Lecture 1 — Groups

Dr. D. Bernhard

In this lecture: groups (axioms and examples) — the group $\mathbb{Z}_n$ — order of a group — division with remainder — equivalence relations — computing in $\mathbb{Z}_n$.

Learning outcomes: After this lecture and revision, you should be able to

- Understand what a group is and check for simple examples whether an example is a group or not.
- Add and invert in the group $\mathbb{Z}_n$ for any integer $n > 0$.
- Divide integers with remainder.
- Understand the C $+$, $-$, / and % operators on unsigned integer datatypes.
- Understand what an equivalence relation is and check for simple examples whether or not a relation is an equivalence relation.

⋄ More advanced material that may help interested students to further understand the topics but can safely be skipped the first few times you read the notes is marked by the ⋄ symbol and set in a smaller font, just like this.

Note on exercises. All exercises are graded with one to three stars according to the following scheme.

⋆ A simple comprehension exercise. If you have understood the topic that this question is about, you should be able to answer it in a matter of minutes or even less than that.

⋆⋆ A standard exercise, may require a bit of thinking but should not take hours to solve if you've understood what it's about. At the end of the course and after revision, you should be able to solve two star exercises without too much trouble.

⋆⋆⋆ An advanced exercise. Three star exercises can require a lot of work or a deeper understanding of the mathematical principles behind an idea. Leave these for last. You are not meant to solve all of these, nor will you be able to unless you have a lot of spare time.

Before we start: in this course, the natural numbers $\mathbb{N}$ are the numbers $0, 1, 2, 3, \ldots$ and the integers $\mathbb{Z}$ are the natural numbers together with the negative whole numbers: $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$

1 Groups

The topic of today's lecture is the mathematical structure called a group and the construction of a particular kind of group called $\mathbb{Z}_n$.

1.1 Motivation

A group is something "a bit like adding numbers" — there are elements which are a bit like numbers and there is something you can do to them that is a bit like adding.

There are a lot of things "like numbers", for example:

- Numbers, including variations such as integers, rational numbers, real numbers, complex numbers etc.
- Integer datatypes as used in most programming languages with a fixed bitlength. Actually, we'll mostly be sticking to unsigned integers in this course.
- Pairs, triples and other tuples or vectors of any of the above kinds of numbers.

Let's recall some things we know about addition:

1. You can add numbers in any order and get the same result. If you're asked to mentally add $5 + 3 + 7$, one way to do this if you know that $3 + 7 = 10$ is can do the sum as $5 + (3 + 7) = 5 + 10 = 15$.
2. You can cancel additions on both sides of equations. If you know that $x + 5 = 10 + 5$, you can write $x \cancel{+5} = 10 \cancel{+5}$ to get $x = 10$.

Anyone writing an optimizing compiler will be spending a lot of time on such rules to turn arithmetic expressions in programs into efficient machine code. But there is a big difference between numbers as we know them and numbers as most languages' int(eger) types use them: computer integers have a fixed size, for example 64 bits, and integers can "wrap around": by computing $1 + 1 + \ldots$, you eventually end up at 0 again.

Let's look at 8-bit unsigned integers. The smallest such value is 0 and the largest $2^8 - 1 = 255$, represented by binary $1111\ 1111$. There are many differences to normal integers. For example, in 8-bit integers it does not make sense to speak of "positive" and "negative" numbers: $128 + 128 = 0$ so we can't call 128 "positive" or "negative" according to the usual rules (the sum of positive/negative numbers is again positive/negative). Multiplying 8-bit integers is even less like working with the usual integers. For example, $16 \cdot 16 = 0$ so the product of two non-zero values can become zero. Also, $16 \cdot 32 = 0$ so we can't cancel in multiplications any more: $16 \cdot x = 16 \cdot y$ does not imply $x = y$. Do the two rules we gave above still hold for 8-bit integers? For example, if we know $16 + x = 16 + y$ in 8-bit integers, can we still conclude that $x = y$? It turns out that we can. Mathematically, this difference between adding and multiplying in 8-bit integers can be expressed by saying that for the usual integers, both addition and multiplication (excluding 0) are group operations; for the 8-bit integers addition is a group operation but multiplication is not (whether or not you exclude 0).

A quick outlook on where we're going in the Algebra section of this course: our big question will be, can we construct operations on the set of 8-bit unsigned integers (or any other bitlength for that matter) such that we can add, subtract, multiply and divide (except by 0) according to more or less the usual rules? It will turn out that we can and that there is "essentially" only one way to do this. What "essentially" means here we will also investigate.

1.2 Definition of a group

Half of learning Algebra is deciphering the notation. We give the definition of a group and an alternative "programmer's notation" version.

> **Definition 1.1 (group).** A group $\mathbb{G} = (G, +)$ consists of a set G and an operation $+ : G \times G \to G$ that has the following properties.
>
> **associative** For any elements g, h, k of G we have $(g + h) + k = g + (h + k)$.
>
> **neutral element** There is an element e of G with the property that for all elements g of G we have $e + g = g$ and $g + e = g$. We call such an element a neutral element of the group $(G, +)$.
>
> **inverses** For any element g of G there is an element h of G such that $g + h = e$ and $h + g = e$ where e is the neutral element. We call h the inverse of g and write $h = (-g)$.

An alternative definition of a group is a datatype or "class" `G` with an equality operation[1]`==` and a function with signature (in C style notation) `G add(Ga, G b)` with the following properties.

associative For any elements `g, h, k` of `G` we have `add(add(g, h), k) == add(g, add(h, k))`.

neutral element There is a function `G neutral(void)` with the property that for all `g` of type `G` we have `g == add(g, neutral())` and `g == add(neutral(), g)`.

inverses There is a function `G invert(G x)` such that for any element `g` of type `G` we have `add(g, invert(g)) == neutral()` and `add(invert(g), g) == neutral()`.

[1] For the usual equality operation that we are used to this "just works". The actual condition is that the equality operation must be an equivalence relation, but this requires a bit of formalism to do correctly which we'll sweep under the carpet for now.

In this definition it is important that the functions `add`, `neutral` and `invert` are functions in the mathematical sense and not “procedures”, i.e. they maintain no state between calls and do not have access to an environment such as a source of randomness. Whether you write the group operation as a function `add` or an infix operator $+$ is a matter of notation and does not change whether or not something forms a group.

These definitions immediately give the following rules for working in groups. (They are not hard to prove but we leave this as an exercise.)

Proposition 1.2. In a group:

1. There can only be one neutral element. If two elements in a group both have the properties of a neutral element, then they are equal.
2. An element can only have one inverse. If $g + h = e$ and $g + k = e$ then $h = k$. The same holds if $h + g = e$ and $k + g = e$.
3. More generally, you can cancel in equations over groups: if $g + h = g + k$ then $h = k$. Similarly, $h + g = k + g$ implies $h = k$ too.

Exercise. ($\star\star\star$) *Rules in groups.* Prove Proposition 1.2.

Note: This is a three-star exercise and therefore strictly OPTIONAL. Being able to do axiomatic proofs is not part of the requirements for this course.

1.3 Associativity and commutativity

If you are given a sum like $1+2+3$, do you stop to wonder if it’s meant to be $(1+2)+3$ or $1+(2+3)$? Probably not, you know that this “doesn’t matter” — which is just another way of saying that normal addition is associative. However, $(1-2)-3$ is not the same as $1-(2-3)$. Compilers need to be told things like this. Most programming languages have a table of all their infix operators somewhere, each listed with its precedence and whether it is left- or right-associative. In the C language, $+$ and $-$ are left-associative so $a+b+c$ will be compiled as $(a+b)+c$ and $a-b-c$ is $(a-b)-c$ as you would expect. Since $+$ on integers is associative, an optimizing compiler is free to rearrange the sum. For the purpose of this course, the terms “left-associative” and “right-associative” don’t exist — either an operation is associative in which case it makes no difference, or an operation is not associative and we will bracket it or not write it “infix” at all.

Something that we have omitted to mention until now is that $a + b = b + a$ on normal and fixed-bitlength integers. This is not the same thing as asssociativity — it is a new property called commutativity and we will see groups later on that are not commutative.

Definition 1.3. A group $(G, +)$ is commutative if for all a, b in G we have $a+b = b+a$.

A group that is commutative is sometimes also called Abelian after the mathematician N. H. Abel.

Exercise. $(\star)$ *Basic groups.*

1. Why is $(\mathbb{N}, +)$ not a group, which group laws does it obey and which ones not? (Note: a structure that obeys the same subset of group laws as $(\mathbb{N}, +)$ is called a monoid.)
2. Why is $(\mathbb{Q}, \cdot)$ — the rational numbers with multiplication — not a group?
3. Why can you not have a group with 0 elements?
4. Can you have a group with one element?
5. Describe what a group with two elements must look like.

1.4 The groups $(\mathbb{Z}, +)$ and $(\mathbb{Z}_n, +)$

The first example of a group is $(\mathbb{Z}, +)$, the usual integers with addition. The group axioms are just some of the usual rules of arithmetic. Taking pairs, triples or any other vectors of fixed length of integers with component-wise addition (i.e. $[2, 3] + [4, 5] = [2 + 4, 3 + 5] = [6, 8]$ etc.) also gives a group. For vectors of integers of length n, this group is called $(\mathbb{Z}^n, +)$. The natural numbers $\mathbb{N}$ with addition are not a group, whether or not you include 0.

A slightly more interesting group is the group $(\mathbb{Z}_n, +)$ for any natural number n which has as its elements the numbers $\{0, 1, \dots, n-1\}$. Addition works as follows: add two numbers normally; if the result is n or larger then you “wrap around”: subtract n until you get a valid group element again. In the group $(\mathbb{Z}_8, +)$ for example, $2 + 2$ is still 4 but $6 + 6$ would normally give 12 so we subtract 8 and also get 4. This is a group (for any $n > 0$) but it takes a bit of number theory to show it. The neutral element is always 0 and the inverse of a number x in $(\mathbb{Z}_n, +)$ is the number $n - x$. The inverse of 7 in $(\mathbb{Z}_8, +)$ for example is $8 - 7 = 1$ since $1 + 7$ is 0 in this group.

An `unsigned int(eger)` datatype with n bits and its addition operation typically implement the group $(\mathbb{Z}_{2^n}, +)$.

WARNING. The $+$ in $(\mathbb{Z}_n, +)$ is a different operation to the $+$ in $(\mathbb{Z}, +)$! For $\mathbb{Z}_8$, the former operation has $6 + 6 = 4$ whereas the latter has $6 + 6 = 12$. If we wanted to be really precise, we could use a different symbol like $+_8$ for the operation in $\mathbb{Z}_8$ but we won’t usually do this.

Exercise. ($\star\star$) *Basic properties of groups.* Over the integers $\mathbb{Z}$, complete the following table, placing a tick or a cross in each cell to indicate whether or not the operator has the property:

	$+$	$-$	$\times$	$\div$
associative				
commutative				
has neutral el.				
has inverses				
is group operation				

Exercise. ($\star\star$) *Addition tables.* For a finite group, we can describe the group operation by means of an addition table. Complete the following tables to make the operations into group operations. Note: there is exactly one way to do it in each case. You may use associativity to find this way but you do not have to check it everywhere!

$+$	♠	♣	♢	♡
♠				
♣	♣			
♢			♠	
♡				♠

$+$	A	B	C	D	E
A			B		
B					
C		C	E		D
D			A		
E					

- ($\star\star$) Which rules have you used to fill in the tables?
- ($\star\star\star$) How would you derive these rules from the definition of a group?

1.5 Order of a group

While we're working on groups we'll throw in the concept of an order:

> **Definition 1.4.** The order of a group $\mathbb{G}$ is defined as follows: if the group's set G has a finite number n of elements, the order is n. If the group has infinitely many elements, the order is infinite (written ∞).

So $(\mathbb{Z}, +)$ has order ∞ and $(\mathbb{Z}_n, +)$ for $n > 0$ has order n.

Exercise. ($\star\star$) *Labelling with numbers.* For the two groups in the last exercise (addition tables), in one of them the elements can be labelled with numbers $0, 1, 2, \ldots$ such that group addition is exactly addition modulo the group order on these labels.

