. The *trivial group* G is a group with one element e and multiplication law $ee = e$.

Example 26.2.2. 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 26.2.3. 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 $v_2 v_1$ is an inverse for $u_1 u_2$ 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, $\mathbb{F}^\times = \mathbb{F} - \{0\}$. 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 26.2.4. *Let $n > 1$ be an integer. The group $\mathbb{Z}/n\mathbb{Z}^\times$ 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 specific 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 26.2.5. 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 26.2.6. 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.

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. Can you see that?

26.3. **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 26.3.1. 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 26.3.2. 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 26.3.3. Let G be a group. G is called *cyclic* if there is an element $g \in G$ such that $G = \{g^n : n \in \mathbb{Z}\}$; that is, any element of G is a power of g. The element g is then called a *generator* of G.

Example 26.3.4. 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 26.3.5. The group $\mathbb{Z}$ is cyclic. As a generator we may take 1 (or -1).

Example 26.3.6. 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.

27. The permutation and dihedral groups

27.1. **Permutation groups.**

Definition 27.1.1. A *permutation* of a set T is a bijective function $f : T \to T$. We shall denote the set of permutations of T by S_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 27.1.2. *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 $\{1, 2, \dots, n\}$ 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 $n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1 = n!$ choices. □

Example 27.1.3. (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 first line is in fact a cyclic subgroup of S_3. It is the subgroup generated by $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$.

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:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 4 & 5 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 4 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 4 & 1 & 3 \end{pmatrix}.$$

27.2. **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 $a_i \in \{1, 2, \dots, n\}$ 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:

$$\begin{tikzcd} a_1 \arrow[r, bend left] & a_2 \arrow[r, bend left] & a_3 \arrow[r, bend left] & \cdots \arrow[r, bend left] & a_t \arrow[llll, bend left] \end{tikzcd}$$

(and elements outside the cycle don't 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 27.2.1. 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 27.2.2. *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 $\{e, g, g^2, \dots, g^r\}$ 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 $g^n = (g^k)^a g^b = e^b g^b = g^b$. 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 27.2.3. 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 27.2.4. *Let $\sigma_1, \dots, \sigma_n$ be disjoint permutations of orders $r_1, \dots, r_n$, respectively. Then the order of the permutation $\sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_n$ is $\mathrm{lcm}(r_1, r_2, \dots, r_n)$.*

Combining this lemma with the following proposition allows us to calculate the order of every permutation very quickly.

Proposition 27.2.5. *Every permutation is a product of disjoint cycles.*

We shall not provide formal proof of this proposition, but illustrate it by examples.

Example 27.2.6. 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 27.2.7. 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 $\text{lcm}(4, 3, 2) = 12$.

Example 27.2.8. 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 $\sigma = (i_1 i_2 \dots i_{k-1} i_k)$, then σ^{-1} is easily seen to be $(i_k i_{k-1} \dots i_2 i_1)$. So,

$$(i_1 i_2 \dots i_{k-1} i_k)^{-1} = (i_k i_{k-1} \dots i_2 i_1).$$

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 and so commute, 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 $(1423)(5\ 8\ 10)(79)$.

27.3. **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 x-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 y given by the cycle $(1\ 2\ 3\ \cdots\ n)$ is an element of the dihedral group rotating the polygon by angle $360°/n$ 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 $\theta - 360°/n$). Another symmetry, x, is reflection through the x-axis. The symmetry x 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, x changes an angle θ to $-\theta$.

Theorem 27.3.1. *The elements of the dihedral group are*

$$D_n = \{e, y, \dots, y^{n-1}, x, yx, y^2x, \dots, y^{n-1}x\},$$

and the relations $x^2 = y^n = 1$ 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

$$\{e, y, \dots, y^{n-1}, yx, y^2x, \dots, y^{n-1}x\}$$

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

$$y^a(1) = a+1, \quad y^a(2) = a+2,$$

and

$$y^ax(1) = a+1, \quad y^ax(2) = y^a(n) = a.$$

This proves our claims.

