Algebra I – MATH235

Course Notes by Dr. Eyal Goren

McGill University

Fall 2006

Last updated: November 28, 2006.

Contents

1. Introduction 4

Part 1. Some Language and Notation of Mathematics 6

2. Sets 6
3. Proofs: idea and technique 8
3.1. Two inequalities 8
3.2. By contradiction and the contrapositive 9
3.3. Induction 10
3.4. Prove or disprove 11
3.5. The pigeonhole principle 11
4. Functions 11
4.1. Injective, surjective, bijective, inverse image 13
4.2. Composition of functions 13
4.3. The inverse function (planned) 13
4.4. The notion of cardinality of a set (planned) 13
5. Number systems 14
5.1. The polar representation 16
5.2. The Fundamental Theorem of Algebra 17
6. Fields and rings - definitions and first examples 19

6.1. Some formal consequences of the axioms 21

Part 2. Arithmetic in $\mathbb{Z}$ 22
7. Division 22
8. GCD and the Euclidean algorithm 23
8.1. GCD 23
8.2. The Euclidean algorithm 24
9. Primes and unique factorization 25
9.1. Some more applications of the Fundamental Theorem of Arithmetic 28

Part 3. Congruences and modular arithmetic 30
10. Relations 30
11. Congruence relations 31
11.1. Fermat's little theorem 34
11.2. Solving equations in $\mathbb{Z}_n$. 35
11.3. Public key cryptography; RSA method 36

Part 4. Polynomials and their arithmetic 38
12. The ring of polynomials 38
13. Division with residue 39
14. Arithmetic in $\mathbb{F}[x]$ 39
14.1. Some remarks about divisibility in a commutative ring T 40
14.2. GCD of polynomials 40
14.3. The Euclidean algorithm for polynomials 41
14.4. Irreducible polynomials and unique factorization 42
14.5. Roots 44
14.6. Eisenstein's criterion 46
14.7. Roots of polynomials in $\mathbb{Z}_p$ 46

Part 5. Rings 47
15. Some basic definitions and examples 47
16. Ideals 49
17. Homomorphisms 51
17.1. Units 53
18. Quotient rings 54
18.1. The quotient ring $\mathbb{F}[x]/(f(x))$ 57
18.2. Every polynomial has a root in a bigger field 58
18.3. Roots of polynomials over $\mathbb{Z}_p$ 58
19. The First Isomorphism Theorem 59
19.1. Isomorphism of rings 59
19.2. The First Isomorphism Theorem 60
19.3. The Chinese Remainder Theorem 61
20. Prime and maximal ideals 62

Part 6. Groups 63
21. First definitions and examples 63

21.1. Definitions and some formal consequences 63
21.2. Examples 63
21.3. Subgroups 65
22. The permutation and dihedral groups 65
22.1. Permutation groups 65
22.2. Cycles 67
22.3. The Dihedral group 67
23. The theorem of Lagrange 68
23.1. Cosets 68
23.2. Lagrange's theorem 69
24. Homomorphisms and isomorphisms 70
24.1. homomorphisms of groups 70
24.2. Isomorphism 71
25. Group actions on sets 72
25.1. Basic definitions 72
25.2. Basic properties 73
25.3. Some examples 74
26. The Cauchy-Frobenius Formula 75
26.1. Some applications to Combinatorics 77
27. Cauchy's theorem: a wonderful proof 79
28. The first isomorphism theorem for groups 80
28.1. Normal subgroups 80
28.2. Quotient groups 81
28.3. The first isomorphism theorem 81
28.4. Groups of low order 83
28.5. Odds and evens 84
28.6. Odds and Ends 85

1. Introduction

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 the book. The book can be considered as the first treatise on algebra. The word "algorithm" is derived from Al-Khwarizmi. The book was very much concerned with recipes for known practical problems: "al-Khwarizmi intended to teach [...]: *... 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 is that most of you will find 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 your family, the material of the course. Work together with your class mates on assignments, but write your own solution 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 help center, to quickly close any gap and clarify any point you're not sure about.

So what is this course about really?

[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. A finite field, perhaps the most 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 slogan "to each, its 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 ;-)

Part 1. Some Language and Notation of Mathematics

2. 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.

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

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

This is a set whose elements are the numbers $2, 3$ and Eyal's (the writer) football.

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

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

(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 is standard. Other standard notation is

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

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, $3 \in S, 2 \notin S$, Eyal's football $\in T$ but $\notin U^+$.

We say that $A \subseteq B$, or simply $A \subset B$, if every element of A is an element of B. For example $\mathbb{N} \subset \mathbb{Z}$. We say that $A = B$ if A is equal to B. Note that $A = B$ holds precisely when both $A \subset B$ and $B \subset A$. The notation ϕ stands for the *empty set*. It $\underline{\text{is}}$ a set but it has no elements. We let

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

be the *intersection* of A and B, 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 index set) and suppose 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") and

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

For example, define for $i \in \mathbb{N}$,

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

(so $A_0 = \{0, 1, 2, 3, \dots\}, A_1 = \{1, 2, 3, 4, \dots\}, A_2 = \{2, 3, 4, 5, \dots\}$ and so on). Then

$$\cup_{i \in N} A_i = \mathbb{N},$$

while

$$\cap_{i \in N} A_i = \emptyset.$$

Yet another operation 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 usually $A \setminus B$ is not equal to $B \setminus A$.

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

$$A_1 \times A_2 \times \dots \times A_n = \{(x_1, x_2, \dots, x_n) : x_i \in A_i\}.$$

In particular,

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

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

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

and

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

Furthermore,

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

and

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

3. Proofs: idea and technique

Proposition 2.1 is not obvious. It is not even clear at first sight whether it’s correct. For that reason we insist on proofs in mathematics. Proofs give us confidence that we are making true statements and they reveal, to a lesser or higher extent, why the statements hold true. The proof should *demonstrate* that the statements made are true. In fact, a few decades ago, people were using the word “demonstration” instead of “proof”. We shall prove some of the statements in the proposition now, leaving the rest as an exercise. Our method of proof for Proposition 2.1 is by a standard technique.

3.1. **Two inequalities.**

When one wants to show two 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

$$x \in A \Rightarrow x \in B, \qquad x \in B \Rightarrow x \in A.$$

(The notation ⇒ 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 3.1. *Let I be a set. Let A be a set and $B_i, i \in I$, be sets as well then*

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

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 *Q.E.D.*)

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

3.2. **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 proceeds to derive logical consequences of this assumption until an obvious contradiction arises. Here is an easy example that illustrates this:

Lemma 3.2. *For every positive integer n we have $n \leq n^2$.*

Proof. Assume not. Then there is a positive integer n such that $n > n^2$. Since n is positive also $n/n > n^2/n$, i.e. $1 > n$. But there is no positive integer smaller than 1. Contradiction. □

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 in the sense that each implies the other. Here is a mathematical example:

Lemma 3.3. *Let S and T be sets. If $S \setminus T \neq \emptyset$ then $S \neq T$.*

Proof. The contrapositive is $S = T$ implies that $S \setminus T = \emptyset$. This is actually ...... obvious!! □

3.3. **Induction.** Induction is perhaps the most fun technique. Its logical foundations also lie deeper. 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 axiom stated above doesn't follow from simpler axioms and, in one form or another, has to be included as an axiom.

Theorem 3.4. *(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.

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. □

Example 3.5. 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 base 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 of $n + 1$. That is, we need to prove

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

(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

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

3.4. **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 integer n, 4 divides n^2-n."

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 ($0^2-0=0, 1^2-1=0, 5^2-5=20$, ...) 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 the dice. 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 of *one n*. If we take $n=2$ we find that 4 doesn't divide 2^2-2 and so the statement is false.

3.5. **The pigeonhole principle.** *(PLANNED)*

4. 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 4.1. 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 4.2. 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 4.3. 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.

The first definition fails because -1 doesn't have a real root. 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.

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

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

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 described as a set

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

with the following property: $\forall a \in A, \exists! b \in B$ such that $(a, b) \in \Gamma$. 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 4.4. Let A be a set and $\Gamma \subset A$ the "diagonal",

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

The function Γ defines is 1_A.

4.1. **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.)

(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 also write $f^{-1}(u)$ and call it the *fibre of f over u.*

Example 4.5. (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.

4.2. **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:

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

Lemma 4.6. *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 $g(b) = g(f(a)) = (g \circ f)(a) = c$. □

4.3. **The inverse function (planned).**

4.4. **The notion of cardinality of a set (planned).**

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 that *natural numbers*

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

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:

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

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 is 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:

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

(This is the addition of the 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. The operations describe satisfy the usual rules of arithmetic, such as

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

(We shall later say that the complex numbers form a *field.*)

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

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

Lemma 5.1. *The complex conjugate has the following properties:*

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

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

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

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

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

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

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

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

We have $\overline{z} = a - bi$ 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,

$$\begin{aligned}|z_1z_2|^2 &= z_1z_2\cdot\overline{z_1z_2}\\ &= z_1z_2\cdot\overline{z_1}\cdot\overline{z_2}\\ &= z_1\cdot\overline{z_1}\cdot z_2\cdot\overline{z_2}\\ &= |z_1|^2\cdot|z_2|^2.\end{aligned}$$

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 5.2. If $z\neq 0$ then z has an inverse. 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 we calculate

$$1+2i+\frac{3-i}{1+5i}.$$

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 by the length r of the vector and the angle θ it forms with the real axis. We have

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

Lemma 5.3. *If z_1 has parameters r_1,θ_1 and z_2 has parameters r_2,θ_2 then z_1z_2 has parameters $r_1r_2, \theta_1+\theta_2$ (up to multiples of 360^0).*

Before the proof we introduce some notation. 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 claim is

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

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

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

where e^a is the usual exponential and e^{ib} is as defined above. We then have

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

We have defined here e^z in a purely formal way. One can show that for every complex number z the series

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

converges and is equal to e^z.