- Find a labelling for one of the two groups.
- Give a counter-argument to show why the other group cannot be labelled this way.

1.6 The "modulo n" operation

To better understand the group $(\mathbb{Z}_8, +)$ and how it relates to the better-known $(\mathbb{Z}, +)$, imagine all natural numbers written out in $n = 8$ columns (this works for any n of course):

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1	2	3	4	5	6	7
8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23
...							

This table gives us a new way to compute in $\mathbb{Z}_8$: to add two numbers, add them normally, look up the result in the table then take the column heading as the result. For example sum above, $6 + 6 = 12$ and 12 is in column **4** so the result is 4. This lookup-column-name operation is important enough that we give it a name: to look up the column of a number x in an 8-column table like this is called "x modulo 8", written "$x \pmod 8$". Since going up one cell in the table subtracts 8, this table does the same job as the "subtract 8 until below 8" rule earlier.

This table is also an example of an equivalence relation, which we will discuss later.

1.7 Division with remainder

Neither the "repeatedly subtract 8" rule nor the table is really useful when we have to deal with large numbers. Instead, we invoke a fact of number theory.

Proposition 1.5 (Division with remainder). For any integer a and any positive natural number b there is exactly one pair (q, r) where q is any integer and r is an element of $\{0, 1, 2, \ldots, b-1\}$ making the equation $a = q \cdot b + r$ hold.

We call q the quotient and r the remainder of dividing a by b and write $q = a \text{ div } b$ and $r = a \bmod b$. In programming languages, these operations are often written $q = a \,/\, b$ and $r = a \,\%\, b$.

If we take a b-column table as above, quotient and remainder become row and column (starting both counts with 0 and including extra rows above for the negative integers). For example, in the 8-column table the number 12 can be found in row 1 and column 4:

	col 0	**col 1**	**col 2**	**col 3**	**col 4**	**col 5**	**col 6**	**col 7**
	...							
row -1	(-8)	(-7)	(-6)	(-5)	(-4)	(-3)	(-2)	(-1)
row 0	0	1	2	3	4	5	6	7
row 1	8	9	10	11	$\boxed{12}$	13	14	15
row 2	16	17	18	19	20	21	22	23
	...							

Which is another way of saying $12 \text{ div } 8 = 1$ and $12 \bmod 8 = 4$, giving us $12 = 1 \cdot 8 + 4$. The operation "12 mod 8" that we defined in proposition 1.5 does the same as $12 \pmod 8$ which we defined with the table.

We have defined three equivalent ways of computing modulo 8 (or modulo any other integer $n > 1$):

- Compute normally then add/subtract 8 repeatedly until you hit something between 0 and 7 (which is $8 - 1$).
- Compute normally then look up the result in an 8-column table and take the column number (starting to count with column 0, of course).
- Compute normally then do a division by 8 and take the remainder.

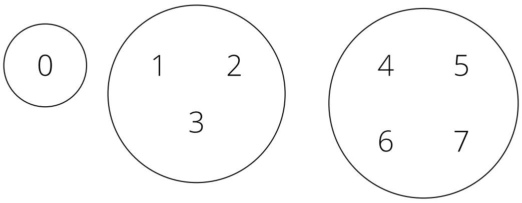
The last method is of course the "official" one. If an integer a leaves remainder 0 when dividing with remainder by b, we say that "a is a divisor of b", "a divides b" or "b is a multiple of a".

1.8 Equivalence relations and classes

As a next step working towards the formal definition of $(\mathbb{Z}_n, +)$ (which will come with a proof that it is actually a group), we introduce equivalence relations and classes.

A relation R on a set S can be thought of as a function that takes two elements a, b of S and outputs a boolean: `true` if a, b are "in relation R" and `false` if a, b are "not in relation R". There are, as always in Algebra, at least two different notations. It should be clear what $R(a, b) = \texttt{true}$ means; this is somtimes simply written $R(a, b)$ and pronounced "a, b are in relation R". The opposite is written $\neg R(a, b)$. Another way of writing a relation is as an infix operator: $a\,R\,b$. This notation is common for relations like $=, \leq, <$: we usually write $a \leq b$ and not $\leq (a, b)$.

An equivalence relation is a relation that divides a set up into different subsets, called "classes". For example, here is one way to divide up the integers 0 to 7 into three classes:



The relation R that creates these classes returns `true` on any two numbers in the same class and `false` otherwise, e.g. $2\,R\,3$ holds but not $2\,R\,4$ (we could write this $2 \not R 4$, in the same way that we write $\neq$ to mean that the $=$ relation does not hold).

For a relation to split a set into classes, it must satisfy three conditions.

Definition 1.6 (equivalence relation). A relation R on a set S is an equivalence relation if it has these three properties.

reflexive For any element a of S we have $a\,R\,a$.

symmetric For any elements a, b of S if $a\,R\,b$ then also $b\,R\,a$.

transitive For any elements a, b, c of S if $a\,R\,b$ and $b\,R\,c$ then also $a\,R\,c$.

The smallest equivalence relation on any set S (that has the fewest number of inputs which produce the output `true`) is the equality relation $=$. The largest equivalence relation is the relation R that always returns `true`, so all elements of S end up in the same class.

If we have a set S and an equivalence relation $\sim$ (a symbol commonly used for equivalence relations), we can form a new set from the classes formed by $\sim$. This is called taking "S modulo $\sim$" and written $S/\sim$. We will soon see that taking an integer modulo another is a special case of taking a set modulo an equivalence relation. For the example relation in the diagram above, if we call the relation $\sim$ then we have

$$\{0, \dots, 7\}/\sim \quad = \quad \{\{0\}, \{1, 2, 3\}, \{4, 5, 6, 7\}\}$$

so we have built a set with three elements, which are themselves sets. We could consider a function c that sends each element of S to its class, so $c(1) = c(2) = \{1, 2, 3\}$. There is a better way to work with classes however: choosing one element to represent each class.

1.9 Representative elements and $\mathbb{Z}_n$

Often it is helpful to pick names for the equivalence classes which are based on one of the elements that they contain. This is called choosing representative elements. Back to our 8-column table for an example:

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
...							
(-8)	(-7)	(-6)	(-5)	(-4)	(-3)	(-2)	(-1)
0	1	2	3	4	5	6	7
8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23
...							

We can interpret this table as an equivalence relation $\sim_8$ which divides the integers into 8 classes (the 8 columns), i.e. $a \sim_8 b$ if and only if a and b are in the same column. In this example, $c(1) = \{\ldots, (-7), 1, 9, 17, \ldots\}$. We can choose the "row zero" as representative elements, i.e. we give the 0-th column the label 0 and so on. We write the map that takes an element x to the representative element of its class as $[x]$.

> **Definition 1.7.** For a fixed modulus n, the operation $[x]$ is the map $\mathbb{Z} \to \mathbb{Z}$ that sends x to its remainder modulo n. If necessary, we can also write $[x]_n$ to make the modulus clear.

◇ Formally, choosing representative elements means choosing a function Z from $\mathbb{Z}/\sim_8$ to $\mathbb{Z}$ that maps each class (which is a subset of $\mathbb{Z}$) to one of its elements. For example, $Z(c(1)) = Z(\{\ldots, (-7), 1, 9, 17, \ldots\})$ $= 1$. If it is clear which representatives are meant, we can omit making this step explicit and just write $\mathbb{Z}/\sim_8 = \{0, 1, 2, 3, 4, 5, 6, 7\}$. The [ ] function is just the composition $Z \circ c$.

1.10 Computing in $\mathbb{Z}_n$

To compute in $\mathbb{Z}_n$, the basic idea is "compute normally and reduce modulo n". We can optimize when we reduce: as long as we reduce at the end, we can reduce as little or as much as we want in between. For example, we can reduce whenever our intermediate results get too big for our liking.

> **Proposition 1.8.** For any $a, b \in \mathbb{Z}$ we have
>
> - $[[a]] = [a]$.
> - $[a + b] = [[a] + b] = [[a] + [b]]$.
> - $[[a + b] + c] = [(a + b) + c]$.

Since addition in $\mathbb{Z}$ is commutative, rules 2 and 3 also give $[a + b] = [a + [b]]$ and $[a + [b + c]] = [a + (b + c)]$. The important thing to remember in all rules is that you can never eliminate the outermost [ ] reduction. The proof of all of these rules is via the official definition of division with remainder. We discuss this in more detail in the next, optional section.

Exercise. *Division with remainder.*

1. ($\star$) Divide 256 by 15 with remainder.
2. ($\star$) Find an example of a modulus n and two numbers a, b such that $[a]+[b] \neq [a + b]$ where [ ] is taking the remainder modulo n.
3. ($\star\star$) For a fixed positive integer n, we know that every number $z \in \mathbb{Z}$ has exactly one pair (q, r) such that $r \in \{0, 1, \ldots, n-1\}$ and $z = q \cdot n + r$. Let's write $[[x]]$ for the operation that sends z to its pair (q, r).

   The pairs (q, r) of this form constitute a group with the "obvious" addition: call the addition in this group ++. Then for any numbers z, z' in $\mathbb{Z}$, if $(q, r) = [[z]]$ and $(q', r') = [[z']]$ then $(q, r){+\!+}(q', r') = [[z + z']]$.

   Describe the addition operation in this group in terms of addition in $\mathbb{Z}$ and reduction modulo n. What is the neutral element, what is the inverse of an element?

1.11 $\diamond$ Formal construction of $\mathbb{Z}_n$

$\diamond$ Using the equivalence relation $\sim_n$ we can define $(\mathbb{Z}_n, +_n)$ formally:

Definition 1.9 ($\mathbb{Z}_n$). Let $n > 0$ be a positive integer. The equivalence relation $\sim_n$ is defined as $a \sim_n b$ if and only if a and b leave the same remainder when dividing by n.

The set $\mathbb{Z}_n$ is this set $\{0, 1, \ldots, n-1\}$ of representatives of the equivalence classes of $\mathbb{Z}/\sim_n$.

The group $(\mathbb{Z}_n, +_n)$ is the set $\mathbb{Z}_n$ with the following addition: $a +_n b := [a + b]$.

We have chosen to write $+_n$ to make clear that we are defining a new operation that differs from the usual $+$ on $\mathbb{Z}$ (in fact, it differs even on $\mathbb{Z}_n$). In the last definition, we claimed that $\sim_n$ was an equivalence relation on $\mathbb{Z}$ and that $(\mathbb{Z}_n, +_n)$ defined this way was a group. This section proves the above claims.

Proposition 1.10. The relation $\sim_n$ is an equivalence relation on $\mathbb{Z}$ for any integer $n > 0$.

Proof. We need to check the three properties. The relation $\sim_n$ is certainly reflexive because every number has the same remainder as itself when dividing by n. It is also symmetric because if r is the a and s is the remainder of b and $r = s$ then $s = r$ too (in other words, $=$ is symmetric). Transitive takes a tiny bit more work. Let's take any integers a, b, c with $a \sim_n b$ and $b \sim_n c$. Then we can write a in exactly one way as $a = \alpha n + r$ for $r \in \{0, \ldots, n-1\}$ and b in exactly one way as $b = \beta n + s$ and c in exactly one way as $c = \gamma n + t$. Since $a \sim_n b$ we have $r = s$ and since $b \sim_n c$ we have $s = t$, from which we conclude $r = t$ (since $=$ is transitive) so $a \sim_n c$ too. QED.

Proposition 1.11. For an integer $n > 0$, $(\mathbb{Z}_n, +_n)$ is a group.

Proof. We have three conditions to check. For associativity, the equation to check is

$$a +_n (b +_n c) = (a +_n b) +_n c$$

We can write this as $[a + [b + c]] = [[a + b] + c]$. The idea is to eliminate the inner [ ] in both cases, then use associativity of $+$ on $\mathbb{Z}$, as we sketched in the rules in the last section.

To show $[x + y] = [x + [y]] = [[x] + [y]]$, write $x = \xi \cdot n + r$ and $y = \eta \cdot n + s$ in the unique way with $r, s \in \mathbb{Z}_n$. Then $[x + y]$ is the remainder of $(\xi + \eta) \cdot n + (r + s)$ modulo n which does not depend on ξ, η but is simply $[r + s]$. (Exercise: why do we need these brackets and can't simply write the remainder as $r + s$?). The term $[x + [y]]$ can be seen to be the remainder of $\xi \cdot n + r + s$ modulo n and again the ξ component does not influence the remainder so we get $[r + s]$, the same for $[[x] + [y]]$.