The relations $x^2 = y^n = 1$ are evident. We check that $xyxy = 1$, by checking that $xyxy(j) = j$ for $j = 1, 2$. We have $xyxy(1) = xyx(2) = xy(n) = x(1) = 1$ and $xyxy(2) = xyx(3) = xy(n-1) = x(n) = 2$. □

The nature of the symmetries $1, y, \dots, y^{n-1}$ is clear: y^j rotates clockwise by angle $j \cdot 360°/n$.

Proposition 27.3.2. *Let $0 \le j < n$. The element y^jx is a reflection through the line forming an angle $-j \cdot 360°/2n$ with the x-axis.*

Proof. The symmetry y^jx is not trivial. If it fixes an angle θ it must be reflection through the line with that angle. Note that y^jx sends the angle θ to $-\theta$ and then adds $-j \cdot 360°/n$ so the equation is $\theta = -\theta - j \cdot 360°/n$ (mod 360). That is $\theta = -j \cdot 360°/2n$. □

Example 27.3.3. We revisit Example 26.2.6. 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 $1, 2, \dots, 8$, 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.

28. The theorem of Lagrange

28.1. **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 28.1.1. 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 28.1.2. *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 28.1.3. The Lemma and its proof should be compared with Lemma 22.0.8. 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 28.1.4. *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 28.1.2 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.

28.2. **Lagrange's theorem.**

Theorem 28.2.1. *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 $g_i H$ has the same number of elements, which is exactly the number of elements in $H = eH$. We get therefore that

$$|G| = |H| \cdot |I|.$$

That completes the proof. □

$\boxed{\text{figure}}$

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 $\mathrm{ord}(g) = |\langle g \rangle|$.

29. Homomorphisms and isomorphisms

29.1. **homomorphisms of groups.**

Definition 29.1.1. 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 29.1.2. *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 29.1.3. *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 $e_G \in \mathrm{Ker}(f)$. Next, if $g_1, g_2 \in \mathrm{Ker}(f)$ 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, $g_1g_2 \in H$. Finally, if $g \in \mathrm{Ker}(f)$ then $f(g^{-1}) = f(g)^{-1} = e_H^{-1} = e_H$ and so $g^{-1} \in \mathrm{Ker}(f)$ 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 $\mathrm{Ker}(f) = \{e_G\}$. Conversely, suppose $\mathrm{Ker}(f) = \{e_G\}$ 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 $g_1^{-1}g_2 \in \mathrm{Ker}(f)$ and so $g_1^{-1}g_2 = e_G$. That is, $g_1 = g_2$. □

29.2. **Isomorphism.**

Definition 29.2.1. A group homomorphism $f\colon G \to H$ is called an *isomorphism* if it is bijective.

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. §23.1.

Example 29.2.2. *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 29.2.3. 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 29.2.4. (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, $1, 2, 3, \dots$), we got a map

$$G \to S_n, \quad g \mapsto \sigma_g.$$

This map is a homomorphism of groups: $\sigma_{gh}(a) = gha = \sigma_g(\sigma_h(a))$. That is, $\sigma_{gh} = \sigma_g \circ \sigma_h$. This homomorphism is injective: if σ_g is the identity permutation then $\sigma_g(e) = e$ 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 29.2.5. We were somewhat informal about identifying the permutations of G with S_n. A more rigorous approach is the following.

Lemma 29.2.6. *Let T, Z be sets and $f : T \to Z$ 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. □

30. Group actions on sets

30.1. **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 30.2.1 states.

30.2. **Basic properties.**

Lemma 30.2.1. (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 $g_1 H = g_2 H$. 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 30.2.2. *The set S is a disjoint union of orbits.*

Proof. The orbits are the equivalence classes of the equivalence relation $\sim$ defined in Lemma 30.2.1. Any equivalence relation partitions the set into disjoint equivalence classes. □

30.3. **Some examples.**

Example 30.3.1. Let $G = S_n$, the symmetric group on $T = \{1, 2, \dots, n\}$. 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 $\sigma \in \text{Stab}(i)$ 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, $\text{Stab}(i) \cong S_{n-1}$. The orbit of i is T; for every j the transposition (ij) shows that $j \in$ Orb(i).