Proof. (Of lemma) We have $r_1r_2 = |z_1||z_2| = |z_1z_2|$ and this shows that r_1r_2 is the length of z_1z_2. Let θ be the angle of z_1z_2 then

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

Similarly we get

$$\cos\theta = \cos(\theta_1+\theta_2).$$

It follows that $\theta = \theta_1 + \theta_2$ up to multiples of 360^0. □

5.2. **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 an 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, a real and a complex polynomial.) The *zero polynomial*, denote 0, is the case when $n = 0$ and $a_0 = 0$.

A polynomial defines a function

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

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 5.4. 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.

Theorem 5.5. *(The Fundamental Theorem of Algebra) Let $f(x)$ be a complex polynomial of degree at least* 1*. Then $f(x)$ has a root in $\mathbb{C}$.*

The proof of the theorem is beyond our scope. It has many proofs. In Honours Algebra 4 MATH371 one sees an algebraic proof; in Complex Variables and Transforms MATH 381 one sees an analytic proof.

Proposition 5.6. *Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial of degree n. Then*

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

for suitable complex numbers, not necessarily distinct, z_i. 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. In $n = 0$ 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

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

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

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

and there is no problem solving these equations. Thus, we can choose $g(x)$ with 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

$$f(z_n) = (x - z_n)g(x), \quad g(x) = a_n x^{n-1} + \ldots.$$

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

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

Furthermore, $f(z_j) = a_n \prod_{i=1}^n (z_j - z_i) = 0$, 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 number is already zero. 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:

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

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

There are also formulas for the roots for degrees 3 and 4, but in degrees 5 and higher no such formulas exist. This follows from Galois theory taught in MATH371.

6. Fields and rings - definitions and first examples

In the examples we have already discussed, or the definitions we've made, there are implicit structures that we want to formally define. At this point we just provide the definitions and reconsider previous examples. Later we'll enter a systematic development of the theory.

An *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 providing 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

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

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

Often, for a general set R we may denote $w(z_1, z_2)$ by $z_1 + z_2$, or z_1z_2, if we want to stress the fact that the operation behaves like addition, or multiplication.

Definition 6.1. A *ring* R is a non-empty set together with two operations, called "addition" and "multiplication" that are denoted, respectively, by

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

One requires the following axioms to hold:

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

We remark that for us a ring always has an identity element with respect to multiplication. In that regard, our conventions differ from Hungerford's.

Note that the multiplication is not assumed to be commutative.

Definition 6.2. 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 $0 \neq 1$, we call R a *division ring.* A commutative division ring is called a *field.*

Example 6.3. $\mathbb{Z}$ is a commutative ring. It is not a division ring and so is not a field.

Example 6.4. The rational numbers $\mathbb{Q}$ form a field. The real numbers $\mathbb{R}$ form a field. The complex numbers form a field; we at some level already used all the axioms implicitly in our calculations, but now we prove it formally.

Proposition 6.5. $\mathbb{C}$ *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 using that $\mathbb{R}$ is a field.
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. $(z_1z_2)z_3 = [(a_1 + b_1 i)(a_2 + b_2 i)](a_3 + b_3 i) = ((a_1a_2 - b_1b_2) + (a_1b_2 + b_1a_2)i)(a_3 + b_3 i) = (a_1a_2 - b_1b_2)a_3 - (a_1b_2 + b_1a_2)b_3 + ((a_1a_2 - b_1b_2)b_3 + (a_1b_2 + b_1a_2)a_3)i = a_1a_2a_3 - b_1b_2a_3 - a_1b_2b_3 - b_1a_2b_3 + (a_1a_2b_3 - b_1b_2b_3 + a_1b_2a_3 + b_1a_2a_3)i$. One now develops the product $z_1(z_2z_3)$ in the same way and checks that the answers match. We don't do that here.

6. Clearly $1 \cdot z_1 = z_1 \cdot 1 = z_1$.
7. $z_1(z_2 + z_3) = (a_1 + b_1 i)((a_2 + a_3) + (b_2 + b_3)i) = a_1(a_2 + a_3) - b_1(b_2 + b_3) + (b_1(a_2 + a_3) + a_1(b_2 + b_3)i) = (a_1a_2 - b_1b_2) + (b_1a_2 + a_1b_2)i + (a_1a_3 - b_1b_3) + (b_1a_3 + a_1b_3)i = (a_1 + b_1 i)(a_2 + b_2 i) + (a_1 + b_1 i)(a_3 + b_3 i) = z_1z_2 + z_1z_3$.

Before proving the next property, we check that $z_1z_2 = z_2z_1$. We have $z_1z_2 = (a_1 + b_1 i)(a_2 + b_2 i) = a_1a_2 - b_1b_2 + (a_1b_2 + b_1a_2)i = a_2a_1 - b_2b_1 + (a_2b_1 + b_2a_1)i = (a_2 + b_2 i)(a_1 + b_1 i) = z_2z_1$. Therefore, $(z_1 + z_2)z_3 = z_3(z_1 + z_2) = z_3z_1 + z_3z_2 = z_1z_3 + z_2z_3$.

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

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 $y = y + (x + y) = y + (x + y') = (y + x) + y' = (x + y) + y' = 0 + y' = y'$. 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 just proved, to show $-x + x = 0$ and that $(x + y) + (-x - y) = 0$. This is easy using commutativity and $x + (-x) = 0, y + (-y) = 0$ (by the very definition of $-x, -y$).

(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 $0 = -y + y = -y + (y + y) = (-y + y) + y = 0 + y = y$.

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

In this part of the course we are going to study arithmetic in $\mathbb{Z}$. We are going to focus on particular properties. Our choice of properties is motivated by a analogy to be drawn later between integers and polynomials. In fact, there even more general classes of rings for which one can extend this analogy, but in this course we shall not discuss them beyond providing some margin notes.

7. Division

Theorem 7.1. *(The division algorithm)*[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 that case $b < 0$.

Consider the set

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

S is a non-empty set, indeed, if $a > 0$ take $x = 0$ and $a \in S$. If $a < 0$ take $x = a$ and $a - bx = a(1 - b) \geq 0$ (because $b > 0$ and so $b \geq 1$). That is, $a(1 - b) \in S$. 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. □

[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 general q, r are not unique.) Thus, the theorem we are proving is that $\mathbb{Z}$ *is a Euclidean ring.*

8. GCD and the Euclidean algorithm

Definition 8.1. Let a, b be integers. We say that $a|b$ 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 $b \pm d = a \cdot (c\pm)$ and so $a|(b \pm d)$. □

Corollary 8.2. *Let $a \neq 0$. $a|b$ if and only if in dividing b in a with a 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 $r = 0$. □

8.1. **GCD.**

Definition 8.3. 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 integer dividing both a and b.

Theorem 8.4. *Let a, b be integers, not both zero, and $d = (a, b)$ their gcd. Then every common divisor of a and b divides d. We can find 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 divides it. □

Corollary 8.5. *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 $a|uac + vbc = c$. □

8.2. **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 uses the prime factorization of the two numbers – we shall discuss that later. The other, *which is much more efficient*, is the Euclidean algorithm.

Theorem 8.6. *(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:*

$$a = bq_0 + r_0, \qquad 0 < r_0 < b,$$
$$b = r_0q_1 + r_1, \qquad 0 \leq r_1 < r_0$$
$$r_0 = r_1q_2 + r_2, \qquad 0 \leq r_1 < r_2$$
$$\vdots$$

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

$$r_{t-2} = r_{t-1}q_t + r_t, \qquad 0 \leq r_t < r_{t-1}$$
$$r_{t-1} = r_tq_{t+1}.$$

Then r_t is the gcd of a and b.

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

$$113 = 54 \cdot 2 + 5$$
$$54 = 5 \cdot 10 + 4$$
$$5 = 4 \cdot 1 + 1$$
$$4 = 4 \cdot 1.$$

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

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

$$442 = 182 \cdot 2 + 78$$
$$182 = 78 \cdot 2 + 26$$
$$78 = 26 \cdot 3$$

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,

$$113 = 54 \cdot 2 + 5$$
$$54 = 5 \cdot 10 + 4$$
$$5 = 4 \cdot 1 + 1$$

$4 = 4 \cdot 1$.
Thus $gcd(113, 54) = 1$. 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
$442 = 182 \cdot 2 + 78$
$182 = 78 \cdot 2 + 26$
$78 = 26 \cdot 3$
and so $\gcd(442, 182) = 26$. Here the process is easier: $26 = 182 - 78 \cdot 2 = 182 - (442 - 182 \cdot 2) \cdot 2 = 5 \cdot 182 - 2 \cdot 442$.

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

9. Primes and unique factorization

Definition 9.1. An integer $p \neq 0, \pm 1$ is called *prime* if its only divisors are $\pm 1, \pm p$.

The phrase "prime number" is usually used to denote a prime positive integer. A positive integer is prime if its only positive divisors are 1 and p.

<u>The sieve of Eratosthenes</u>:[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 denoting 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, $\underline{50}$

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 denoting crossing out by an underline):

[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

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, $\underline{24}$, 25, $\underline{26}$, $\underline{27}$, $\underline{28}$, 29, $\underline{30}$, 31, $\underline{32}$, $\underline{33}$, $\underline{34}$, 35, $\underline{36}$, 37, $\underline{38}$, $\underline{39}$, $\underline{40}$, 41, $\underline{42}$, 43, $\underline{44}$, $\underline{45}$, $\underline{46}$, 47, $\underline{48}$, 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 denoting crossing out by an underline):

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, $\underline{48}$, 49, $\underline{50}$

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 denoting crossing out by an underline):

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, $\underline{48}$, $\underline{49}$, $\underline{50}$

The next number (11) is already greater than $\sqrt{N} = \sqrt{50} \sim 7.1$. So we stop. Any number left 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 9.2. *(The Fundamental Theorem of Arithmetic) Every integer $n \neq 0, \pm 1$ is a product of primes. 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. 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 for $-n$, $-n = p_1 p_2 \cdots p_m$ and thus $n = -p_1 p_2 \cdots p_m$).