Back to our equation $[a + [b + c]] = [[a + b] + c]$: if r, s, t are the remainders of a, b, c modulo n then as a direct consequence of the above rule, this equation is equivalent to $[r + (s + t)] = [(r + s) + t]$ and the two sides are the same since $+$ on $\mathbb{Z}$ is associative.

The neutral element is the class $[0] = 0$: for any element a of $\mathbb{Z}_n$, we have $a +_n [0] = [a + [0]] = [a + 0] = [a]$ and $[0] +_n a = [[0] + a] = [0 + a] = [a]$, using that 0 was netural under $+$ in $\mathbb{Z}$.

The inverse of an element a in $\mathbb{Z}_n$ is $[n - a]$. We show $a +_n [n - a] = [0]$: this sum is $[a + n - a] = [n] = [0]$ and the same argument shows $[n - a] +_n a = [0]$. QED.

Lecture 2 — Subgroups

Dr. D. Bernhard

In this lecture: subgroups — generators — order of elements — Lagrange's theorem — permutations as groups — cycles

Learning outcomes: After this lecture and revision, you should:

- Understand subgroups and their possible sizes and be able to apply this knowledge to compute them in any finite group.
- Be able to check whether or not a subset of a group is a subgroup, and extend a subset to the smallest subgroup that contains it if not.
- Understand the order of elements in a group and be able to compute the possible orders of elements of a finite group and find an element of a given order (in particular a generator) if one exists.
- Understand the structure of the set S of permutations on a set A as a group and be able to compose and invert permutations.
- Be able to decompose permutations into disjoint cycles, compose cycles and compute permutations from cycles.

2 Subgroups

We continue our quest to understand groups and how they work. If we are given any structure $(G, +)$ and told that is a group, we already know a lot about how the $+$ operation can behave — today, we look at what happens when the same $+$ operation is used on subsets of G and we look at another way to get examples of groups, namely permutations.

2.1 Opening example

Start with the group $\mathbb{G} = (\mathbb{Z}_{32}, +_{32})$. If we pick a subset H of the set $\mathbb{Z}_{32}$ with the same operation, do we get a group? Here's a table of elements:

0	1	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

Let's try the first column $H = \{0, 8, 16, 24\}$. The group operation on these elements has the table

$+_{32}$	0	8	16	24
0	0	8	16	24
8	8	16	24	0
16	16	24	0	8
24	24	0	8	16

This is a group operation, so $(H, +_{32})$ is a group. But if we tried the subset $H' = \{0, 1, 2, 3, 4, 5, 6, 7\}$, we would not get a group: $6 +_{32} 6 = 12$ which is not in H' anymore, for example.

2.2 Definition of a subgroup

> **Definition 2.1.** Let $\mathbb{G} = (G, +)$ be any group. Let H be a subset of G. If $\mathbb{H} = (H, +)$ is a group for the same operation $+$, we say that $\mathbb{H}$ is a subgroup of $\mathbb{G}$.

When we defined groups, we defined $+$ to be an operation $G \times G \to G$ so by definition, adding two group elements produced another group element. If we take a subset H but the same operation, even if we add two elements of H we have no guarantee that we will end up in H again. So what the definition of subgroups is really about is that doing group operations on H-elements lands in H again. To check that something is a subgroup we do not need to check that the operation is associative etc. — we already know this because it is a group operation. Instead, we have three new rules:

> **Proposition 2.2.** If $(G, +)$ is a group and H is a subset of G then $(H, +)$ is a subgroup of G if and only if the following three conditions hold.
>
> 1. The neutral element e is in H.
> 2. For any two elements a, b in H their sum $a + b$ is also in H.
> 3. For any element a in H, its inverse $(-a)$ is also in H.

Another way of stating this is saying that a subgroup is formed by a subset that is closed under the group operation (using the group operation you cannot "escape").

Let's take the group $(\mathbb{Z}, +)$ and the subset $H = \{0, 1, \ldots, n-1\}$ for some $n > 0$. Using the same operation $+$, we see that 1 and $n-1$ are both elements of H but $1 + (n-1) = n$ which is not an element of H. So H cannot be a subgroup. This brings us to a

WARNING: $(\mathbb{Z}_n, +)$ is NOT a subgroup of $(\mathbb{Z}, +)$!

The two groups have different $+$ operations. As in the last lecture, we could write $+$ and $+_n$ to make this clear.

Something that is a subgroup of $(\mathbb{Z}, +)$ is the subset of even numbers with the usual addition, as can be seen by checking the three conditions above: 0 is even, the sum of two even numbers is even and the inverse of an even number is even. the same holds for multiples of any number: for $n \in \mathbb{N}$, the subset

$$n\mathbb{Z} := \{n \cdot z \mid z \in \mathbb{Z}\}$$

forms a subgroup of $\mathbb{Z}$ with addition. Note that $\mathbb{Z} = 1\mathbb{Z}$ and $0\mathbb{Z} = \{0\}$ which is a subgroup with one element. More generally, every group $(G, +)$ has two trivial subgroups:

> **Definition 2.3.** The two trivial subgroups of a group $(G, +)$ are $(G, +)$ itself and $(\{e\}, +)$ where e is the neutral element. All other subgroups are called non-trivial subgroups.

2.3 Generators

How do we find subgroups of a group? The example with the even numbers $2\mathbb{Z}$ in $\mathbb{Z}$ suggests that we can try the following: for a given group $\mathbb{G} = (G, +)$, pick any element g and look at the "multiples" $g, g + g, g + g + g, \ldots$. Actually we have to add the neutral element e and the inverses $(-g), (-g) + (-g), \ldots$ too if we want to end up with a group.

> **Definition 2.4 (subgroup generated by element).** Let $(G, +)$ be a group and g any element of G. The subgroup generated by g, written $\langle g \rangle$, is the group containing exactly the following elements:
>
> - g is an element of $\langle g \rangle$.
> - The neutral element e is an element of $\langle g \rangle$.
> - If a, b are elements of $\langle g \rangle$ then so is $a + b$.
> - If a is an element of $\langle g \rangle$ then so is $(-a)$.

The way we wrote this definition automatically makes it into a subgroup. As an example, for any integer z we have $\langle z \rangle = z\mathbb{Z}$ in the group $\mathbb{Z}$, i.e. $\langle 2 \rangle$ really is the subgroup containing exactly the multiples of 2. In any group, the neutral element generates a subgroup with only one element — itself: $\langle e \rangle = (\{e\}, +)$.

In $\mathbb{Z}$, there happens to be an element 1 with the property that $\langle 1 \rangle = \mathbb{Z}$, i.e. 1 generates the whole group. We call such elements, if they exist, generators:

> **Definition 2.5 (generator, cyclic group).** Elements g of G in a group $(G, +)$ with $\langle g \rangle = G$ are called generators. If a group has a generator, it is called a cyclic group.

Not all groups have generators but we will have to wait a bit to see an example of a group that does not have one.

Exercise. $(\star)$ $\mathbb{Z}$ has one other generator besides 1; which?

The definition of generating a subgroup by one element generalises to several elements.

> **Definition 2.6 (subgroup generated by multiple elements).** Let $(G, +)$ be a group and $g_1, \dots g_n$ be elements of G. The subgroup $\langle g_1, \dots, g_n \rangle$ is the subgroup containing exactly the following elements:
>
> - $g_1, \dots, g_n$ are all elements of $\langle g_1, \dots, g_n \rangle$.
> - The neutral element e is an element of $\langle g_1, \dots, g_n \rangle$.
> - If a, b are two elements of $\langle g_1, \dots, g_n \rangle$ then so is their sum $a + b$.
> - If a is an element of $\langle g_1, \dots, g_n \rangle$ then so is $(-a)$.
>
> If $\langle g_1, \dots, g_n \rangle = G$ then we say that G is generated by the elements $g_1, \dots, g_n$.

The group $\langle g_1, \dots, g_n \rangle$ is the smallest subgroup of G that contains the elements $g_1, \dots, g_n$. Every group is generated by some set of elements: one can always take all of them to get $\langle G \rangle = G$. Whether an infinite group is generated by a finite subset of its elements is a more interesting question, but one that we will not investigate any further here.

2.4 Order of elements and Lagrange's theorem

We can define the order of an element by looking at the subgroup that it generates:

> **Definition 2.7 (order of an element).** The order of an element g in a group $\mathbb{G} = (G, +)$ is the order of the subgroup $\langle g \rangle$.

We can predict what the orders of elements in a finite group can be. In fact, if we know that any $\mathbb{H}$ is a subgroup of $\mathbb{G}$ then we know a lot about what $\mathbb{H}$ can look like. One of the most important facts if $\mathbb{G}$ is finite is Lagrange's theorem:

> **Theorem 2.8 (Lagrange).** If $\mathbb{G} = (G, +)$ is a finite group then the order of any subgroup of $\mathbb{G}$ divides the order of $\mathbb{G}$.

In our opening example, we found a subgroup H of 4 elements in $(Z_{32}, +)$ and indeed, 4 divides 32. Actually, H was the subgroup $\langle 8 \rangle$. We will not prove Lagrange's theorem here but a proof can be found at the end of these notes.

If $\mathbb{G}$ is finite, Lagrange's theorem tells us that the order of each element must divide the order of $\mathbb{G}$. For example in $(\mathbb{Z}_6, +)$, the only possible orders of elements are $1, 2, 3$ and 6. Elements of order 6, the same as the order of the whole group, are exactly the generators.

Exercise. *Small orders and small groups.*

1. ($\star$) Find the orders of all elements in the groups $(\mathbb{Z}_4, +_4)$, $(\mathbb{Z}_5, +_5)$ and $(\mathbb{Z}_6, +_6)$.
2. ($\star\star$) In any group, the only element of order 1 is the neutral element. Why?
3. ($\star\star$) A prime number p is a natural number greater than 1 whose only divisors are 1 and p. What can you say about the order of elements in $(\mathbb{Z}_p, +_p)$ when p is prime? (Look at $(\mathbb{Z}_5, +_5)$ again for a hint; compute all orders in $(\mathbb{Z}_7, +_7)$ if you need an extra hint since both 5 and 7 are prime.)

Exercise. ($\star\star$) *The L block.* An L block consists of four squares arranged in the shape of the letter capital L. Suppose we have an L block and two operations A and B that rotate it by 90 degrees clockwise and anticlockwise respectively as in Figure 1.

We can describe these actions on the L block by a group. Its elements are the possible rotations of the L block starting in the upright position. We can describe each operation by a (possibly empty) sequence using the letters A and B as shown in the diagram.

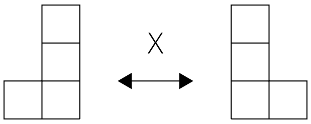
Two sequences of letters may describe the same operation, such as $ABABA$ and BBB. So let's introduce an equivalence relation: two sequences are equivalent if they produce the same rotation of the L block.

1. Describe a sensible set of representatives for these equivalence classes (hint: take the alphabetically first, shortest sequence from each class).

2. Give the rules to compute the representative element $[s]$ given a sequence s. (Hint: the rules will say that you can cancel certain subsequences.) Find the representatives of AAA, $ABABA$, $AAAA$ and $BBBBA$ using these rules.

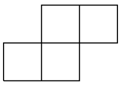
3. Label the representative elements with the numbers $0, 1, 2, 3$ in such a way that composing sequences matches addition in $\mathbb{Z}_4$. For example, if some sequence s gets the label 1 and t gets 2, then representative of the composition st should get the label $3 = 1 + 2$.

4. Suppose we added an extra move X that "mirrors" the L block horizontally:

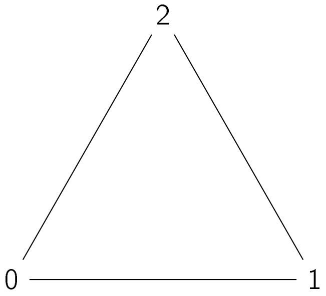


Redo the construction of a group of operations from before. How many elements do you end up with? Find sensible representatives. Why can you not label these elements with numbers $0, 1, \ldots, n-1$ anymore in such a way that addition of labels in $(\mathbb{Z}_n, +)$ matches composing transformations?