Example 30.3.2. 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 30.3.3. 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: $e_Hg = g$ 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 30.3.4. We consider a roulette with n sectors and write $n = i_1 + \dots + i_k$, for some positive (and fixed) integers $i_1, \dots, i_k$. We suppose we have different colors $c_1, \dots, c_k$ 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 30.3.5. Let G be the dihedral group D_8. Recall that G is the group of symmetries of a regular octagon in the plane.

$$G = \{e, y, y^2, \dots, y^7, x, yx, y^2x, \dots, y^7x\},$$

where y is rotation clockwise by angle $2\pi/8$ and x is reflection through the x-axis. We have the relations

$$x^2 = y^8 = e, \quad xyxy = 1.$$

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 x and under y^2. Therefore, the stabilizer of s_0 contains at least the set of eight elements

$$\{e, y^2, y^4, y^6, x, y^2x, y^4x, y^6x\}. \tag{30.1}$$

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 $y \notin \mathrm{Stab}(s_0)$. It follows that the stabilizer has exactly 8 elements and is equal to the set in (30.1).

Let H be the stabilizer of s_0. According to Lemma 30.2.1 the orbit of s_0 is in bijection with the left cosets of $H = \{e, y^2, y^4, y^6, x, y^2x, y^4x, y^6x\}$. By Lagrange's theorem there are two cosets. For example, H and gH are distinct cosets. The proof of Lemma 30.2.1 tells us how to find the orbit: it is the set $\{s_0, gs_0\}$, which is of course quite clear if you think about it.

31. The Cauchy-Frobenius Formula

Theorem 31.0.6. (CFF)[11] *Let G be a finite group acting on a finite set S. Let N be the number of orbits of G in S. Define*

$$I(g) = |\{s \in S : g \star s = s\}|$$

(the number of elements of S fixed by the action of g). Then [12]

$$N = \frac{1}{|G|}\sum_{g\in G} I(g).$$

Remark 31.0.7. 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).

[11] This is also sometimes called Burnside's formula.

[12] The sum appearing in the formula means just that: If you write $G = \{g_1, \dots, g_n\}$ 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 $S = \{s_1, \dots, s_m\}$ then the double sum is $T(g_1, s_1) + T(g_1, s_2) + \cdots + T(g_1, s_m) + T(g_2, s_1) + T(g_2, s_2) + \cdots + T(g_2, s_m) + \cdots + T(g_n, s_1) + T(g_n, s_2) + \cdots + T(g_n, s_m)$.

Proof. We define a function

$$T : G \times S \to \{0, 1\}, \quad T(g, s) = \begin{cases} 1 & g \star s = s, \\ 0 & g \star s \neq s. \end{cases}$$

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 $s_1, \dots, s_N$ 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 31.0.8. 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 31.0.9. *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. □

31.1. **Some applications to Combinatorics.**

Example 31.1.1. 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 $1, \dots, 11$. 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 31.1.2. 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 31.1.3. 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. [13]

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 rational 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

$$(i_1\; i_2)(i_3\; i_4)(i_5\; i_6)(i_7\; i_8).$$

[13] $y^{(3^2)} = g^9 = g$ because $y^8 = e$, etc.

(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 31.1.4. 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 27.3.3) 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

$$N = \frac{1}{12}(6 + 3 \times 2) = 1.$$

32. 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 32.0.5. (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 $n|\ |S|$, 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. □

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* Ker(f) *is a normal subgroup of G.*

Proof. We proved already that Ker(f) is a subgroup. Let $g \in G, h \in$ Ker(f); we need to show that $ghg^{-1} \in$ Ker(f), that is $f(ghg^{-1}) = e_H$. We calculate $f(ghg^{-1}) = f(g)f(h)f(g^{-1}) = f(g)e_Hf(g^{-1}) = f(g)f(g^{-1}) = f(g)f(g)^{-1} = e_H$. □

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:*

$$\begin{tikzcd}
G \arrow[rr, "f"] \arrow[dr, "\pi"'] & & H . \\
& G/\mathrm{Ker}(f) \arrow[ur, "F"'] &
\end{tikzcd}$$

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

$$F : G/\mathrm{Ker}(f) \to H, \qquad F(\bar{a}) = f(a).$$

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 $\text{ord}(g) = |\langle g \rangle|$ and it divides $|G|$. So, in our case, $\text{ord}(g) = 1, 2$ or 4. If $\text{ord}(g) = 4$, we get that G is cyclic and if $\text{ord}(g) = 1$ then $g = e_G$. Thus, we must have $\text{ord}(g) = 2$.