Lemma 9.3. *Every positive integer is a product of primes numbers. (We allow the empty product, equal by definition to 1).*

Proof. Suppose not. Then the set of integers S that are not a product of prime numbers has a minimal element, say n_0. n_0 is not one, or a prime, because in those cases it is a product of primes. 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 9.4. *Let p be a positive integer. The following are equivalent: (i) p is a prime number; (ii) 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've already seen, $p|b$.

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. So p has no proper divisors and hence is prime. □

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). 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

$$n/p_1 = p_2 \cdots p_m = q_2 \cdots q_t.$$

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 such building blocks, such primes.

Theorem 9.5. *(Euclid[5]) There are infinitely many prime numbers.*

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

Consider the integer $n = p_1 p_2 \cdots p_n + 1$ and its prime factorization. Let q be a prime dividing n. If $q \in \{p_1, p_2, \ldots, p_n\}$ then $q|p_1 p_2 \cdots p_n$ and so $q|(n - p_1 p_2 \cdots p_n) = 1$, which is a contradiction. □

[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

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.

– *How small can the gaps be?* For example, we have $(3, 5)$, $(5, 7)$, $(11, 13)$, $(17, 19)$, ... are there infinitely many such pairs? The answered is believed to be yes but no one has proved it yet (Fall 2006). This is called the *Twin Prime Conjecture.*

– *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.

– *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 2006) up to $n \leq 4 \times 10^{17}$.

9.1. Some more applications of the Fundamental Theorem of Arithmetic.

Proposition 9.6. *Let a, b be non-zero integers. Then $a|b$ if and only if $a = \epsilon p_1^{a_1} \cdots p_m^{r_m}$ 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 $i = 1, \ldots, m$.*

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 $a = \epsilon p_1^{a_1} \cdots p_m^{r_m}$ 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 9.7. *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, though there are infinitely many

[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

rational numbers, even in the interval $[0,1]$, they are still a rather meagre set inside the real numbers. The advantage of the following proposition is that it shows that a *specific* number is irrational.

Proposition 9.8. *$\sqrt{2}$ 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, $m = 1$. 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.

The proof above has a flaw that we didn't prove that there is a unique factorization when we allow also negative powers. This is not hard to show, given what we have proven. 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 so $2|n$. 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. □

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 2006) if 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!).

Part 3. Congruences and modular arithmetic

10. 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 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 only 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, 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.

Lemma 10.1. *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 equivalent classes.*

Conversely, if S is a disjoint union $S = \cup_{i \in I} U_i$ of non-empty sets (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 $S = \cup_{i \in I}[x_i]$.

11. Congruence relations

Let n be a positive integer. Define a relation $x \sim y$ 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 11.1. *Congruence modulo n is an equivalence relation on $\mathbb{Z}$. The set $\{0, 1, \dots, n-1\}$ is a complete set of representatives.*

Proof. First $n|(x-x)$ so $x \equiv x$ and the relation is 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 $0 \leq r < n$. 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 $0 \leq i < j < n$ we have $i \equiv j$, that is, $n|(j-i)$. But $0 < j - i < n$ and we get a contradiction. □

Theorem 11.2. *Denote the equivalence classes of congruence modulo n by $\bar{0}, \bar{1}, \dots, \overline{n-1}$. Denote this set by $\mathbb{Z}_n$. The set $\mathbb{Z}_n$ 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 11.3. We take $n = 13$ and calculate $\bar{5} \cdot \bar{6} - \bar{5}$. First, $\bar{5} \cdot \bar{6} = \overline{30} = \bar{4}$. Then $\bar{4} - \bar{5} = \bar{4} + \overline{-5} = \overline{4-5} = \overline{-1} = \overline{12}$. Note that we could have also calculated $\bar{4} - \bar{5} = \bar{4} + \bar{8} = \overline{12}$, or $\bar{5} \cdot \bar{6} - \bar{5} = \bar{5}(\bar{6} - \bar{1}) = \bar{5} \cdot \bar{5} = \overline{25} = \overline{12}$.

Modular arithmetic, that is calculating in the ring $\mathbb{Z}_n$, 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 $\mathbb{Z}_{13}$. 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.

Example 11.4. As another example, we give that addition and multiplication table of $\mathbb{Z}_5$.

$+$	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 $\overline{i} = \overline{i'}$, $\overline{j} = \overline{j'}$, where i, i', j, j' need not be in the set $\{0, 1, 2, \dots, n-1\}$. We defined $\overline{i} + \overline{j} = \overline{i+j}$. We need to check that this is the same as $\overline{i'+j'}$. Since $\overline{i} = \overline{i'}, n|(i - i')$ 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) $\overline{i} + \overline{j} = \overline{i+j} = \overline{j+i} = \overline{j} + \overline{i}$.

(2) $(\overline{i} + \overline{j}) + \overline{k} = \overline{i+j} + \overline{k} = \overline{(i+j)+k}$. 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, $\overline{(i+j)+k} = \overline{i+(j+k)} = \overline{i} + \overline{j+k} = \overline{i} + (\overline{j} + \overline{k})$.

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

(4) $\overline{i} + \overline{-i} = \overline{i+(-i)} = \overline{0}$. 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 $\mathbb{Z}_n$, 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 11.5. *$\mathbb{Z}_n$ 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 11.6. *Let R be a commutative ring. If R has zero divisors then R is not a field.*

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

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

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

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

We claim that they are distinct elements of $\mathbb{Z}_n$. Suppose that $\bar{i}\cdot\bar{a} = \bar{j}\cdot\bar{a}$, for some $0 \leq i \leq j \leq n-1$ then $\overline{ia} = \overline{ja}$, which means that $n|(ia - ja) = (i-j)a$. Since $(n, a) = 1$, it follows that $n|(i-j)$ but that means $i = j$. Thus, the list $\bar{0}\cdot\bar{a}, \bar{1}\cdot\bar{a}, \dots, \overline{n-1}\cdot\bar{a}$ contains n distinct elements of $\mathbb{Z}_n$ 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 $\mathbb{Z}_p$ 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.

11.1. Fermat's little theorem.

Theorem 11.7. *(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 prove state two auxiliary statements whose proofs are delegated to the assignments.

Lemma 11.8. *We have $p | \binom{p}{i}$ for every $1 \leq i \leq p-1$.*

Lemma 11.9. *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 $y(1+a)^p \equiv y(1+a)$, that is $(1+a)^{p-1} \equiv 1$. □

Example 11.10. 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

[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

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

test divisibility by small primes available to the program as pre-computed 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 I'm particularly proud in drawing this conclusion...

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

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

11.2.1. *Linear equations.* We want to consider the equation $ax + b$ in $\mathbb{Z}_n$. 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 $\mathbb{Z}_n$ then x solves the equation $uax + ub = 0 \pmod{n}$, that is $x + ub = 0$ and so $x = -ub$ in $\mathbb{Z}_n$. Conversely, if $x = -ub$ in $\mathbb{Z}_n$ where $ua = 1$ in $\mathbb{Z}_n$ then $ax = a(-ub) = -aub = -b$ in $\mathbb{Z}_n$.

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 $(12, 17) = 1$. 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 $x = 7 * 3 = 21 = 4 \pmod{17}$.

11.2.2. *Quadratic equations.* Consider the equation $ax^2 + bx + c = 0$ in $\mathbb{Z}_n$ 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 solutions of this equation are given by the usual formula:

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

Those are solutions in $\mathbb{Z}_n$ if and only if $b^2 - 4ac$ is a square in $\mathbb{Z}_n$ (which may or may not be the case).

For example, the equation $x^2 + x + 1$ has no solution in $\mathbb{Z}_5$ because the discriminant $b^2 - 4ac$ is in this case $1^2 - 4 = -3 = 2$ in $\mathbb{Z}_5$ and 2 can be checked not to be a square in $\mathbb{Z}_5$ (one just tries: $0^2 = 0, 1^2 = 1, 2^2 = 4, 3^2 = 4, 4^2 = 1$ in $\mathbb{Z}_5$). One the other hand, $x^2 + x + 1$ can be solved in $\mathbb{Z}_7$. The solutions are $4(-1 \pm \sqrt{-3}) = -4 \pm 4\sqrt{4} = -4 \pm 8 = -4 \pm 1 = \{2, 4\}$.

When n is not prime, we shall not study the problem in this course, beyond remarking that one can proceed by trying all possibilities if n is small and that the number of solutions can be very large. For example: consider the equation $x^3 - x$ in $\mathbb{Z}_8$. 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.

11.3. **Public key cryptography; RSA method.** 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 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 $b^e \pmod{n}$ to X, allowing anyone interested to see that message. The point is, and this is called the *discrete log problem*, that it is very difficult to find what b is, even when one knows b^e and n. Thus, someone seeing Y's message cannot find the secret b from it.

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

Lemma 11.11. *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.

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

Part 4. Polynomials and their arithmetic

12. 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 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 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 12.1. 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 12.2. *With the operations defined above $R[x]$ is a commutative ring, with zero being the zero polynomial and* 1 *being the constant polynomial* 1*. The additive inverse of $a_n x^n + \cdots + a_1 x + a_0$ is $-a_n x^n - \cdots - a_1 x - a_0$.*

Since the proof is straightforward we leave it as an exercise.

Proposition 12.3. *If R is an integral domain then $R[x]$ is an integral domain. If $f(x), g(x) \in R[x]$ are non-zero polynomials,*

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

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 $\deg(f(x)g(x)) = n + m$. □

13. 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 13.1. *Let $f(x), g(x)$ be two polynomials in $\mathbb{F}[x]$, $g(x) \neq 0$. Then, there exist unique polynomials $q(x), r(x)$ in $\mathbb{F}[x]$ such that*

$$f(x) = q(x)g(x) + r(x), \qquad r(x) = 0 \text{ or } \deg(r(x)) < \deg g(x).$$

Proof. We first show the existence and later the uniqueness. Consider the set

$$S = \{f(x) - q(x)g(x) : q(x) \in R[x]\}.$$

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 $r(x) \in S$. 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 that $g(x)$. 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. □

14. 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 $\mathbb{F}$.