5. If you redo the same construction (with the X operation) for the S block (below), you get a group with four elements. Find a way to represent elements of this group as 2-bit numbers such that the group operation becomes the exclusive or (XOR) operation.



Note: in addition to L and S blocks, the same construction can be done with any regular n-gon where the vertices are labelled with the numbers $0, 1, \ldots, n-1$. Here is the regular 3-gon, a.k.a. triangle:



The group of rotations of such a n-gon around its centre, where two rotations are equivalent if they place all labels in the same place, is called the n-th cyclic group C_n and is another way or writing $(\mathbb{Z}_n, +)$ (for example by tracking the position of the 0 label). If we add the "mirroring" operation, we get a non-Abelian group of $2n$ elements called D_n. C_n is a subgroup of D_n and both are subgroups of S_n, the group of all permutations of the labels 0 through $n - 1$.

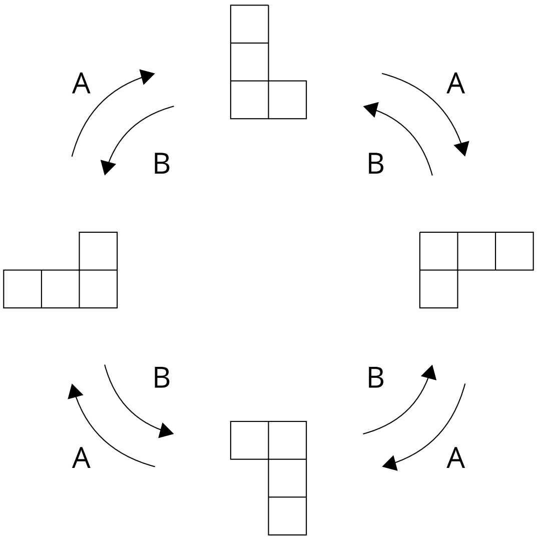


Figure 1: The L block (see exercises).

Exercise. *Free groups.* Here is another way to construct groups, similar to the idea with the L block and sequences of operations above. Pick any finite alphabet $\{A, B, C, \ldots\}$ of letters. For each letter, add an anti-letter which we denote by $\{\bar{A}, \bar{B}, \bar{C}, \ldots\}$ (none of the original letters is allowed to be the same as any of the anti-letters).

Our group elements are sequences (including the empty one) of letters and anti-letters where no letter is allowed to stand next to its anti-letter. The group operation $s + t$ on two sequences s, t is done by writing the sequence s followed by t, then repeatedly cancelling any pairs of a letter with its anti-letter. For n letters, this is called the free group on n generators.

1. $(\star)$ What is the neutral element? How do you compute the inverse of a sequence?

2. ($\star\star$) Convince yourself informally that this is a group (a full proof is not required!)
3. ($\star$) Convince yourself that the set of original letters really is a set of generators for this group.
4. ($\star$) For the free group on two generators A, B, find two elements (words) v, w such that $v + w \neq w + v$.
5. ($\star\star$) The free group on one generator is just another way of describing $(\mathbb{Z}, +)$. Explain how you could label sequences with integers such that the sum of two sequences is labelled by the sum of their labels.

Exercise. ($\star\star\star$) *Free groups via equivalence relations.* This is the "official" way to define free groups (and prove that they are groups):

- Take the set of all words containing the letters or their anti-letters (including the empty word) with no restrictions.
- Define an equivalence relation $\sim$ under which two words are equivalent if one can be obtained from the other by adding or removing pairs of a letter followed by its anti-letter (or the other way round).
- Take the set of equivalence classes of all words under $\sim$. Each class contains exactly one word that has no pairs of a letter and its anti-letter; take this word as the representative of the class.

Check that the relation $\sim$ described here really is an equivalence relation and that every equivalence class contains a word that has no pairs of a letter next to its anti-letter. (You don't need to prove that each class contains only one such word.)

The following sections are not presented in the lectures but are intended for self-study.

2.5 Permutation groups

Besides groups derived from $\mathbb{Z}$ there are many other ways to construct groups. Permutations are another example of a structure that produces groups.

Definition 2.9 (permutation). A permutation on a finite set S of elements is a bijective map p from S to itself. If the elements of S are $s_1, \ldots, s_n$ then a permutation p can be written in the form

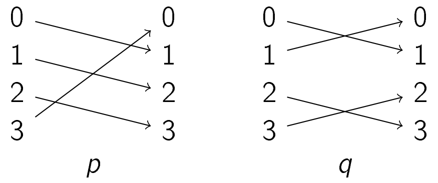
$$\begin{pmatrix} s_1 & s_2 & \ldots & s_n \\ p(s_1) & p(s_2) & \ldots & p(s_n) \end{pmatrix}$$

In the rest of this lecture we will use $\{0, 1, \ldots, n-1\}$ as an example set of size n. Indeed, to define permutations it makes no difference which set of a given size we pick — the elements are just "labels" for the permutations to work on.

We can also draw permutations as diagrams. For example, here are two permutations on the set $S = \{0, 1, 2, 3\}$:

$$p = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 2 & 3 & 0 \end{pmatrix} \quad q = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 3 & 2 \end{pmatrix}$$

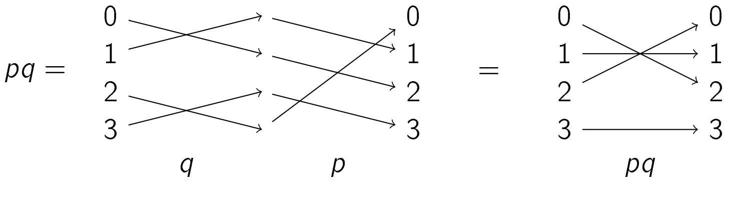
As diagrams, the permutations look like this:



What makes permutations into a group is the fact that they can be composed. This is easiest to see if we take two permutation diagrams and glue the arrows together in the middle. Unfortunately, there are two different ways we can do this and while our way is perhaps the more common, both ways can be found in different textbooks. For us, composing p and q means that we write the diagram of q on the left and that of p on the right, then glue the arrows together.

We also need a name for this operation. We choose the more standard pq (rather than $p + q$). The choice of symbol to mean "compose", unlike the order in which we compose things, is just a matter of notation and does not change what we are doing.

At the moment it might look "backwards" to use pq to mean the object you get when you compose with q on the left. We will explain this in a moment but first let's look at an example with the permutations p and q from above. We compose two permutation diagrams by glueing the arrows together in the middle:



This gives us the permutation

$$pq = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 3 \end{pmatrix}$$

WARNING. The order in which one writes a composed permutation varies from textbook to textbook. We call this object pq. Some authors would call it qp instead. As we will see in a moment, this is important: $pq \neq qp$.

The reason that we take pq to mean "q glued to p" and not the other way round is that permutations are formally functions on a set of elements: for the permutations p and q in the example, we could try and evaluate $p(q(0))$. Since $q(0) = 1$, this becomes $p(1)$ which is 2. In other words, to calculate $p(q(0))$ we apply q first, then p. This is actually the formal definition of our way of composing permutations:

> **Definition 2.10 (composition of permutations).** For two permutations p and q on the same set X, their composition pq is the permutation such that for all $x \in X$ we have $pq(x) = p(q(x))$.

Whichever way round we define composition, we get a group:

> **Definition 2.11 (symmetric group).** For any positive integer n, the symmetric group S_n is the group whose elements are the permutations of the set $\{0, 1, \ldots, n-1\}$ and the group operation is composition.

Let's check that this is actually a group (this is not a full proof). For associativity, look at the diagrams of any three permutations: it does not matter which two you glue together first as long as you keep the three diagrams in the same order. The neutral element is the permutation that maps every element to itself, sometimes written id (for "identity"). The inverse of a permutation can be found by reversing all arrows in the diagram and taking the mirror image.

Of course, one can define the group of permutations over any set S. But the structure of this group depends only on the number of elements that S has, so for finite sets S one can consider only the sets $\{0, 1, \ldots, n-1\}$ "without loss of generality". What we are actually doing is taking this set as representative element of the "class of all sets with n elements".

Permutation groups are an example of non-commutative groups. (Easy exercise: find qp for the p, q given above and check that $pq \neq qp$.)

2.6 Subgroups of permutation groups

Take the group S_n of permutations on any set A of n elements (the set $\{0, 1, \ldots, n-1\}$ will do fine). Pick any subset B of A and consider only the permutations S' that leave all elements outside B alone. For example, if $A = \{0, 1, 2, 3\}$ and $B = \{0, 1, 2\}$ then the permutation

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 2 & 1 & 3 \end{pmatrix}$$

leaves the elements outside B alone — the only such element is 3. (It is irrelevant for S' whether or not the permutation leaves anything inside B alone.) It turns out that S' is a subgroup of S_n. The neutral element id of S_n leaves every element alone so it certainly leaves those outside B alone; if two permutations both leave some element alone then so does their composition — as can be seen on their diagrams, the composition of two horizontal lines is still a horisontal line. The same argument on diagrams shows that if some permutation p leaves an element x alone then so does its inverse: reversing a horizontal arrow gives another horizontal arrow. We have informally shown the following:

> **Proposition 2.12.** For any positive integers $m \leq n$, the group S_m is a subgroup of S_n.

2.7 Cycles

Another way to look at permutations is to consider cycles.

> **Definition 2.13.** On a set A, a cycle of length k is a list of elements of A with no repetitions $(a_1, a_2, \ldots, a_k)$. The empty cycle (of length 0) is written $()$.

A cycle defines a permutation by sending a_1 to a_2 etc., and a_k back to a_1. For example, on the set $A = \{0, 1, 2, 3, 4, 5, 6, 7\}$ the cycle $(5, 1, 2)$ defines the permutation

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 2 & 5 & 3 & 4 & 1 & 6 & 7 \end{pmatrix}$$

The empty cycle () gives the identity permutation id. For any cycle or set of cycles c, we can look at the group $\langle c \rangle$ that they generate. This is by definition a subgroup of S_n.

The notation for cycles does not uniquely describe a permuation: $(5, 1, 2)$ and $(1, 2, 5)$ both give the same permutation. (Exercise: there is one other way to write the same permutation as a cycle, which one?) However $(5, 2, 1)$ describes a different permutation.

The interesting thing about cycles is that they can be used to build all permutations.

Proposition 2.14. Any permutation $p \in S_n$ can be written as a composition of cycles with disjoint elements. If we take the underlying set to be $\{0, \ldots, n-1\}$ then there is an ordering under which each permutation has exactly one normal form as a composition of disjoint cycles:

- Shorter cycles come before longer ones.
- For two cycles of the same length, the one with the smallest element comes first.
- Within each cycle, the smallest element is the first.

There are several useful rules for computing with cycles which we will use to prove this proposition. The composition of two cycles is the composition of their permutations (in the same order as the cycles are given), so the expression $(0,1)(2,3)$ on the set $\{0,1,2,3\}$ is the composition

$$(0,1)(2,3) = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 2 & 3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 3 & 2 \end{pmatrix}$$

1. If you rotate the elements in a cycle around by taking the last element and placing it as the new first element, you get the same permutation. Conversely, two cycles describe the same permutation if and only if one can be rotated to give the other.

   For example, $(1,2,3)$ and $(3,1,2)$ both describe the same permutation.

2. If c and d are two disjoint cycles (they do not share any elements) then $cd = dc$.

   In the example above, one can check that $(1,0)(2,3) = (2,3)(1,0)$.

3. If two cycles c, d share exactly one element x, their composition is the cycle obtained by rotating x to be in the last place of c and the first place of d, and "glueing together" the x-es, dropping the middle parentheses.

   For example, to compute $(1,2,3)(2,4)$ we write this as $(3,1,2)(2,4)$ and glue the 2-s to get $(3,1,2,4)$.

Rules 1 and 2 together say that if we have any way of representing a permutation as a composition of disjoint cycles then we can reorder the cycles and the elements within cycles to get the unique representation from proposition 2.14.