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 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, $\mathbb{Z}/p\mathbb{Z}^\times = \{1, \dots, p-1\}$ 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 $\mathbb{H}$, discussed in a previous assignment (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: $1/3 = 0.33333\dots$ has period 1 and the order of 10 (mod 3) $= 1$ (mod 3), with respect to multiplication, is 1. We have $1/7 = 0.142857142857142857142857142 9\cdots$, which has period 6; the order of 10 (mod 7) $= 3$ (mod 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 $\mathbb{Z}_6 \not\cong S_3$, 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. 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 $g \in G$, $I(g) = |\{s \in S : gs = s\}|$.

(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.)

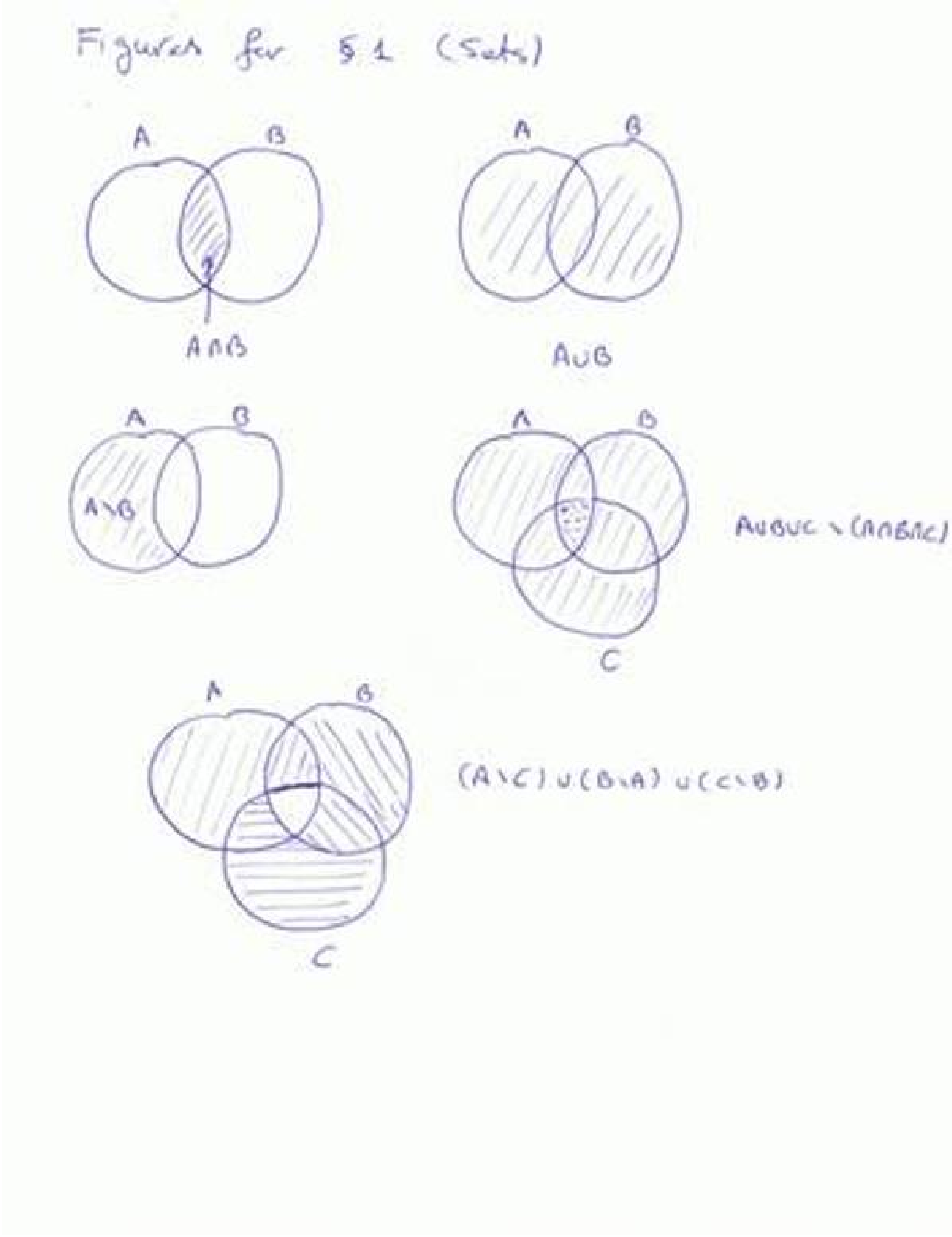
(16) Let R be the ring of matrices

$$\left\{\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33}\end{pmatrix} : a_{ij} \in \mathbb{Z}_2\right\}.$$

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 $G \cong D_4$.

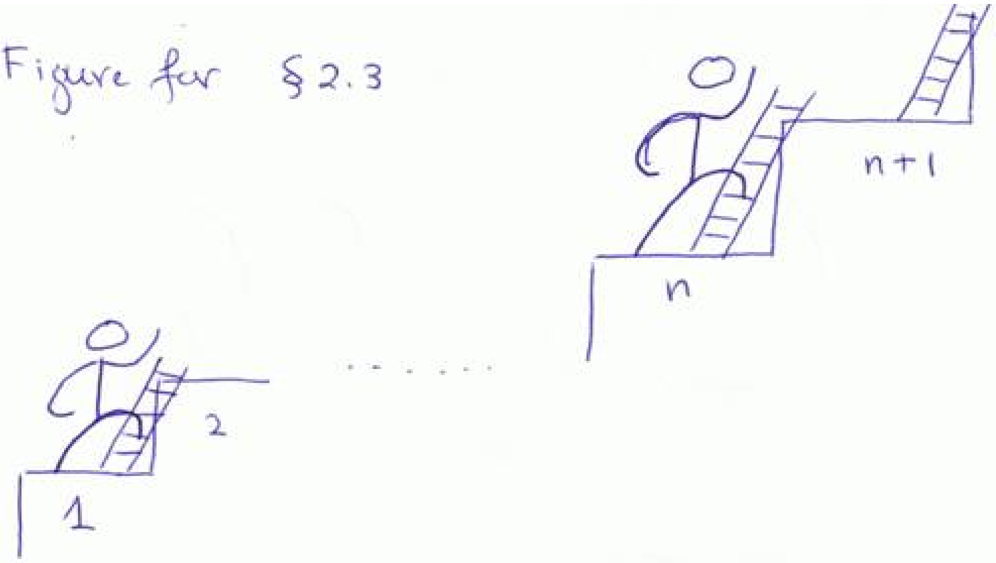
(17) Let H be a subgroup of index 2 of a group G. Prove that H is normal in G.

Part 7. Figures



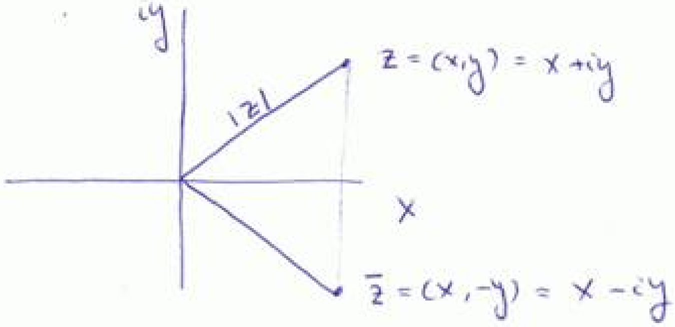
Figures for § 1 (Sets)