14.1. **Some remarks about divisibility in a commutative ring T.** The definitions we made in § 8 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 $c \in R$. 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, these definition and properties hold for the ring of polynomials $R[x]$, where R is a commutative ring.

14.2. **GCD of polynomials.**

Definition 14.1. 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 14.2. *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 $f(x) = bx^n + l.o.t.$[9] 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 $h(x) = h_1(x)b(x)$, but what we just 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. □

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

14.3. **The Euclidean algorithm for polynomials.**

Theorem 14.3. *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,*

$$\begin{aligned}
&f(x) = q_0(x)g(x) + r_0(x), \qquad \deg(r_0) < \deg(g)\\
&g(x) = q_1(x)r_0(x) + r_1(x), \qquad \deg(r_1) < \deg(r_0)\\
&r_0(x) = q_2(x)r_1(x) + r_2(x), \qquad \deg(r_2) < \deg(r_1)\\
&\vdots\\
&r_{t-2}(x) = q_t(x)r_{t-1}(x) + r_t(x), \qquad \deg(r_t) < \deg(r_{t-1}))\\
&r_{t-1}(x) = q_{t+1}(x)r_t(x).
\end{aligned}$$

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 13.1. 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 $r_{t-1}, r_{t-2}, \dots, r_a$ 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 $r_0, r_1, \dots, r_t$. 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 14.4.

(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 $\mathbb{F} = \mathbb{Z}_3$, 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

$$\begin{aligned}f(x) &= 1\cdot g(x) + 4x^2+5x+1\\ x^3+x^2-x-1 &= (\frac{1}{4}x - \frac{1}{16})(4x^2+5x+1) - \frac{15}{16}x - \frac{15}{16}\\ (4x^2+5x+1) &= \frac{-16}{15}(4x+1)(-\frac{15}{16}x - \frac{15}{16}).\end{aligned}$$

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 $\mathbb{F} = \mathbb{Z}_3$. We now have:

$$\begin{aligned}f(x) &= 1\cdot g(x) + x^2+2x+1\\ x^3+x^2-x-1 &= (x-1)(x^2+2x+1).\end{aligned}$$

Therefore, now we have $(f(x), g(x)) = x^2+2x+1 = (x+1)^2$.

14.4. **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 14.5. *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 $g \sim f$. That's as close we can get to the notion of a prime. We cannot expect $g(x) = 1$, or $g(x) = f(x)$, of course. Note that if $g|h$ and $g_1 \sim g$ then $g_1|h$.

Proposition 14.6. *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 $f \sim g$. □

Theorem 14.7. *(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 monic irreducible polynomials $f_1, \cdots, f_g$ of positive degree and positive integers $r_1, \ldots, r_g$ 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 monic irreducible polynomials of positive degree and $s_i > 0$, then $a = b$, $g = t$, and after re-naming the h_i's we have $h_i = f_i$ for all i and $r_i = s_i$ for all i.

Proof. The proof is very similar to the proof for integers. We first prove the existence of factorization. Suppose that there is a non-zero polynomial $f(x)$ with no such factorization. Choose then a non-zero polynomial $f(x)$ of minimal degree for which no such factorization exists. Then $f(x)$ is not a constant polynomial and is not an irreducible polynomial either, else $f(x) = a_n x^n + \cdots + a_0 = a_n \cdot (a_n^{-1} f(x))$ is a suitable factorization. It follows that $f(x) = f_1(x) f_2(x)$, where each $f_i(x)$ has degree less than that of $f(x)$.

Therefore, each $f_i(x)$ has a factorization

$$f_1(x) = c_1 a_1(x) \cdots a_m(x), \qquad f_2(x) = c_2 b_1(x) \cdots b_n(x),$$

with $c_i \in \mathbb{F}$ and a_i, b_j monic irreducible polynomials. 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. 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. 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 $\leq n$ and $\deg(f) = n+1$. Since $a_1(x)|c_2b_1(x)\cdots b_n(x)$ and $a_1(x)$ is irreducible, it follows that $a_1(x)|c_2$ (which is impossible because c_2 is a constant) or $a_1(x)|b_i(x)$ for some i (and in particular we must have $n \geq 1$). 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, $a_i(x) = b_i(x), i = 2, 3, \ldots, n$. □

14.5. **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 14.8. *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 then f has no roots in $\mathbb{F}$.*
(2) *Let $\deg(f) = d$ then f has at most d roots.*

Proof. Suppose that $f(a) = 0$ and divide f by $x - a$ with a residue, getting $f(x) = g(x)(x - a) + r(x)$, where $r(x)$ is either zero or a polynomial of degree less than that of $x - a$. That is, in either case, $r(x)$ is a constant. Substitute $x = a$. We get $0 = f(a) = g(a)(a - a) + r = r$ and so $f(x) = (x - a)g(x)$.

Consider the factorization of f into irreducible monic polynomials:

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

where the f_i are irreducible polynomials of degree larger than 1, 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 $\mathbb{F}$. The Fundamental Theorem of Algebra says that the field of complex numbers $\mathbb{C}$ 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, but we have some partial answers in special cases.

Proposition 14.9. *Let $f(x) = a_nx^n + \cdots + a_1x + a_0$ be a non-constant polynomial with integer coefficients. If $a = s/t$, $(s, t) = 1$, is a rational root of f then $s|a_0$ and $t|a_n$.*

Proof. We have $a_n(s/t)^n + \cdots + a_1(s/t) + a_0 = 0$ and so

$$a_ns^n + a_{n-1}s^{n-1}t + \cdots + a_1st^{n-1} + a_0t^n = 0.$$

Since s divides 0 and s divides $a_ns^n + a_{n-1}s^{n-1}t + \cdots + a_1st^{n-1}$, it follows that $s|a_0t^n$. Then, since $(s, t) = 1$, we get that $s|a_0$. Similarly, t divides 0 and t divides $a_{n-1}s^{n-1}t + \cdots + a_1st^{n-1} + a_0t^n$, so t divides a_ns^n. Now $(s, t) = 1$ implies that $t|a_n$. □

Example 14.10. *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 $2x^4 - 7x^3 + 5x^2 - 7x + 3$. There are thus of the form s/t, where $s = \pm1, \pm3, t = \pm1, \pm2$. We have the possibilities $\pm1, \pm1/2, \pm3, \pm3/2$. By checking each case, we find the roots are $1/2$ and 3. We remark that after having found the root $1/2$ we can divide the polynomial $2x^4 - 7x^3 + 5x^2 - 7x + 3$ by $x - 1/2$ finding $2x^3 - 6x^2 + 2x - 6$, whose roots are the roots of $x^3 - 3x^2 + x - 3$. So, in fact, the only possibilities for additional roots are ±3. We saved this way the need to check if $\pm3/2$ are roots.

Here is another example. *Is the polynomial $x^3 + 2x^2 + 5$ irreducible over $\mathbb{Q}$?* In this case, if it is reducible then one of the factors would have to have degree 1 (this type of argument only works for degrees $1, 2, 3$ polynomials. For higher degree, we might have a reducible polynomial with no linear factor, e.g., $(x^2 + 1)(x^2 + 3)$). Namely, the polynomial would have a rational root. But the rational roots can only be $\pm1, \pm5$ and one verifies those are not roots. Thus, the polynomial is irreducible.

Proposition 14.11. *If $f(x) \in \mathbb{R}[x]$ is a polynomial of odd degree then f has a root in $\mathbb{R}$.*

Proof. Since the roots of f are the roots of $-f$, we may assume that $f(x) = a_nx^n + \cdots + a_1x + a_0$, $a_i \in \mathbb{R}$, $a_n > 0$. An easy estimate shows that there is an $N > 0$ such that $f(N) > 0$ and $f(-N) < 0$. By the intermediate value theorem there is some a, $-N \leq a \leq N$ such that $f(a) = 0$. □

14.6. **Eisenstein's criterion.** (Planned)

14.7. **Roots of polynomials in $\mathbb{Z}_p$.** Let p be a prime and let $\mathbb{Z}_p$ be the field with p elements whose elements are congruence classes modulo p. By Fermat's little theorem, every element of $\mathbb{Z}_p^\times$ 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 $\mathbb{Z}_p$ again with multiplicity one. That is,

$$x^p - x = \prod_{a=0}^{p-1}(x-\bar{a}).$$

Proposition 14.12. *Let $f(x)$ be any polynomial in $\mathbb{Z}_p[x]$. Then $f(x)$ has a root in $\mathbb{Z}_p$ if and only if* $\gcd(f(x), x^p - x) \neq 1$.

Proof. If $f(a) = 0$ for some $a \in \mathbb{Z}_p$ 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 $\mathbb{Z}_p$. In particular, each such a_i is a root of $f(x)$. □

The straightforward way to check if $f(x)$ has a root in $\mathbb{Z}_p$ is just to try all possibilities for x. Suppose that $f(x)$ has a small degree relative to p, say 5, to fix ideas. 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 14.12 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 then 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 $\mathbb{Z}_n$ in the context of $\mathbb{F}[x]$. This is one of our main motivation for studying rings in much more detail.

Part 5. Rings

15. Some basic definitions and examples

Recall our definition of a ring

Definition 15.1. A *ring* R is a non-empty set together with two operations, called "addition" and "multiplication" that are denoted, respectively, by

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

One requires the following axioms to hold:

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

Example 15.2. $\mathbb{Z}$ 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 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 $-(x+y) = -x-y$, where, technically $-x-y$ means $-x+(-y)$.
(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 15.3. Let $\mathbb{F}$ be a field and $n \geq 1$ an integer. Consider the set of $n \times n$ matrices:

$$M_n(\mathbb{F}) = \left\{ \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} : a_{ij} \in \mathbb{F} \right\}.$$

For example:

(1) for $n = 1$ we just get $(a_{11}), a_{11} \in \mathbb{F}$;
(2) for $n = 2$, $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$;

(3) for $n = 3$ we get $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$.