As an example, consider the set $A = \mathbb{Z}_8$ and the permutation

$$p = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 2 & 4 & 0 & 1 & 6 & 5 & 7 \end{pmatrix}$$

Starting with 0, we find $p(0) = 3$ and $p(3) = 0$ so we have our first cycle $(0,3)$. As we move through the elements we get the other cycles $(1,2,4)$ and $(5,6)$, which we can order to write

$$p = (0,3)(5,6)(1,2,4)$$

Exercise. *Permutations — the case $n = 3$.*

1. ($\star$) Write out all permutations on the set of three elements $\{0, 1, 2\}$.
2. ($\star$) Find two elements p, q of S_3 such that $pq \neq qp$.
3. ($\star$) Decompose all the elements of S_3 into cycles.
4. ($\star\star$) Write out the "addition table" (composition table) for S_3, representing the elements as cycles.
5. ($\star\star$) You have just shown that S_3 is not commutative. Conclude that for any $n \geq 3$, S_n is not commutative either. What about S_1 and S_2?

Exercise. ($\star$) *Permutations in general.* Consider

$$p = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 5 & 4 \end{pmatrix} \quad q = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 3 & 5 \end{pmatrix}$$

- Compute pq and qp.
- What are the orders of p and q?
- Decompose p and q into cycles.
- Write the following as permutations: (154), $(12)(54)$ and $(1254)(4531)$.

2.8 $\diamond$ Cosets and a proof of Lagrange's theorem.

$\diamond$ For any group $\mathbb{G} = (G, +)$ with a subgroup $\mathbb{H} = (H, +)$ we can define an equivalence relation $\sim_H$ on G: let $g \sim_H k$ hold if and only if there is an element h of H such that $g + h = k$. $\sim_H$ is an equivalence relation because it is (i) reflexive: $g + e = g$ for any g and e is an element of H; (ii) symmetrical because if $g + h = k$ then $k + (-h) = g$ and $(-h)$ must also be in H; (iii) transitive because if $g + h = k$ and $k + h' = l$ for elements g, k, l of G and h, h' of H then $g + (h + h') = l$ and $h + h'$ must also be in H. In fact, the three conditions for H being a subgroup match exactly with the three conditions for $\sim_H$ being an equivalence relation, so we could define subgroups via equivalence relations if we wanted to.

In our opening example, the relation $\sim_H$ that we get this way is exactly the relation $\sim_8$ that we know from the last lecture; this is not a coincidence of course. What we are doing now is more general since we can build an equivalence relation out of a subgroup even if the elements of these groups are not "numbers".

Every element of G thus lands in exactly one equivalence class of $\sim_H$ and all of H lands in the same class. We claim that all equivalence classes have the same number of elements. This way, if we write $|\mathbb{G}|$ for the number of elements of G and $|\mathbb{H}|$ for the number of elements in H then $|\mathbb{G}| = |\mathbb{H}| \cdot c$ where c is the number of classes that $\sim_H$ creates, proving the theorem. This number of classes is important enough that it gets its own name:

Definition 2.15. The index of $\mathbb{H}$ in $\mathbb{G}$, written $[\mathbb{G} : \mathbb{H}]$, is the number of classes that the relation $\sim_H$ creates in G.

Lagrange's theorem in a more general form actually says:

$$|\mathbb{G}| = |\mathbb{H}| \cdot [\mathbb{G} : \mathbb{H}]$$

Cosets. Another way to define these equivalence classes is via left cosets. We will use cosets to prove Lagrange's theorem.

Definition 2.16 (left coset). Let $\mathbb{H} = (H, +)$ be a subgroup of $\mathbb{G} = (G, +)$. The left coset of H at an element $g \in G$ is the set

$$g + H := \{g + h \mid h \in H\}$$

It is not hard to see that for any $g \in G$, the coset $g + H$ is exactly the equivalence class $c(g)$ under $\sim_H$. (Left as an exercise for the mathematically proficient.) What we are trying to prove is that all cosets have the same size and we can do this by giving a bijection between them: let g and k be any two elements of G. Then the function $f : x \mapsto k + (-g) + x$ is a bijective map from $g + H$ to $k + H$ with inverse $f^{-1} : x \mapsto g + (-k) + x$. To see this, pick an element of $g + H$ and write it as $g + h$ for $h \in H$ as per its definition. Then $f(g + h) = k + (-g) + g + h = k + h$ is an element of $k + H$, showing that the function maps $g + H$ into $k + H$. Further for any element x we have $f^{-1}(f(x)) = g + (-k) + k + (-g) + x = x$ and the same holds for $f(f^{-1}(x))$. So any two left cosets of H are in bijection and therefore of the same size. Since $H = e + H$ for the neutral element e, this proves Lagrange's theorem.

We close with a few comments. We could do the same argument with right cosets as well as left cosets; in an Abelian group the two are the same anyway.

Infinite groups. The whole construction of cosets and indices works equally well for infinite groups, for example for $n > 0$ we have $[\mathbb{Z} : n\mathbb{Z}] = n$. An index can also become infinite: $[\mathbb{Z} : 0\mathbb{Z}] = \infty$ since the resulting equivalence relation is just the = (equals) relation so every element is its own equivalence class.

A version of Lagrange's theorem holds for infinite groups too but the reason that we did not give it here is that we would have first to define what "divides" means when infinities are involved, which is beyond what we want to cover in this course.

Lecture 3 — Rings and Multiplication

Dr. D. Bernhard

In this lecture: rings — units and zero divisors — the group $\mathbb{Z}_n^\times$ — Euler's totient function — Euclid's algorithm — fields — exponent arithmetic

Learning outcomes. After this lecture and revision, you should be able to:

- Find the order of the multiplicative group modulo any n.
- Apply Euclid's algorithm to integers m, n to find a, b with $a \cdot n + b \cdot m = \texttt{gcd}(m, n)$.
- Multiply and invert in $\mathbb{Z}_n^\times$.
- Tell if a structure is a ring.
- Classify ring elements as units, zero divisors or neither.
- Solve equations of the form $a \cdot x = b \pmod{n}$.
- Perform exponent arithmetic.

3 Rings and multiplication

The last two lectures, we have looked at groups which are a generalisation of addition. Today we look at multiplication.

3.1 Rings

We began this course by saying that a group is a bit like adding numbers. This is not quite true — a group is a bit like adding and subtracting numbers, since there must always be inverses. If we introduce multiplication as a group operation, this gives us division too but sometimes that is more than we need. In $\mathbb{Z}$ for example, you can multiply but you cannot always divide. A ring is a structure a bit like numbers with addition, subtraction and multiplication but you cannot necessarily divide (even when there is no 0 involved).

> **Definition 3.1.** A structure $\mathcal{R} = (R, +, \cdot)$ where R is a set and $+, \cdot$ are two operations $R \times R \to R$ is a ring if the following hold.

additive group The structure $(R, +)$ is an Abelian group. We call its neutral element the zero of R and write it with the symbol 0 or sometimes 0_R.

multiplication The structure $(R, \cdot)$ is associative and has a neutral element, which we call the one of the ring and write as 1 or 1_R.

distributive law For any elements a, b, c in R we have $(a + b) \cdot c = a \cdot c + b \cdot c$ and $c \cdot (a + b) = c \cdot a + c \cdot b$.

The description of a ring requires the additive group to be commutative but not ncessarily the multiplication. We call a ring commutative if the multiplication is commutative too.

The standard example of a ring is $(\mathbb{Z}, +, \cdot)$. For any positive integer n, the space $(\mathbb{Z}_n, +_n, \cdot_n)$ is a ring too. If we take n-tuples of integers with componentwise addition and multiplication, this gives us yet another ring.

We state some basic facts about rings:

- There can only be one neutral element for multiplication.
- There is a ring with one element, which is the neutral element for both addition and multiplication ($0 = 1$!). As soon as a ring has at least two elements, the neutral elements 0 for addition and 1 for multiplication must be different however.
- Multiplying any element x in a ring with 0, the neutral element of addition, gives 0 (from both the left and the right, if the ring is not commutative).

3.2 Units and zero divisors

In a ring, things can happen that we might not expect from the usual integers. For example, the structure $(\mathbb{Z}_8, +_8, \cdot_8)$ is a ring in which $2 \cdot_8 4 = 0$, so a product of two nonzero elements can be zero even if neither of the factors are. If we take the ring of pairs of integers with component-wise addition and multiplication, $(0, 1) \cdot (1, 0) = (0, 0)$: another case where the product of two nonzero elements becomes zero.

We can classify all elements in a ring into the following classes.

Definition 3.2 (unit, zero divisor). In a ring $(R, +, \cdot)$:

- The neutral element of addition is called the zero of the ring and written with the symbol 0.
- An element a in the ring which has a multiplicative inverse b, i.e. $a \cdot b = 1$ and $b \cdot a = 1$, is called a unit of the ring. The neutral element of multiplication is always a unit.

- An element $a \neq 0$ for which there is some other $b \neq 0$ such that $a \cdot b = 0$ or $b \cdot a = 0$ is called a zero divisor.
- An element can also be neither of the above.

Most of the time, the classification zero/unit/zero divisor/none of the previous is unique, i.e. each element of the ring is in exactly one class. The only exception is that in the ring with one element, the element is both the zero and a unit; as soon as $1 \neq 0$ a unit can neither be zero nor a zero divisor.

3.3 The group $\mathbb{Z}_n^\times$

Let's try and build a structure where multiplication is a group operation. Start by looking at multiplication in $\mathbb{Z}$. Multiplication is associative and 1 is the neutral element. However, 0 causes problems if we try and find inverses: $0 \cdot x = 0$ for all x so there cannot be an x with $0 \cdot x = 1$. So let's get rid of 0 and look at $\mathbb{Z}^\times = \mathbb{Z} \setminus \{0\}$. We still don't have inverses: there is no integer z satisfying $3 \cdot z = 1$ for example. The usual way to carry on at this point is to introduce the fractions (without zero) $\mathbb{Q}^\times$ which form a group under multiplication.

Now let's look at the group $(\mathbb{Z}_{2^8}, +) = (\mathbb{Z}_{256}, +)$; this group is sometimes called `byte`, `uint8` or `unsigned char` in programming. What happens if we introduce multiplication on this group with the rule that if you exceed 255, you subtract 256 until you are back in range (equivalently, you ignore all "higher bits" of a number)?

Interestingly, we can solve the equation $3 \cdot_{256} z = 1$ where $\cdot_{256}$ is multiplication modulo 256. The solution is $z = 171$ since $3 \cdot 171 = 513 = 2 \cdot 256 + 1$ (division with remainder) so $3 \cdot_{256} 171 = 1$. So in some cases we can find inverses without resorting to fractions.

We quickly hit another problem though: $16 \cdot 16 = 256$, so $16 \cdot_{256} 16 = 0$ so 16 cannot have an inverse, fractions or no: if $16 \cdot_{256} x = 1$ had any solution for x, we could multiply both sides of the equation to get $0 = 16$ which is nonsense, even modulo 256.

Let's throw out all elements that cause problems from $\mathbb{Z}_n$. 0 is right out. Anything that gives 0 when multiplied with another number from $\mathbb{Z}_n$ except 0 is out too (that is, all zero divisors). This leaves us with a group:

Proposition 3.3 (the group $\mathbb{Z}_n^\times$). For a positive integer n, the set $\mathbb{Z}_n^\times$ is the subset of $\mathbb{Z}_n$ containing all elements x for which $x \cdot_n y \neq 0$ for all other elements $y \neq 0 \in \mathbb{Z}_n$. Together with multiplication modulo n, this forms a group $(\mathbb{Z}_n^\times, \cdot_n)$.

Let's check that this really is a group. Associativity follows from that of $\cdot$ on $\mathbb{Z}$, the same way we did it for $+_n$ in the first lecture. The neutral element is 1 since $1 \cdot_n y = y$

for any $y \neq 0 \in \mathbb{Z}_n$ and if $y \neq 0$ then $1 \cdot_n y = y$ is not 0 either, so we can't lose the element 1 in the definition of $\mathbb{Z}_n^\times$. To find inverses we have to invoke a bit more number theory.