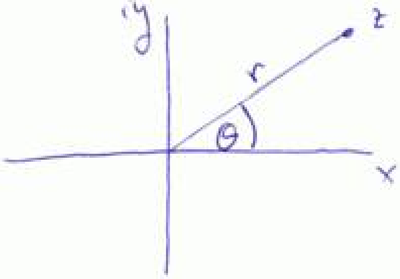
Figure for §2.3



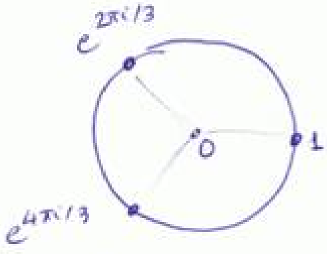
(2).jpg

Figures for §5

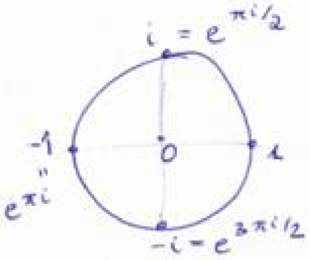




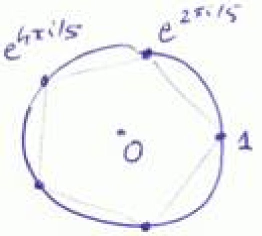
Figures for §6



3rd roots of 1



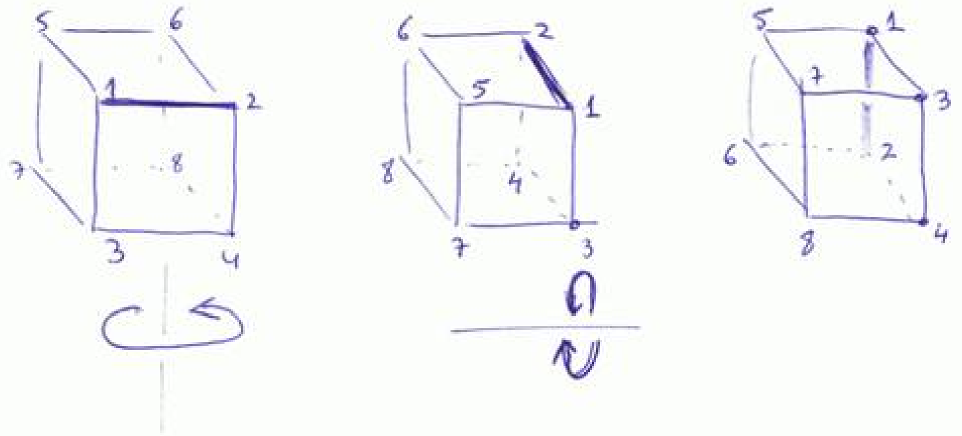
4th roots of 1



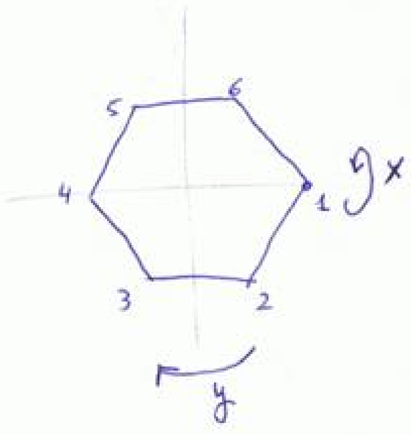
5th roots of 1.

(3).jpg

Figures for § 26.2

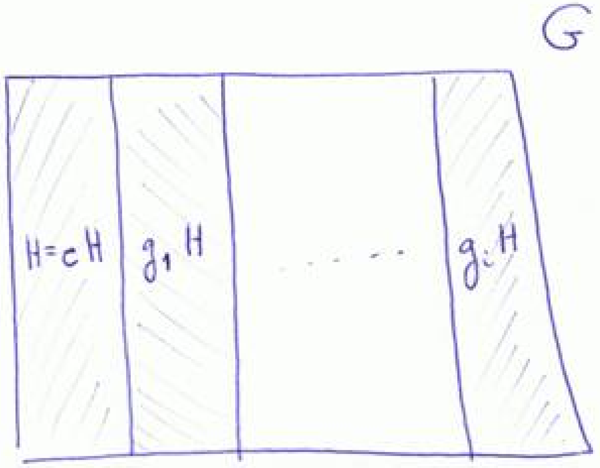


Figures for § 26.3



(4).jpg

Figures for § 28.2



(5).jpg

Index