In general we shall write an $n \times n$ matrix as (a_{ij}), or $(a_{ij})_{i,j=1}^n$. 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 15.4. Let ϵ be a formal symbol and $\mathbb{F}$ a field. The *ring of dual numbers*, $\mathbb{F}[\epsilon]$ is defined as

$$\mathbb{F}[\epsilon] = \{a + b\epsilon : a, b \in \mathbb{F}\},$$

with the following addition and multiplication:

$$(a + b\epsilon) + (c + d\epsilon) = a + c + (b + d)\epsilon, \quad (a + b\epsilon)(c + d\epsilon) = ac + (ad + bc)\epsilon.$$

Note that ϵ is a zero divisor: $\epsilon \neq 0$ but $\epsilon^2 = 0$.

Example 15.5. 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 such that $p^2 \nmid n$. 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.

Example 15.6. Let R_1, R_2 be rings. Then $R_1 \times R_2$ is a ring with the following operations:

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2), \quad (a_1, b_1)(a_2, b_2) = (a_1a_2, b_1b_2).$$

The zero element is $(0_{R_1}, 0_{R_2})$ and the identity element is $(1_{R_1}, 1_{R_2})$.

Definition 15.7. 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) $s_1, s_2 \in S \Rightarrow s_1 + s_2 \in S$;
(3) $s_1, s_2 \in S \Rightarrow s_1s_2 \in S$.

Note that in this case S is a ring in its own right. We remark that we require axiom (1) and Hungerford does not.

Example 15.8. 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 this set.

16. Ideals

Definition 16.1. Let R be a ring. A (two-sided) *ideal* I of R is a subset 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 $ar \in I$.

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

Example 16.2. $I = \{0\}$ and $I = R$ are always ideals. They are called the *trivial ideals.*

Example 16.3. Suppose that R is a division ring (e.g., a field) and $I \lhd R$ is a non-zero ideal. Then $I = R$. Indeed, there is an element $a \in I$ such that $a \neq 0$. Then $1 = a^{-1}a \in I$ and so for every $r \in R$ we have $r = r \cdot 1 \in I$. That is, $I = R$. We conclude that a division ring has only the trivial ideals. (Note also that the argument shows for any ring R that if an ideal I contains an invertible element of R then $I = R$.)

Example 16.4. Let R be a commutative ring. Let $r \in R$. The *principal ideal* (r) is defined as

$$(r) = \{ra : a \in R\}.$$

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!), $(ra_1)s = r(a_1s) \in R$.

If every ideal of R is principal, one calls R a *principal ideal ring.*

Example 16.5. *$\mathbb{Z}$ is a principal ideal ring. In fact, the list*

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

is a complete list of the ideals of $\mathbb{Z}$.

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, $I = (i)$.

Example 16.6. *Let $\mathbb{F}$ be a field. The ring $\mathbb{F}[x]$ is a principal ideal ring.*

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, $I \subseteq (f(x))$.

Example 16.7. 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 16.8. The ring of polynomials $\mathbb{C}[x, y]$ in two variables with complex coefficients is not a principal ideal ring. We claim that the set of polynomials $I = \{f(x, y) : f(0, 0) = 0\}$, namely, polynomials with zero constant term, is an ideal that is not principal. We leave that as an exercise.

17. 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) $f(r_1 r_2) = f(r_1) f(r_2)$.

Again, we insist on axiom (1), which Hungerford omits. It doesn't follow from the other requirements: indeed, consider the map

$$f : \mathbb{R} \to \mathbb{R} \times \mathbb{R}, \quad f(r) = (r, 0).$$

This map satisfies $f(r_1 + r_2) = f(r_1) + f(r_2)$ and $f(r_1 r_2) = f(r_1) f(r_2)$, but $f(1) = (1, 0)$ is not the identity element of $\mathbb{R} \times \mathbb{R}$. So this is not a ring homomorphism.

On the other hand, if $S \subset R$ is a subring then the inclusion map $i : S \to R, i(s) = s$, is a ring homomorphism. Note that this explains why in the definition of a subring we insisted on $1_R \in S$.

Proposition 17.1. *Let $f : R \to S$ be a homomorphism of rings. The image of f is a subring of S.*

Proof. It is a consequence of the axioms that $f(0_R) = 0_S$. Indeed, put $s = f(0_R)$ then $s + s = f(0_R) + f(0_R) = f(0_R + 0_R) = f(0_R) = s$. So $-s + (s + s) = -s + s$. That is, $s = 0_S$. Also, by definition $f(1_R) = 1_S$ and so $0_S, 1_S \in \mathrm{Im}(f)$.

Let now $s_1, s_2 \in \mathrm{Im}(f)$, say $s_i = f(r_i)$. Then, $s_1 + s_2 = f(r_1) + f(r_2) = f(r_1 + r_2)$ and so $s_1 + s_2 \in \mathrm{Im}(f)$. Similarly, $s_1 s_2 = f(r_1 r_2)$ and so $s_1 s_2 \in \mathrm{Im}(f)$. □

Definition 17.2. Let $f : R \to S$ be a homomorphism of rings. The *kernel* of f, $\mathrm{Ker}(f)$, is defined as follows:

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

Proposition 17.3. $\mathrm{Ker}(f)$ *is an ideal of R. The map f is injective if and only if* $\mathrm{Ker}(f) = \{0\}$.

Proof. First, since $f(0_R) = 0_S$ we have $0_R \in \mathrm{Ker}(f)$. Suppose that $r_1, r_2 \in \mathrm{Ker}(f)$ then $f(r_i) = 0_S$ and we find that $f(r_1 + r_2) = f(r_1) + f(r_2) = 0_S + 0_S = 0_S$ so $r_1 + r_2 \in \mathrm{Ker}(f)$.

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

We have so far proven that $\mathrm{Ker}(f)$ is an ideal. Suppose now that f is injective. Then $f(r) = 0_S$ implies $f(r) = f(0_R)$ and so $r = 0_R$. That is, $\mathrm{Ker}(f) = \{0_R\}$. Suppose conversely that $\mathrm{Ker}(f) = \{0_R\}$. We first need the following useful fact:

$$f(-r) = -f(r), \quad r \in R.$$

Indeed, we only need to show that $f(r) + f(-r) = 0_S$, but $f(r) + f(-r) = f(r + (-r)) = f(0_R) = 0_S$.

Now, if $f(r_1) = f(r_2)$ then $0_S = f(r_1) - f(r_2) = f(r_1) + (f(-r_2)) = f(r_1 - r_2)$ and so $r_1 - r_2 \in \mathrm{Ker}(f)$. Since $\mathrm{Ker}(f) = \{0_R\}$, 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 17.4. Let $n \geq 1$ be an integer. Define a function,

$$f : \mathbb{Z} \to \mathbb{Z}_n,$$

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 $\mathbb{Z}_n$;

(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 17.5. 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) $f((r_1,s_1)(r_2,s_2)) = f((r_1r_2, s_1s_2)) = r_1r_2 = f((r_1,s_1))f((r_2,s_2))$.

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

Example 17.6. 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 $\mathrm{Ker}(f) \subseteq (\epsilon)$. On the other hand $(c + d\epsilon)\epsilon = c\epsilon$ and that shows $(\epsilon) \subseteq \mathrm{Ker}(f)$.

Example 17.7. 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 the *specialization homomorphism* or the *evaluation homomorphism.* The kernel of α is $\{f \in \mathbb{F}[x] : f(a) = 0\}$ and is equal to the principle 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 14.8 says that $g(x) = (x - a)g_1(x)$ for some polynomial $g_1(x)$ and so $g(x) \in (x - a)$.

17.1. **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 17.8. *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 $r_1y_1 = y_1r_1 = 1, r_2y_2 = y_2r_2 = 1$. Let $y = y_2y_1$ then $(r_1r_2)y = r_1(r_2y_2)y_1 = r_1 \cdot 1 \cdot y_1 = r_1y_1 = 1$. A similar computation gives $y(r_1r_2) = 1$ and so $r_1r_2 \in R^\times$.

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 17.9. We have $\mathbb{Z}^\times = \{\pm 1\}$. We have $\mathbb{Q}^\times = \mathbb{Q} - \{0\}$.

Example 17.10. We have $\mathbb{F}[\epsilon] = \{a + b\epsilon : a \neq 0\}$. Indeed, if $a \neq 0$ then $(a+b\epsilon)(a^{-1} - a^{-2}b\epsilon) = 1$. Conversely, if $(a + b\epsilon)(c + d\epsilon) = 1$ then $ac = 1$ and so $a \neq 0$.

Example 17.11. Let n be a square free integer. We have $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a^2 - b^2n = \pm 1\}$. Indeed, if $a^2 - b^2n = \pm 1$ 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 $ad - bc = 1$ 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 $(a + b\sqrt{n})(a - b\sqrt{n})$ and $(c + d\sqrt{n})(c - d\sqrt{n})$ are integers. So $(a + b\sqrt{n})(a - b\sqrt{n}) = a^2 - b^2n = \pm 1$.

Example 17.12. 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 $(ad-bc)^{-1}\begin{pmatrix} d & -b \\ -c & a\end{pmatrix}$, but we don't really need this fact now).

18. Quotient rings

Consider a surjective ring homomorphism $f : R \to S$. Given an element $s \in S$ let r be an element of R such that $f(r) = s$. How unique is r? If $a \in I := \operatorname{Ker}(f)$ 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

$$r + I = \{r + i : i \in I\}$$

(the is called a coset of I). Then we have proven that

$$f^{-1}(s) = r + I.$$

Thus, in a sense, we may identify elements of S with cosets of R and from this point of view we may 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 18.1. Let R be a ring and $I \lhd R$ a two sided ideal. A *coset* of I is a set of the form

$$a + I := \{a + i : i \in I\},$$

where is a is an element of R.

Example 18.2. Suppose that $R = \mathbb{Z}$ and $I = (n)$ for some positive integer n. Then $a + (n) = \{\dots, a - n, a, a + n, a + 2n, \dots\}$ are precisely the numbers congruent to a modulo n.