Exercise. ($\star$) *Arithmetic modulo primes.* In this exercise, we choose the prime $p = 1009$ as our modulus and look at the ring $\mathcal{R} = (\mathbb{Z}_{1009}, +, \cdot)$.

1. Compute $824 + 3 \cdot 632$ in $\mathcal{R}$.
2. What is the order of 7 in $(\mathbb{Z}_{1009}, +)$?

Exercise. ($\star$) *Addition and multiplication tables.*

- Compute the addition and multiplication tables for $(\mathbb{Z}_5, +, \cdot)$.
- Do the same for $(\mathbb{Z}_4, +, \cdot)$.
- Write the group table for the group $(\mathbb{Z}_4^\times, \cdot)$.

Exercise. ($\star\star$) *Arithmetic modulo n.* Consider $n = 16$. Find all solutions modulo n of the equations

1. $5x = 1$
2. $6y = 1$
3. $8z = 0$

3.4 Euler's totient function

How many elements are left over in $\mathbb{Z}_n^\times$? We give this quantity a name.

Definition 3.4 (Euler totient function). Euler's totient function ϕ is the function $\phi(n) := |\mathbb{Z}_n^\times|$ for $n > 0 \in \mathbb{N}$ that maps a positive integer n to the number of elements in $\mathbb{Z}_n^\times$.

To compute this function, we can ask the related question, which elements have we got rid of? Let's take an element $x \neq 0$ that was in $\mathbb{Z}_n$ but not in $\mathbb{Z}_n^\times$, which means there is some $y \neq 0$ in $\mathbb{Z}_n$ such that $x \cdot_n y = 0$. The definition of $x \cdot_n y$ is the remainder when dividing $x \cdot y$ by n, so we must have $x \cdot y = c \cdot n + 0$, i.e. the remainder is 0 so $x \cdot y$ is a multiple of n. Since we assumed $x \neq 0$ and $y \neq 0$ then c cannot be 0 either: the product of two nonzero elements in $\mathbb{Z}$ is never zero. Further if y is not a multiple of n on its own, which it cannot be because $y \in \mathbb{Z}_n$ so $y < n$, then x and n must have a factor in common (that is higher than 1). Conversely, if x and n have the common factor k then we can write $x = ak$, $n = bk$ for some nonzero integers a, b and $b \cdot x = a \cdot b \cdot k = a \cdot n$ which is a multiple of n so $x \cdot_n b = 0$. So a number in $\mathbb{Z}_n$ is excluded from $\mathbb{Z}_n^\times$ if and only if it has a common factor (other than 1) with n.

This tells us $\phi(p)$ whenever p is a prime. Since primes are exactly those numbers that have no factors except 1 and themselves, the only element of $\mathbb{Z}_p$ that is excluded from $\mathbb{Z}_p^\times$ is 0 and we get $\phi(p) = p - 1$ and can say that $\mathbb{Z}_p^\times = \{1, 2, \dots, p - 1\}$.

If m and n are two positive integers that have no factors other than 1 in common (another way of saying this is that m, n are coprime) then the only elements of $\mathbb{Z}_{m \cdot n}$ that have a factor in common are those that already had a factor in common with m or with n. In other words,

> **Proposition 3.5.** Two integers m, n are coprime if their greatest common divisor is 1. For coprime positive integers m, n we have $\phi(m \cdot n) = \phi(m) \cdot \phi(n)$.

In particular this holds when m, n are distinct primes. For primes, we can say even more: if we look at the numbers sharing a factor with p^k for p a prime and k a positive integer, these numbers must already share a factor with p since the factors of p^k are exactly $1, p, p^2, \dots, p^k$. If we write the numbers from 1 to p^k in a p-column table, the table has $p^k/p = p^{k-1}$ rows and each row contains exactly one multiple of p in the last column, so there are $(p-1)$ columns with numbers that are coprime to p per row. This lets us find $\phi(p^k)$:

> **Proposition 3.6.** For a prime p and a positive integer k we have $\phi(p^k) = p^{k-1}(p-1)$.

The last two propositions allow us to find $\phi(n)$ for any n. We factor n as $p_1^{a_1} \cdot \dots p_k^{a_k}$ where the $p_1, \dots, p_k$ are distinct primes and a_i is the number of times the prime p_i appears. Then we have

$$\phi(n) = \prod_{i=1}^{k} p_i^{a_i - 1}(p_i - 1)$$

This concludes our little excursion to investigate Euler's ϕ function and we return to showing that $(\mathbb{Z}_n^\times, \cdot_n)$, whose size we can now compute, is a group.

Exercise. ($\star$) *Euler's ϕ function.* How many elements are there in the group $(\mathbb{Z}_n^\times, \cdot)$ for

1. $n = 1009$ (this is a prime)
2. $n = 64$
3. $n = 60$

3.5 Euclid's algorithm

Our next step is a lemma by Euclid, a variation on division with remainder, that comes with an algorithm to find the numbers in question:

> **Lemma 3.7 (Euclidean algorithm).** If m, n are two nonzero integers then there are unique integers a, b such that
>
> $$a \cdot n + b \cdot m = \mathtt{gcd}(m, n)$$

This gives us the inverses we want in $\mathbb{Z}_n^\times$. If m is coprime to n then the $\mathtt{gcd}$ is 1 (or m would not be in $\mathbb{Z}_n^\times$ at all) and we find a, b such that $a \cdot n + b \cdot m = 1$ which is equivalent to saying that $b \cdot m = (-a) \cdot n + 1$ so $b \cdot m$ leaves remainder 1 when dividing by n and $b \pmod{n}$ is the inverse we are looking for. (We take b modulo n again because Euclid's algorithm can return an b outside the range $\mathbb{Z}_n$ in some cases.) This concludes the proof that $(\mathbb{Z}_n^\times, \cdot_n)$ is a group.

Here's one way to compute the extended Euclidean algorithm. Suppose we want to invert 17 modulo 256. That is, we want to find b such that $17b = 1 \pmod{256}$, for which we find a, b with Euclid's algorithm such that $256a + 17b = 1$. Make a four-column table with headings **q** (quotient), **r** (remainder), **a** and **b**. Write the top two rows as in the table below with the numbers in the **r** column and the rest as shown.

q	**r**	**a**	**b**
0	256	1	0
0	17	0	1
15	1	1	-15
17	0	-17	256

For all further rows, calculate as follows. Let q', r', a', b' be the values from the last row and q'', r'', a'', b'' be the second-to-last row (so for the third row, $r' = 17$ and $r'' = 256$).

- Divide r'' by r' with remainder. Put the quotient and remainder as the new q, r values.

- Set $a := a'' - q \cdot a'$ and $b := b'' - q \cdot b'$.

For row 3, we get $256 = 15 \cdot 17 + 1$ so we start the third row with $q = 15, r = 1$. Then $a := 1 - 0 \cdot 15 = 1$ and $b := 0 - 1 \cdot 15 = -15$.

Repeat until the remainder becomes $r = 0$. When this happens, the last row with a nonzero remainder contains the important information: the r value in this row is the $\texttt{gcd}$ of the two original numbers (in our case this is 1, without this we could not do modular inversion at all) and the a, b values in this row are the ones we are looking for. In our case $a = 1$ and $b = -15$ and indeed, $1 \cdot 256 - 15 \cdot 17 = 1$ so $-15 \mod 256 = 241$ is the inverse of 17 in $(\mathbb{Z}_{256}^{\times}, \cdot)$.

Exercise. ($\star\star$) *Euclid's algorithm.*

1. Find integers a, b such that $13a + 64b = 1$.
2. Find the inverse of 5 under multiplication modulo 1009.
3. Find all solutions of the equations $101 \cdot x = 1$ and $25 \cdot y + 6 = 98$ in $\mathbb{F} = (\mathbb{Z}_{1009}, +, \cdot)$.

Exercise. ($\star\star$) *Classification of ring elements.* Classify all elements of the rings $(\mathbb{Z}_n, +, \cdot)$ as zero, unit, zero divisor or neither for the following values of n.

1. $n = 1009$ (this is still a prime)
2. $n = 64$
3. $n = 60$
4. Do the same for $(\mathbb{Z}, +, \cdot)$.

3.6 Fields

A commutative ring in which every nonzero element is a unit is called a field. (Fields are therefore automatically integral domains.)

> **Definition 3.8 (field).** A structure $\mathbb{F} = (F, +, \cdot)$ is a field if it is a commutative ring and every element except the zero is a unit.

Another way of saying that $\mathbb{F}$ is a field is that $\mathbb{F} \setminus \{0\}$ forms a (commutative) group under multiplication — with the one exception of $(\{0\}, +, \cdot)$ which is also a field. We will return to fields and finite fields in particular in a later lecture. For now, we summarise the basic algebraic structures:

- A monoid[1] is a structure in which you can add.
- A group is a structure in which you can add and subtract.
- A ring is a structure in which you can add, subtract and multiply.
- A field is a structure in which you can add, subtract, multiply and divide[2].

Exercise. ($\star\star$) *Fields modulo primes.* We know that $p = 1009$ is a prime. This means $(\mathbb{Z}_{1009}, +, \cdot)$ is a field (and the same holds for every other prime). Why?

The following section is for self-study.

3.7 Exponent arithmetic

We have learnt to add and multiply modulo n. Writing $[\,]$ for reduction modulo n, we have $[a + b] = [[a] + [b]]$ and $[a \cdot b] = [[a] \cdot [b]]$. What about exponentiation? Since it is usually defined by repeated multiplication

$$a^n := \underbrace{a \cdot a \cdot \ldots \cdot a}_{n \text{ times}} \qquad (n \in \mathbb{N})$$

we obviously get $[a^n] = [[a]^n]$ and the usual rules such as $a^n \cdot b^n = (a \cdot b)^n, (a^n)^m = a^{n \cdot m}$ in any commutative ring. Note however that while a, b can be elements of any commutative ring, m, n are integers — exponentiation with exponents in any ring is not defined! Thus, in the last formula, the multiplication in the exponent is normal integer multiplication. In a field, we can define exponentiation with negative exponents the usual way $a^{-n} := (1/a)^n$ for any base except 0 (we leave 0^0 undefined).

In some cases, we can reduce exponents as well to calculate more efficiently. Let's try and calculate $3^{10} \pmod 5$. Since we expect a result in $\mathbb{Z}_5$, surely we can reduce that 10? The wrong way to do this is to note that $[10] = 0$ modulo 5, so $3^0 = 1$. Wrong, because $3^{10} = 59049 = 11809 \cdot 5 + 4$ so the correct answer is 4, not 1. What went wrong? The problem is that exponents, as we mentioned, are integers and not ring elements — even if ring elements are sometimes integers too. We summarise:

WARNING: you can reduce $\pmod n$ at any time in addition, subtraction, multiplication and (where defined) division in $\mathbb{Z}_n$. You cannot reduce exponents this way.

There is a way to reduce the exponent correctly though. Recall that $(Z_n^\times, \cdot)$ is a group, so each element a of this group has an order k for which $a^k = 1$ in the group.

[1] We haven't formally covered monoids in this course but they are structures $(M, +)$ where $+$ is associative and has a neutral element, but not necessarily inverses. If $(R, +, \cdot)$ is a ring then $(R, \cdot)$ is a monoid.

[2] Except in the field with one element (the ring with one element is a field!), you cannot of course divide by zero.

And all these element orders must divide the group order by Lagrange's theorem; the group order is of course $\phi(n)$. So the correct way to reduce exponents is taking them modulo $\phi(n)$ instead of n. In our example $3^{10} \pmod 5$ we have $\phi(5) = 4$ so $3^{10} = 3^{10 \pmod{\phi(5)}} = 3^{10 \pmod 4} = 3^2 = 4 \pmod 5$. So the correct formula is

$$a^k = (a \mod n)^{(k \mod \phi(n))} \pmod n$$

using $[\;]_n$ to denote reduction modulo n we can also write this as

$$[a^k]_n = [\,([a]_n)^{[k]_{\phi(n)}}\,]_n$$

that is, you reduce group elements or bases modulo n and exponents modulo $\phi(n)$. In the special case where n is a prime, $\phi(n) = n - 1$.