Lemma 18.3. *We have the following facts:*

(1) *Every element of R belongs to a coset of I.*
(2) *Two cosets are either equal or disjoint.*
(3) *The following are equivalent: (i) $a + I = b + I$; (ii) $a \in b + I$; (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 $a = a + 0 \in a + I$. 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 $a = a + 0 \in a + I$, but also $a = b + i \in b + I$. That is, $a + I \cap b + I \neq \emptyset$ and so $a + I = b + I$. □

Theorem 18.4. *Let R be a ring and $I \lhd R$ a two-sided ideal. Denote the collection of cosets of I in R by R/I. Define addition by*

$$(a + I) + (b + I) = a + b + I,$$

and multiplication by

$$(a + I)(b + I) = ab + I.$$

These operations are well defined and make R/I into a ring 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 a such that the coset is $a + I$. Namely, if

$$a + I = a' + I, b + I = b' + I,$$

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 the lemma 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 when is it that we use that the axioms hold in R we use $\overset{!}{=}$ at that point.

(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 18.5. *The natural map,*

$$\pi : R \to R/I, \qquad a \mapsto \pi(a) := \bar{a}$$

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 18.6. Consider the ring $\mathbb{Z}$. If we take the ideal $\{0\}$ then $\mathbb{Z}/\{0\}$ can be identified with $\mathbb{Z}$; the map $\mathbb{Z} \to \mathbb{Z}/\{0\}$ is a bijective ring homomorphism. Let $n > 0$ then. The ring $\mathbb{Z}/(n)$ has as elements the cosets $a + (n)$. Two cosets $a + (n), b + (n)$ are equal if and only if $a - b \in (n)$, that is, precisely when $n|(a-b)$. We see that the elements of $\mathbb{Z}/(n)$ are just the congruence classes modulo n and the operations on $\mathbb{Z}/(n)$ are then just the operations we defined on congruence classes.

18.1. **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.

Lemma 18.7. *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 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 18.8. *Let $\mathbb{F}$ be a field , $f(x) \in \mathbb{F}[x]$ a non-constant irreducible polynomial of degree n. Then $\mathbb{F}[x]/(f(x))$ 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 $f(x) \nmid g(x)$ 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 $a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. If $\mathbb{F}$ has q elements, we get q^n such polynomials. □

Example 18.9. A field with 4 elements. Take the field $\mathbb{F}$ to be $\mathbb{Z}_2$ and consider the polynomial $x^2 + x + 1$ over that field. Because it is of degree 2 and has no root in $\mathbb{Z}_2$ it must be irreducible. Therefore, $\mathbb{Z}_2[x]/(x^2 + x + 1)$ is a field $\mathbb{K}$ with 4 elements. Let us list its elements:

$$\mathbb{K} = \{\bar{0}, \bar{1}, \bar{x}, \overline{x+1}\}.$$

We can give tables of addition and multiplication:

$+$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{1}$	$\bar{1}$	$\bar{0}$	$\overline{x+1}$	$\bar{x}$
$\bar{x}$	$\bar{x}$	$\overline{x+1}$	$\bar{0}$	$\bar{1}$
$\overline{x+1}$	$\overline{x+1}$	$\bar{x}$	$\bar{1}$	$\bar{0}$

,

$\cdot$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{x}$	$\overline{x+1}$
$\bar{x}$	$\bar{0}$	$\bar{x}$	$\overline{x+1}$	$\bar{1}$
$\overline{x+1}$	$\bar{0}$	$\overline{x+1}$	$\bar{1}$	$\bar{x}$

Example 18.10. A field with 9 elements. Consider the polynomial $x^2 + 1$ over $\mathbb{Z}_3$. It is quadratic and has no root in $\mathbb{Z}_3$, hence is irreducible over $\mathbb{Z}_3$. We conclude that $\mathbb{Z}_3[x]/(x^2 + 1)$ is a field with 9 elements. Note that in $\mathbb{Z}_3$ the element $-1 = 2$ is not a square. However, in $\mathbb{Z}_3[x]/(x^2 + 1)$ we have $x^2 = x^2 - (x^2 + 1) = -1$ and so -1 is a square now – it's root is x (viewed as an element of $\mathbb{Z}_3[x]/(x^2 + 1)$). In fact, any quadratic polynomial over $\mathbb{Z}_3$ has a root in $\mathbb{Z}_3[x]/(x^2 + 1)$, because the discriminant "$b^2 - 4ac$" is either $0, 1, 2$ and all those are squares in $\mathbb{Z}_3[x]/(x^2 + 1)$.

Example 18.11. 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 $\mathbb{Z}_2$ and conclude that $\mathbb{Z}_2[x]/(x^3 + x + 1)$ 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

$$\mathbb{L} = \mathbb{K}[t]/(t^2 + t + \bar{x}).$$

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

Theorem 18.12. *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 $L \supset F$.

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

18.3. **Roots of polynomials over $\mathbb{Z}_p$.**

We can now continue our discussion, begun in § 14.7, of the efficient determination of whether a small degree polynomial $f(x)$ over $\mathbb{Z}_p$ has a root in $\mathbb{Z}_p$. 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 the 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. 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.

19. The First Isomorphism Theorem

19.1. **Isomorphism of rings.**

Definition 19.1. Let R, S be rings. A ring homomorphism $f : R \to S$ is called an *isomorphism* if f is bijective.

Lemma 19.2. *If $f : R \to S$ is a ring isomorphism then the inverse function $g = f^{-1}$, $g : S \to R$ 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 19.3. Let R, S be rings. We say that R and S are isomorphic if there is a ring isomorphism $R \to S$.

Lemma 19.4. *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 $f : R \to S$ is an isomorphism then $g : S \to R$ is an isomorphism, where g is the inverse function to f. Thus, the relation is symmetric. Now suppose $f : R \to S$ and $g : S \to T$ are ring isomorphisms between the rings R, S, T. To show the relation is transitive we need to prove that $g \circ f : R \to T$ 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. $(g \circ f)(r_1 r_2) = g(f(r_1 r_2)) = g(f(r_1) f(r_2)) = g(f(r_1)) g(f(r_2)) = (g \circ f)(r_1)(g \circ f)(r_2)$.

□

We shall denote the fact that R is isomorphic to S by $R \cong S$.

19.2. **The First Isomorphism Theorem.**

Theorem 19.5. *Let $f : R \to S$ be a surjective homomorphism of rings. Let $I = \ker(f)$ then there is an isomorphism $F : R/I \to S$, 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 $\pi : R \to R/I$ is the canonical map $g \mapsto \bar{g}$.

Proof. We define a function,

$$F : R/I \to S,$$

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 19.6. We consider again the homomorphism $\mathbb{Z} \to \mathbb{Z}_n$. It is a surjective ring homomorphism with kernel (n) and we conclude that

$$\mathbb{Z}/(n) \cong \mathbb{Z}_n,$$

a fact we have noticed somewhat informally before.

Example 19.7. We have $\mathbb{R}[x]/(x^2+1) \cong \mathbb{C}$. To show that, define a ring homomrophism

$$\mathbb{R}[x] \to \mathbb{C},$$

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 also clear that this homomorphism is surjective

The kernel I definitely contains x^2+1, and so all its multiples. That is, I contains x^2+1. Now, $I = (f)$ for some polynomial f. If f is linear then $f = ax + b$ for some $a, b \in \mathbb{R}$ and we get that $ai + b = 0$ and so that $i = -b/a$ is real. Contradiction. It is obvious that f cannot be a constant polynomial either. Because $x^2+1 \in (f)$, $f|(x^2+1)$ and so f must be quadratic and, in fact, $f \sim x^2+1$. Thus, $I = (x^2+1)$ and by the first isomorphism theorem we have

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

19.3. **The Chinese Remainder Theorem.**

Theorem 19.8. *Let m, n be positive integers such that $(m,n) = 1$. Then*

$$\mathbb{Z}_{mn} \cong \mathbb{Z}_m \times \mathbb{Z}_n.$$

Proof. We define a function

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

This function is a ring homomorphism:

- $f(1) = (1 \pmod{m}, 1 \pmod{n}) = (1_{\mathbb{Z}_m}, 1_{\mathbb{Z}_n})$;
- $f(a+b) = (a+b \pmod{m}, a+b \pmod{n}) = (a \pmod{m}, a \pmod{n}) + (b \pmod{m}, b \pmod{n}) = f(a) + f(b)$;
- $f(ab) = (ab \pmod{m}, ab \pmod{n}) = (a \pmod{m}, a \pmod{n}) \cdot (b \pmod{m}, b \pmod{n}) = f(a)f(b)$. The kernel of the map is the set $\{a : m|a, n|a\} = \{a : mn|a\}$ (using that $(m,n) = 1$), that is (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

  $$\mathbb{Z}/(mn) \cong \mathbb{Z}_m \times \mathbb{Z}_n,$$

  and we know already that $\mathbb{Z}/(mn) \cong \mathbb{Z}_{mn}$.

□

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 $a_nA^n + \cdots + a_1A + a_0 \cong 0 \pmod{mn}$ if and only if we have $a_nA^n + \cdots + a_1A + a_0 \cong 0 \pmod{m}$ and $a_nA^n + \cdots + a_1A + a_0 \cong 0 \pmod{n}$. Here is an example:

Example 19.9. *Solve the equation* $5x + 2 = 0$ *modulo* 6.

We consider the equation modulo 2 and get $x = 0 \pmod{2}$; we consider it modulo 3 and get $2x + 2 = 0 \pmod{3}$ and get that $x = 2 \pmod{3}$. There is an $x \in \mathbb{Z}$ such that $x \pmod{2} = 0, x \pmod{3} = 2$ and in fact x is unique modulo 6 (this is the CRT). We can guess that $x = 2$ will do in this case, but it raises the general problem of finding the inverse isomorphism to

$$\mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n.$$

19.3.1. *Inverting* $\mathbb{Z}_{mn} \rightarrow \mathbb{Z}_m \times \mathbb{Z}_n$. It is easy to see that if we know to find elements 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 element $ae_1 + be_2$.