For large values of n and its prime factors, computing $\phi(n)$ is much more time-consuming than simple computation modulo n: essentially, you have to perform a task similar to factoring n to get hold of $\phi(n)$. This forms the basis of the famous RSA cryptosystem and much of modern cryptography that has been developed since.

Exercise. $(\star\star)$ *Exponentiation modulo n.* Compute the following value:

$$2^{128} - 1 \pmod{1009}$$

Hint: you definitely do not want to compute all the digits of 2^{128}! You know that $[a + b] = [[a] + [b]]$ and $[a \cdot b] = [[a] \cdot [b]]$ where $[\;]$ is reduction modulo 1009. You can also avoid doing 128 individual calculations and do 8 instead.

3.8 ◇ Cancellation and integral domains

◇ In a group, you can "cancel" in additions: $x+a = y+a$ implies $x = y$. This holds because group elements have inverses: given $x + a = y + a$ you can add the inverse of a on both sides to get $(x + a) + (-a) = (y + a) + (-a)$, swap the brackets round with the associative law to get $x + (a + (-a)) = y + (a + (-a))$, use the property of inverses to get $x + 0 = y + 0$ and the property of the neutral element to get $x = y$.

If an element in a ring is not zero or a zero-divisor, you can still cancel it — but the reason is a slightly different one. Suppose a is such an element and $a \cdot x = a \cdot y$. This implies that $a \cdot (x - y) = 0$ (subtract $a \cdot y$ on both sides and use the distributive law) but since a is not a zero divisor, this means that $x - y = 0$ and therefore $x = y$.

A ring in which there are no zero divisors and you can cancel multiplication with anything except 0 is called an integral domain. The standard example of an integral domain is $\mathbb{Z}$: $3 \cdot x = 3 \cdot y$ implies $x = y$ even without you having to extend to $\mathbb{Q}$ in order to invert 3.

Definition 3.9 (integral domain). A ring $\mathcal{R}$ in which there are no zero divisors is called an integral domain.

Lecture 4 — Polynomials

Dr. D. Bernhard

In this lecture: definition of polynomials — degree — polynomials over rings — polynomial arithmetic — calculation modulo polynomials

Learning outcomes. After this lecture and revision, you should be able to:

- Convert polynomials between a representation as sequences and a representation using a formal variable X.
- Divide polynomials with remainder, over both $\mathbb{Q}$ and any finite field.
- Add and multiply polynomials over any field and in polynomial rings over fields.

4 Polynomials

If we have any group $\mathbb{G} = (G, +)$ we can make a new group $\mathbb{G}^n$ by taking tuples of length n of G-elements and adding them component-wise. We can do the same for rings but the result is not that interesting: any tuple with a 0 element anywhere is a zero divisor (except the all-zero tuple, which is the zero element). Polynomials are a much "richer" structure to look at tuple of ring elements.

4.1 Definition of polynomials

Let's look at polyomials over $\mathbb{Z}$ in one variable X to start with. A polynomial is a finite sequence of monomials, each of which is a coefficient multiplied with a power of X such as $p = 2X^2 + X - 1$ or $q = 2X$. We can add polynomials, this means we add the coefficients of the same powers so $p + q = 2X^2 + 3X - 1$. We can also multiply polynomials, the rule here is that you multiply every monomial of p with every one of q, multiplying coefficients and adding powers: $p \cdot q = 4X^3 + 2X^2 - 2X$.

Polynomials over $\mathbb{Z}$ are of course functions from $\mathbb{Z}$ to $\mathbb{Z}$, since you can stick an integer in the variable and get an integer out. In Algebra, we are more interested in polynomials as objects in their own right rather than their effects as functions. The variable X is not important: all the information about polynomials is contained in the coefficients. We can write a polynomial over the integers as a sequence of integers, starting with the coefficient for the power 0 and proceeding in order of ascending powers, so $p = (-1, 1, 2, 0, \ldots)$ and $q = (0, 2, 0, \ldots)$. This way, a polynomial is an infinite sequence of which at most finitely many elements are nonzero. To save ourselves from always writing out dots, we can break off as soon as no more nozero elements follow and write for example $q = (0, 2)$.

With this example in mind we give the definition of a polynomial ring over an arbitrary ring $(R, +, \cdot)$:

Definition 4.1 (polynomial ring). Let $\mathcal{R} = (R, +, \cdot)$ be any ring. The polynomial ring in one variable $\mathcal{R}[X]$ over $\mathcal{R}$ is the following ring:

- Elements are (countably) infinite sequences of R-elements, of which at most finitely many in a sequence are nonzero.
- Addition is element-wise, so $(a_0, a_1, \ldots) + (b_0, b_1, \ldots) = (a_0 + b_0, a_1 + b_1, \ldots)$.
- Multiplication is defined as follows. For two elements $a = (a_0, a_1, \ldots)$ and $b = (b_0, b_1, \ldots)$ the product is the element $c = (c_0, c_1, \ldots)$ with

$$c_j := \sum_{i \in \mathbb{N}} \sum_{j=0}^{i} a_i \cdot b_{j-i}$$

The formula for multiplication describes each coefficient of c in terms of addition and multiplication in the ring $\mathcal{R}$ and does exactly what the informal definition over "powers of X" says. Note that the element X in the name $\mathcal{R}[X]$ never appears later in the definition: it is just a label and it is beat not to think of it as a "variable" too much. Of course we can write polynomials using a formal variable X as well, if we want to.

The zero of $\mathcal{R}[X]$ is the sequence that is zero (the zero of $\mathbb{R}$) everywhere and the one is the sequence $(1, 0, 0, \ldots)$; written as a polynomial in a formal variable X this sequence is simply $1 + 0X + 0X^2 \ldots = 1$ (which is the one of the ring $\mathcal{R}$).

Exercise. ($\star$) *Polynomials modulo* 7. Compute the following in $\mathbb{Z}_7[X]$:

1. $(4X^3 + 5X + 2) + (6X^3 + 2X^2 + 3)$
2. $(X^2 + 6X + 4) \cdot (3X^3 + 2X^2 + 1)$

4.2 Degree

We should check that the product of two polynomials (as sequences) really is another polynomial, that is only finitely many elements end up non-zero. We introduce the degree:

> **Definition 4.2.** The degree of a non-zero polynomial is the index of the highest coefficient that is non-zero, starting the count at 0. The degree of p is denoted by $\mathtt{deg}(p)$.

So the degree of $p = 4X^2+2X+1 = (1, 2, 4)$ is 2 since we start counting at coefficient zero (which has the value 1 in p). The degree of the zero polynomial can either be left undefined or we can define it to be minus infinity.

> **Proposition 4.3.** For polynomials p, q we have $\mathtt{deg}(p+q) \leq \mathtt{max}(\mathtt{deg}(p), \mathtt{deg}(q))$ and $\mathtt{deg}(p \cdot q) \leq \mathtt{deg}(p) \cdot \mathtt{deg}(q)$.

This is almost the rule we are used to from polynomials over $\mathbb{Z}$ and it proves that the sum and product of polynomials is again a polynomial, since the resulting sequence has a finite degree. The reason for $\leq$ instead of $=$ is that zero divisors can cancel the highest coefficients. In $\mathbb{Z}_6[X]$, for polynomials $p = 2X = (0, 2)$ and $q = 3X^2 = (0, 0, 3)$ we have $pq = (0)$ since $2 \cdot 3 = 0 \pmod 6$. An even more "unusual" event happens in $\mathbb{Z}_4[X]$ for the polynomial $2X + 1 = (1, 2)$: we compute $(2X + 1) \cdot (2X + 1) = [4X^2 + 4X + 1] = 1$ (i.e. compute normally and reduce modulo 4) so this polynomial is a unit and also its own inverse.

4.3 Polynomial arithmetic

In Algebra, we often treat polynomials as "another kind of number". We can perform operations like greatest common divisor or division with remainder on polynomials too, as long as the ring over which we're building them is "nice" enough. In this course we only consider the case when the base ring is a field.

> **Proposition 4.4 (polynomial division with remainder).** Let $\mathbb{F}$ be a field and consider the polynomial ring $\mathbb{F}[X]$. For any two polynomials a, b with $b \neq 0$, there are unique polynomials q and r such that $a = q \cdot b + r$ and $\mathtt{deg}(r) < \mathtt{deg}(b)$.

Here the condition that the remainder be less than the modulus is replaced by the new condition that it be of lesser degree. The usual "long division" algorithm works fine for polynomials over any finite field (division of coefficients is performed in the field; this is why we need a field not just any ring). We will revisit this when we discuss finite fields in a later lecture.

Time for some actual polynomial arithmetic. Let's take the field $\mathbb{F}_7 = (\mathbb{Z}_7, +, \cdot)$ with the following addition/multiplication tables:

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

·	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

- Divide $(X^2 + 5X + 1)$ by $(3X + 2)$ with remainder.

$$\begin{array}{rrrrrrl}
(& X^2 & + & 5X & + & 1) & :(3X+2) = 5X+3 \\
-(& X^2 & + & 3X) & & & \\
\cline{1-4}
 & & & 2X & + & 1 & \\
 & & -(& 2X & + & 6) & \\
\cline{3-6}
 & & & & & 2 &
\end{array}$$

  To match up the highest coefficients, we need to solve $X^2 = 3X \cdot a$ which gives $a = 5X$ since $3 \cdot 5 = 1 \pmod 7$. We compute $5X(3X+2) = X^2+3X$ and subtract $(X^2+5X+1)-(X^2+3X) = (2X+1)$ giving $(X^2+5X+1) = 5X(3X+2)+(2X+1)$. The remainder $2X+1$ still has degree 1 so we divide with remainder again: first, solve $2X = 3X \cdot b$ to get $b = 3$, $3 \cdot (3X+2) = 2X+6$, $(2X+1)-(2X+6) = 2$ so

$$(X^2 + 5X + 1) = (5X + 3) \cdot (3X + 2) + 2$$

- Find the greatest common divisor of $(2X^2 + 4X + 5)$ and $(X^2 + 3X + 3)$.

  Euclid's algorithm for greatest common divisors says to repeatedly divide with remainder until a remainder becomes 0, the last non-zero remainder is then the greatest common divisor. So:

$$\begin{array}{rrrrrrl}
(& 2X^2 & + & 4X & + & 5) & :(X^2+3X+3) = 2 \\
-(& 2X^2 & + & 6X & + & 6) & \\
\cline{1-6}
 & & & 5X & + & 6 &
\end{array}$$

$$\begin{array}{rrrrrrl}
(& X^2 & + & 2X & + & 6) & :(5X+6) = 3X+1 \\
-(& X^2 & + & 4X) & & & \\
\cline{1-4}
 & & & 5X & + & 6 & \\
 & & -(& 5X & + & 6) & \\
\cline{3-6}
 & & & & & 0 &
\end{array}$$

  This gives us a greatest common divisor of $(5X + 6)$ and indeed $2X^2 + 4X + 5 = (5X + 6)(6X + 2)$ and $X^2 + 3X + 3 = (5X + 6)(3X + 4)$.

NOTE — factorisations, gcds etc. of polynomials are unique "up to units", just like in $\mathbb{Z}$. For example, what is the gcd of 12 and -9? For sure, 3 is a candidate but also -3 — both are common factors of the same magnitude. It's just that in $\mathbb{Z}$ we have the convention that we take the positive one (since the only units are 1 and -1, so there are only two to choose from). In $\mathbb{F}_7[X]$, there is a whole field to choose from. We could take $2X^2 + 4X + 5 = (5X + 6)(6X + 2)$ and multiply the first bracket with 2 and the second with the inverse of 2, which happens to be 4, to get $2X^2 + 4X + 5 = (3X + 5)(3X + 1)$ which is an equally correct factorisation.

Exercise. ($\star\star$) *Division with remainder.* Over $\mathbb{Z}_7[X]$, divide $p(X)$ by $q(X)$ with remainder:

- $p(X) = 2X^5 + 2X^2 + X + 4, \quad q(X) = 2X^2 + 1$
- $p(X) = 3X^3 + 2X^2 + 5X + 1, \quad q(X) = 3X + 2$
- $p(X) = 3X^3 + 2X^2 + 5X + 1, \quad q(X) = 2X + 6$

Exercise. ($\star\star$) *Division without remainder.* We know that a polynomial of degree $n > 0$ can have at most n distinct roots (values of a such that $p(a) = 0$) and this holds over any finite field. Further, we know that $(X - a)$ divides $p(X)$ (leaves remainder 0 in division with remainder) if and only if a is a root of p — this too holds over any field.