Since $(m, n) = 1$ we may find u, v such that $1 = um + vn$. Put

$$e_1 = 1 - um, \quad e_2 = 1 - vn.$$

Example 19.10. *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.

20. Prime and maximal ideals

(planned)

Part 6. **Groups**

21. First definitions and examples

21.1. **Definitions and some formal consequences.**

Definition 21.1. 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) e 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) d 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) Cancelation: $ab = cb \Rightarrow a = c$, 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) $(ab)^{-1} = b^{-1}a^{-1}$.
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) $(a^{-1})^{-1} = a$.
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

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

21.2. **Examples.**

Example 21.2. The *trivial group* G is a group with one element e and multiplication law $ee = e$.

Example 21.3. 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.

These examples thus include $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{F}, \mathbb{F}[\epsilon], M_2(\mathbb{F}), \mathbb{Z}_n$ with the addition operation.

Example 21.4. Let R be a ring. Recall that the *units* $R^\times$ of R are defined as

$$\{u \in R : \exists v \in R, uv = vu = 1\}.$$

This is a group. If $u_1, u_2 \in R$ with inverses v_1, v_2, respectively, then, as above, one checks that v_2v_1 is an inverse for u_1u_2 and so $R^\times$ is closed under the product operation. The associative law holds because it holds in R; 1_R serves as the identity. If R is not commutative there is no reason for $R^\times$ to be commutative, though in certain cases it may be.

Thus we get the examples of $\mathbb{Z}^\times = \{\pm 1\}, \mathbb{Q}^\times = \mathbb{Q} - \{0\}, \mathbb{R}^\times = \mathbb{R} - 0, \mathbb{C}^\times = \mathbb{C} - \{0\}$, and more generally, $\mathbb{F}^\times = \mathbb{F} - \{0\}$. In addition, $\mathrm{GL}_2(\mathbb{F}) = \{M \in M_2(\mathbb{F}) : \det(M) \neq 0\}, \mathbb{F}[\epsilon]^\times = \{a + b\epsilon : a \neq 0\}$.

Proposition 21.5. *Let $n > 1$ be an integer. The group $\mathbb{Z}_n^\times$ 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

$$\varphi(n) = \begin{cases} 1 & n = 1 \\ |\mathbb{Z}_n^\times| & n > 1. \end{cases}$$

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 that based on the Chinese Remainder Theorem.

Here are some specific examples:

n	$\mathbb{Z}_n^\times$	$\varphi(n)$
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 21.6. If G, H are groups then $G \times H$ is a group with the operation

$$(g_1, h_1)(g_2, h_2) = (g_1g_2, h_1h_2).$$

The identity is (e_G, e_H) and $(g,h)^{-1} = (g^{-1}, h^{-1})$.

21.3. **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 21.7. 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 21.8. 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 21.9. 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 21.10. 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 21.11. The group $\mathbb{Z}$ is cyclic. As a generator we may take 1 (or -1).

Example 21.12. The group $\mathbb{Z}_5^\times = \{1, 2, 3, 4\}$ is cyclic. The elements $2, 3$ are generators. The group $\mathbb{Z}_8^\times$ is not cyclic. One can check that the square of any element is 1.

22. The permutation and dihedral groups

22.1. **Permutation groups.**

Definition 22.1. 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.

Proposition 22.2. *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 $[(f \circ g) \circ h](t) = (f \circ g)(h(g)) = f(g(h(t))) = f((g \circ h)(t)) = [f \circ (g \circ h)](t)$. Thus, as functions, we have $(f \circ g) \circ h = f \circ (g \circ h)$ and so our operation 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 $n \cdot (n-1) \cdot (n-1) \cdots 2 \cdot 1 = n!$ choices. □

Example 22.3. (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.* $Id(1) = 1, Id(2) = 2$.

(b) σ. $\sigma(1) = 2, \sigma(2) = 1$.

We may also represent those permutations as 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:

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

22.2. **Cycles.** There is still more efficient notation for permutations in S_n. Fix $n \geq 1$. A cycle (in S_n) is an expression of the form

$$(a_1\; a_2 \cdots a_t),$$

where $a_i \in \{1, 2, \ldots, n\}$ are distinct elements. This expression is understood as the permutation σ given by

$$\sigma(a) = \begin{cases} a_{i+1} & a = a_i, i < n, \\ a_1 & a = a_n, \\ a & \text{else.} \end{cases}$$

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 22.4. 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 22.5. *Let $g \in G$, $o(g)$ is the minimal positive integer k such that $g^k = e$.*

Proof. Let k be the minimal integer such that $g^k = e$ (∞ if such doesn't exist).

Suppose first that $o(g)$ is finite, say equals r. Then the $r+1$ elements $\{e, g, g^2, \ldots, g^r\}$ cannot be distinct and so $g^i = g^j$ for some $0 \leq i < j \leq r$. It follows that $g^{j-i} = e$ and so $k \leq j - i \leq r$. In particular k is also finite. So r is finite implies k is finite and $k \leq r$.

Suppose now that k is finite. Let n be an integer and write $n = ak + b$ where $0 \leq b < k$. Then $g^n = (g^k)^a g^b = e^b g^b = g^b$. 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 22.6. Let $(a_1\; a_2 \cdots a_t)$ be a cycle. Its order is t.

22.3. **The Dihedral group.** Consider a regular polygon with n sides in the plane, $n \geq 3$, symmetric around $(0,0)$. The *dihedral group D_n* is defined as the symmetries of the polygon. Let us number the vertices of the polygon by $1, 2, \ldots, n$ 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. 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 $360^0/n$ 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 $\theta - 360^0/n$). 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 22.7. *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

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

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 $j \cdot 360^0/n$.

Proposition 22.8. *Let $0 \leq j < n$. The element $y^j x$ is a reflection through the line forming an angle $-j \cdot 360^0/2n$ with the x-axis.*

Proof. The symmetry $y^j x$ is not trivial. If it fixes an angle θ it must be reflection through the line with that angle. Note that $y^j x$ sends the angle θ to $-\theta$ and then adds $-j \cdot 360^0/n$ so the equation is $\theta = -\theta - j \cdot 360^0/n \pmod{360}$. That is $\theta = -j \cdot 360^0/2n$. □

23. The theorem of Lagrange

23.1. **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 23.1. 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 23.2. *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 23.3. The Lemma and its proof should be compared with Lemma 18.3. In fact, since R is an abelian group and an ideal I is a subgroup, that lemma is special case of the lemma above.

Corollary 23.4. *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 23.2 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.

23.2. **Lagrange's theorem.**

Theorem 23.5. *Let G be a finite group and H a subgroup of G. Then,*

$$|H| \mid |G|.$$

Moreover, let $\{g_i : i \in I\}$ be a complete set of representatives for the cosets of H, then $|I| = \frac{|G|}{|H|}$. In particular, the cardinality of I does not depend on the choice of a complete set of representatives. It is called the index *of H in G.*

Proof. We have,

$$G = \coprod_{i \in I} g_i H.$$

Let $a, b \in G$. We claim that the function

$$f : aH \to bH, \quad x \mapsto ba^{-1}x,$$

is a well defined bijection. First, $x = ah$ for some h and so $ba^{-1}x = bh \in bH$ and so the map is well defined. It is surjective, because given an element $y \in bH$, say $y = bh$ it is the image of ah. The map is also injective: if $ba^{-1}x_1 = ba^{-1}x_2$ then multiplying both sides by ab^{-1} we get $x_1 = x_2$.

We conclude that each coset g_iH has the same number of elements, which is exactly the number of elements in $H = eH$. We get therefore that

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

That completes the proof. □

Here are some applications of Lagrange's theorem:

(1) Let G be a finite group of prime order p. Then G is cyclic; in fact, every element of G that is not the identity generates G.

Indeed, Let $g \neq e$. Then $H = \langle g \rangle$ is a non-trivial subgroup. So $|H| > 1$ and divides p. It follows that $|H| = |G|$ and so that $\langle g \rangle = G$.

(2) In a similar vein, we conclude that a group of order 6 say, cannot have elements of order 4, or 5, or any order not dividing 6. This follows immediately from Lagrange's theorem, keeping in mind that $\mathrm{ord}(g) = |\langle g \rangle|$.

24. Homomorphisms and isomorphisms

24.1. **homomorphisms of groups.**

Definition 24.1. Let G, H be groups and

$$f : G \to H,$$

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 24.2. *Let $f : G \to H$ 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 24.3. *Let $f : G \to H$ 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, $g_1 = g_2$. □

24.2. **Isomorphism.**

Definition 24.4. A group homomorphism $f : G \to H$ 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. §19.1.

Example 24.5. *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 $\mathbb{Z}_n$ with addition.

Theorem 24.6. (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 24.7. We were somewhat informal about identifying the permutations of G with S_n. A more rigorous approach is the following.

Lemma 24.8. *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. □

25. Group actions on sets

25.1. **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 $\text{Orb}(s)$ is a subset of S, equal to all the images of the element s under the action of the elements of the group G. We also define the *stabilizer* of s to be

$$\text{Stab}(s) = \{g \in G : g \star s = s\}.$$

Note that $\text{Stab}(s)$ is a subset of G. In fact, it is a subgroup, as Lemma 25.1 states.

25.2. **Basic properties.**

Lemma 25.1. (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* $\text{Orb}(s_1)$.

(2) *Let $s \in S$. The set* $\text{Stab}(s)$ *is a subgroup of G.*

(3) *Suppose that both G and S have finitely many elements. Then*

$$|\text{Orb}(s)| = \frac{|G|}{|\text{Stab}(s)|}.$$

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, $(g_2g_1) \star s_1 = g_2 \star (g_1 \star s_1) = g_2 \star s_2 = s_3$, and hence $s_1 \sim s_3$.

Moreover, by the 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, $\text{Orb}(s_1)$.

(2) Let $H = \text{Stab}(s)$. We have to show that: (i) $e \in H$, (2) if $g_1, g_2 \in H$ then $g_1g_2 \in H$, and (iii) if $g \in H$ then $g^{-1} \in H$.

First, by 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, $(g_1g_2) \star s = g_1 \star (g_2 \star s) = g_1 \star s = s$, which proves that $g_1g_2 \in H$. 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 and the orbit of s. If we show that, then by Lagrange's theorem,

$$|\text{Orb}(s)| = \text{no. of left cosets of } H = \text{index of } H = |G|/|H|.$$

Define a function

$$\{\text{left cosets of } H\} \xrightarrow{\phi} \text{Orb}(s),$$

by

$$\phi(gH) = g \star s.$$

We claim that ϕ is a well-defined bijection. First
Well-defined: Suppose that $g_1H = g_2H$. We need to show that the rule ϕ would give 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, $(g_2^{-1}g_1) \star s = (g_2^{-1}g_2) \star s$ and so $(g_2^{-1}g_1) \star s = e \star s = s$. Therefore, $g_2^{-1}g_1 \in \mathrm{Stab}(s) = H$ and hence $g_1H = g_2H$. □

Corollary 25.2. *The set S is a disjoint union of orbits.*

Proof. The orbits are the equivalence classes of the equivalence relation $\sim$ defined in Lemma 25.1. Any equivalence relation partitions the set into disjoint equivalence classes. □

25.3. **Some examples.**

Example 25.3. Let G be the group of real numbers $\mathbb{R}$. The group operation is addition. Let S be the the set of points on 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 degree r. 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 25.4. 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 25.5. We consider a roulette with n sectors and write $n = i_1 + \cdots + i_k$, for some positive (and fixed) integers $i_1, \ldots, i_k$. We suppose we have different colors $c_1, \ldots, c_k$ and we color i_1 sectors of the roulette by the color c_1, i_2 sectors by the color c_2 and so on. The sectors can be chosen as we wish and so there are many possibilities. We get a set S of colored roulettes.

Now, we turn the roulette a steps clockwise, say, then 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 $\mathbb{Z}_n$ acts on S, a acts on a colored roulette by turning it a steps clockwise, and by saying that we are interested in the number of orbits for this action.

Example 25.6. 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

$$\{e, y^2, y^4, y^6, x, y^2x, y^4x, y^6x\}. \tag{25.1}$$

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 (25.1).

Let H be the stabilizer of s_0. According to Lemma 25.1 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 25.1 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.

26. The Cauchy-Frobenius Formula

Theorem 26.1. (CFF) *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 [10]

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

[10] 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 26.2. 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 $g \star s_1 = s_2$.

Corollary 26.3. *Let G be a group acting transitively on 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. □

26.1. **Some applications to Combinatorics.**

Example 26.4. 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 $\mathbb{Z}_{11}$. 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 $\langle i \rangle = \mathbb{Z}_{11}$ (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 26.5. 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 $\mathbb{Z}_{12}$. 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 26.6. 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 $i = 0, \ldots, 7$	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. [11]

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 $y^6 = (8\ 6\ 4\ \ 2)(7\ 5\ 3\ 1)$.

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

Consider now y^4. It may written in permutation notation as $(1\ 5)(2\ 6)(3\ 7)(4\ 8)$. In any coloring fixed by y^4 each of the cycles $(1\ 5)(2\ 6)(3\ 7)$ and $(4\ 8)$ must be single colored. There are thus $\binom{4}{2} = 6$ possibilities (Choosing which 2 out of the four cycles would be green).

It remains to deal with the elements xy^i. We recall that these are all reflections. There are two kinds of reflections. One may be written using permutation notation as

$$(i_1\ i_2)(i_3\ i_4)(i_5\ i_6)$$

(with the other two vertices being fixed. For example $x = (2\ 8)(3\ 7)(4\ 6)$ is of this form). The other kind is of the form

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

(For example $xy = (1\ 8)(2\ 7)(3\ 6)(4\ 5)$ is of this sort). Whatever is 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.

27. 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 27.1. (Cauchy) *Let G be a finite group of order n and let p be a prime dividing n. Then G has an element of order p.*

Proof. [12] 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 $\mathbb{Z}_p$ on T by cyclic shifts (✠). One may therefore apply **CFF** and get

$$|S| = \frac{n^p - n}{p} + n. \quad (✠)$$

Note that $n \not| |S|$ (✠).

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 a *well defined* action (✠).

[12] Some details are left out. We use the symbol ✠ to point out that a detail is left out and the interested person should fill it in.

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

$$g^p = e.$$

Let k be the order of g. Then $2 \leq k \leq p$. Write $p = qk + r$ with $0 \leq r < k$. Then

$$g^r = g^p \cdot (g^k)^{-q} = e \cdot (e)^q = e.$$

Since k is the minimal positive power such that $g^k = e$ we must have $k|p$. This implies $k = p$ and therefore we found an element of order p (namely, g).

□

28. The first isomorphism theorem for groups

28.1. **Normal subgroups.**

Definition 28.1. 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 28.2. *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 28.3. *Let $f : G \to H$ be a group homomorphism. Then* $\mathrm{Ker}(f)$ *is a normal subgroup of G.*

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

Example 28.4. 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.

28.2. **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 $hb_1 = h''b_1$ 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, $(aH)^{-1} = a^{-1}H$).

28.3. **The first isomorphism theorem.**

Theorem 28.5. *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 28.6. 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 28.7. *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 is 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 $\mathbb{Z}_2$.

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 $S_3/A_3 \cong \mathbb{Z}_2$

28.4. **Groups of low order.**

28.4.1. *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.

28.4.2. *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 28.8. *Every group G of prime order p is isomorphic to $\mathbb{Z}_p$; it has no subgroups apart from the trivial subgroups $\{e_g\}, G$.*

In particular, this corollary applies to groups of order $2, 3, 5, 7$.

28.4.3. *Groups of order 4.* Let G be a group of order 4.
First case: G is cyclic.

In this case we have $G \cong \mathbb{Z}_4$. Its subgroups are $\{0\}, \mathbb{Z}_4$ and $H = \langle 2 \rangle = \{0, 2\}$. There are no other subgroups because if a subgroup J contains an element g is 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 $G/H \cong \mathbb{Z}_2$.

Second case. G is not cyclic.
Claim: *Every element of G different from e_G has order* 2.
Proof: we have $\mathrm{ord}(g) = |\langle g \rangle|$ and it divides $|G|$. So, in our case, $\mathrm{ord}(g) = 1, 2$ or 4. If $\mathrm{ord}(g) = 4$, we get that G is cyclic and if $\mathrm{ord}(g) = 1$ then $g = e_G$. Thus, we must have $\mathrm{ord}(g) = 2$.

Claim: *Let G be a group in which every element different from the identity has order* 2. *Then G is commutative.*
Proof: Note first that if $a \in G$ has order 2 (or is the identity) then $aa = e_G$ and so $a^{-1} = a$. Now, we need to show that for every $a, b \in G$ we have $ab = ba$. But this is equivalent to $ab = b^{-1}a^{-1}$. Multiply both sides by ab and we see that we need to prove that $abab = e_G$. But, $abab = (ab)^2$ and so is equal to e_G, by assumption.

One example of a group of order 4 satisfying all these properties is $\mathbb{Z}_2 \times \mathbb{Z}_2$. We claim that $G \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. Pick two distinct elements g_1, g_2 of G that aren't the identity either. Define a map

$$f : \mathbb{Z}_2 \times \mathbb{Z}_2 \to G, \qquad f(a, b) = g_1^a g_2^b.$$

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 is also injective.

The non-trivial subgroups of $\mathbb{Z}_2\times\mathbb{Z}_2$ 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 $\mathbb{Z}_2$.

28.4.4. *Groups of order 6.* We know three candidates already $\mathbb{Z}_6,\mathbb{Z}_2\times\mathbb{Z}_3$ and S_3. Now, in fact, $\mathbb{Z}_6\cong\mathbb{Z}_2\times\mathbb{Z}_3$ (say by CRT). And since S_3 is not commutative it is not isomorphic to $\mathbb{Z}_6$. In fact, every group of order 6 is isomorphic to either $\mathbb{Z}_6$ or S_3. We don't prove it here.

The subgroups of $\mathbb{Z}_6$: Let n be a positive integer. We have a surjective group homomorphism $\pi:\mathbb{Z}\to\mathbb{Z}_n$. 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 $\mathbb{Z}_n$. 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 $\mathbb{Z}_n$ 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 $\mathbb{Z}_6,\{0,2,4\},\{0,3\},\{0\}$. They are all normal and the quotients are isomorphic respectively to $\{0\},\mathbb{Z}_2,\mathbb{Z}_3,\mathbb{Z}_6$.

The subgroups of S_3: Those were classified above.

28.5. **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

$$\mathrm{sgn}: S_n\longrightarrow\{\pm1\}$$

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).

28.6. **Odds and Ends.**

Example 28.9. We prove that *in $\mathbb{Z}_p$ any element is a sum of two squares.*

Clearly this holds for $p = 2$, so we assume $p > 2$. To begin with, $\mathbb{Z}_p^\times = \{1, \dots, p-1\}$ is a group under multiplication; it has $p-1$ elements. Consider the homomorphism:

$$sq : \mathbb{Z}_p^\times \to \mathbb{Z}_p^\times, \qquad sq(x) = x^2.$$

Let H be its image – a subgroup of $\mathbb{Z}_p^\times$. The kernel of sq is the solutions to $x^2 = 1$, which are precisely ± 1. Note that $1 \neq -1$. It follows that $H \cong \mathbb{Z}_p^\times / \{\pm 1\}$ 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 $\mathbb{Z}_p$ with $(p+1)/2$ elements consisting of all squares.

Let $a \in \mathbb{Z}_p$ then the two sets H^* and $a - H^* := \{a - h : h \in H^*\}$ have size $(p+1)/2$ and so must intersect (because $\mathbb{Z}_p$ 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 28.10. *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, $g_1 \star (g_2 \star t) = \rho(g_1)(\rho(g_2)(t)) = (\rho(g_1) \circ \rho(g_2))(t) = \rho(g_1 g_2)(t) = (g_1 g_2) \star t$. □