Over $\mathbb{Z}$, any polynomial has a unique factorisation. However, over $\mathbb{F}_7$, the polynomial $X^2 + 5X - 1$ divides all of the following: $(X + 2)$, $(X + 3)$, $(2X + 4)$, $(4X + 5)$, $(3X + 6)$, $(3X - 1)$, $(5X + 1)$.

1. Find the roots of $X^2 + 5X - 1$ in $\mathbb{F}_7$. Explain why these many different linear factors do not give more than 2 roots.
2. Give an example where a degree-2 polynomial can be written in several different ways as products of degree-1 polynomials without using any "modulus" operations, i.e. over a field where $1 + 1 + 1 + \ldots$ can never become 0.

4.4 Fields modulo polynomials

Just like we produced $\mathbb{Z}_n$ from $\mathbb{Z}$ by repeatedly subtracting n from anything that is n or above, we can take a polynomial ring modulo a polynomial p to get a ring of elements "less than" p. This is called taking the polynomial ring modulo p. While this works for any ring $\mathcal{R}$, it only really gives a nice structure to work in if we start with a field.

Formally, we start with any field $\mathbb{F}$ and any polynomial p in $\mathbb{F}[X]$ and define an equivalence relation $\sim_p$ under which two polynomials a, b are equivalent if there is a polynomial q such that $a + p \cdot q = b$, i.e. you can add a multiple of p to a to get b. This is an equivalence relation. We can do division with remainder in $\mathbb{F}[X]$ and the equivalence relation is the same as saying two elements are equivalent if and only if they leave the same remainder when dividing by p. Each equivalence class contains exactly one polynomial of degree less than $\mathtt{deg}(p)$ and we choose this as the representative of the class.

Definition 4.5 (field modulo polynomial). For a field $\mathbb{F}$, we mean by $\mathbb{F}[X]/p(X)$ the ring of equivalence classes of polynomials under $\sim_p$ where the representative of each class is the unique polynomial in that class of degree less than $\mathtt{deg}(p)$.

As an example, let's compute $(X^2 + X + 4) \cdot (2X + 3)$ in $\mathbb{F}_7[X]/(X^3 + X + 1)$.

$$\begin{aligned}
\left[(X^2 + X + 4)(2X + 3)\right] &= \\
\left[2X^3 + 5X^2 + 4X + 5\right] &= \\
\left[2(X^3 + X + 1) + (5X^2 + 2X + 3)\right] &= \\
5X^2 + 2X + 3 &
\end{aligned}$$

We reduce $\pmod{7}$ along the way; the last step is division with remainder by the modulus polynomial $X^3 + X + 1$.

From this example we see that since the modulus polynomial has degree 3, every polynomial in $\mathbb{F}_7[X]/(X^3+X+1)$ will have degree at most 2. So the size of the resulting ring will be 7^3 elements as polynomials in this ring can be represented by length-3 vectors of elements of $\mathbb{F}_7$.

We will soon use this idea to construct finite fields. Before we do so however we will use the next lecture to introduce homomorphisms which will help us understand better what is going on in these fields and when two diffferent-looking fields are "essentially" the same.

Exercise. $(\star\star)$ *The relation* $\sim_p$. Check that $\sim_p$ used in this section really is an equivalence relation. For an added challenge, check the same when the construction is done over an arbitrary ring instead of a field.

4.5 $\diamond$ Representatives modulo $\sim_p$

$\diamond$ Let's check that the representatives modulo $\sim_p$ do what we claim, i.e. that there is a unique element of each class with degree less than that of p. For any class C, pick any element c in the class and divide c by p with remainder; the remainder is a member of the same class but has degree less than $\mathtt{deg}(p)$. If two members a, b of the same class both have degree less than the degree of p then we divide a by p with remainder and must get b, but a already had degree less than that of p so $a = b$.

Note that over an arbitrary ring $\mathcal{R}$, equivalence classes can contain several different elements of low degree so there is no longer such an obvious choice of representatives.

4.6 $\diamond$ Euclidean domains

$\diamond$ Just like we can cancel $\cancel{a\cdot} x = \cancel{a\cdot} y$ in integral domains without being able to invert a, we can sometimes perform division with remainder without being able to invert ring elements. In this case we call the ring an Euclidean domain:

> **Definition 4.6 (Euclidean domain).** A ring $(R, +, \cdot)$ with a function $\mathtt{deg} : R\backslash\{0\} \to \mathbb{N}$ is an Euclidean domain if the function $\mathtt{deg}$ has the property that for all nonzero $a, b \in R$ we have $\mathtt{deg}(a \cdot b) \geq \mathtt{deg}(a)$ and for all $a, b \in R$ with $b \neq 0$ there exist q, r with $\mathtt{deg}(r) < \mathtt{deg}(b)$ so that we can write $a = q \cdot b + r$.

The degree function may be any function that satisfies the conditions given, not just the usual one for polynomials. We could define the degree of the ring's zero to be minus infinity if we wanted. The general definition just says that q, r exist but not that they are unique. If we take the ring $(\mathbb{Z}, +, \cdot)$ with the degree function $\mathtt{deg}(x) := |x|$, the absolute value, we still have a Euclidean domain but remainders are no longer unique (exercise: why? What are we doing differently to Lecture 1?)

If we take any field $\mathbb{F}$, the ring of polynomials $\mathbb{F}[X]$ is a Euclidean domain with the usual degree function and quotients and remainders are unique.

Lecture 5 — Homomorphisms

Dr. D. Bernhard

In this lecture: homomorphisms — isomorphisms — automorphisms and their group — application to groups, rings and fields

Learning outcomes. After this lecture and revision, you should be able to:

- Define homomorphisms and isomorphisms and give examples.
- Check, for simple examples, whether two structures are isomorphic and find an isomorphism if they are.
- Determine the isomorphism class of finite Abelian groups.
- Find integers with given remainders modulo different coprime moduli.
- Compute the Frobenius map in a ring and the inverse of an element in a finite field.

5 Homomorphisms

Imagine you're trying to solve a sudoku puzzle. Halfway through, you take out your laptop and open a sudoku solver program to check whether one of your guesses is correct. Except — this sudoku is in a paper that tries to be special and uses the letters A–I instead of the numbers 1–9, but your sudoku solver only accepts numbers as input. Does this make any difference? Of course not: you can map the letters to numbers, for example A=1, B=2 etc., then solve the numeric sudoku and map back again: if the program says that a certain field is a 3, you can write a C in the original sudoku there.

The moral of this story is that you have a mapping that respects the structure of a sudoku: you not only map the elements (letters) to other elements (numbers) and back again but any statement that you can make about the original "in the language of sudokus" maps to a statement about the numeric version and back again too. For example, "the remaining field in row one of the top left square must be an A or a B" is a statement that translates to the numeric version by replacing A and B with 1 and 2. Such a mapping lets us say that the two versions of the sudoku are "essentially the same, just written differently"; mathematicians would say the two are isomorphic.

Apart from isomorphisms, there are more general structure-respecting mappings that are not reversible: these are called homomorphisms.

5.1 Group homomorphisms

A homomorphism from a group $(G, +_G)$ to a group $(H, +_H)$ is a map from G to H that preserves the structure or language or groups. The language of groups revolves around the noun "neutral" and the verbs "add" and "invert". Note that to be pedantic, we are using different symbols for addition in the two groups.

> **Definition 5.1 (group homomorphism).** A function $f : G \to H$ is a group homomorphism between the groups $(G, +_G)$ and $(H, +_H)$ if it satisfies these three conditions.
>
> - It preserves neutral elements: if 0_G is the neutral element of $(G, +_G)$ and 0_H is the neutral element of $(H, +_H)$ then $f(0_G) = 0_H$.
> - It preserves addition. For any two elements a, b of G we have $f(a +_G b) = f(a) +_H f(b)$.
> - It preserves inverses: for any element a of G with inverse $-a$, $f(-a)$ is the inverse of $f(a)$.

We have discussed one particular group homomorphism a lot: for any positive integer n, the map $[\,]$ from $(\mathbb{Z}, +)$ to $(\mathbb{Z}_n, +_n)$ is a homomorphism. Indeed, $[0] = 0$, $[a+b] = [a]+_n[b]$ and $[a]+_n[-a] = 0$ so $[-a]$ is the inverse of $[a]$. This is one of the times when it really pays off to be pedantic with the addition symbol. The statement $[a+b] = [a]+_n[b]$ is the same as $[a + b] = [[a] + [b]]$ which we stated as a rule in lecture 1; it would be wrong to write $[a + b] = [a] + [b]$ for the usual addition in $\mathbb{Z}$ however. (Exercise: find a counter-example.)

Exercise. $(\star)$ *Trivial homomorphisms.*

- For any $n > 0$, describe the group homomorphisms between $(\mathbb{Z}_n, +)$ and $(\{0\}, +_0)$ in both directions, where $0 +_0 0 = 0$.
- Check that for any groups $(G, +_G)$ and $(H, +_H)$ the map $f : G \to H$ that sends every element of G to the neutral element of $(H, +_H)$ is a group homomorphism.
- Which other trivial homomorphism is there from any group $(G, +)$ to itself?

Exercise. $(\star\star)$ *Group homomorphisms and orders.*

- Explain why there are no non-trivial homomorphisms (that do not send everything to the neutral element) from $(\mathbb{Z}_2, +_2)$ to $(\mathbb{Z}_3, +_3)$ or back again.

- What about $(\mathbb{Z}_2, +_2)$ to $(\mathbb{Z}_4, +_4)$ (and back again)?
- Consider the following situation. $(G, +_G)$ and $(H, +_H)$ are groups and $g \in G$ is an element of finite order $n > 0$, in particular $\underbrace{g +_G \ldots +_G g}_{n \text{ times}} = 0_G$. Show that if $f : G \to H$ is a homomorphism between these groups and $h = f(g)$ then $\underbrace{h +_H \ldots +_H h}_{n \text{ times}} = 0_H$.
- Show that group homomorphisms do not have to preserve the order of elements: find an example of two groups as above, an element $g \in G$ and a group homomorphism f such that the order of $f(a)$ is not the same as the order of a.

$(\star\star\star)$ The precise rule is that if $(G, +_G)$ and $(H, +_H)$ are groups and $f : G \to H$ is a group homomorphism between them then the order of $f(g)$ divides the order of g for any $g \in G$. Prove this.

5.2 Isomorphisms

A homomorphism allows you to translate statements one way but not back. For example, the statement "adding any element to itself gives 0" holds in the group $(\mathbb{Z}_2, +_2)$ but not in $(\mathbb{Z}, +)$ so there is no way to translate all possible statements in group-language back again. If a homomorphism does have an inverse, it is called an isomorphism.

Definition 5.2 (isomorphism). A homomorphism f from $\mathbb{G} = (G, +_G)$ to $\mathbb{H} = (H, +_H)$ is called an isomorphism if there is a homomorphism s from $\mathbb{H}$ to $\mathbb{G}$ such that f and s are inverses, i.e. for any $g \in G$ we have $s(f(g)) = g$ and for any $h \in H$ we have $f(s(h)) = h$.

Two groups are called isomorphic if there is an isomorphism between them (being isomorphic is an equivalence relation).

For example, consider the following three groups.

$\mathbb{G} = (\{0, 1, 2, 3\}, +)$

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

$\mathbb{H} = (\{A, B, C, D\}, \oplus)$

$\oplus$	A	B	C	D
A	A	B	C	D
B	B	C	D	A
C	C	D	A	B
D	D	A	B	C

$\mathbb{I} = (\{0, 1, 2, 3\}, \boxplus)$

$\boxplus$	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

The groups $\mathbb{G}$ and $\mathbb{H}$ are isomorphic: the isomorphism from $\mathbb{G}$ to $\mathbb{H}$ is the function f with $f(0) = \mathrm{A}$, $f(1) = \mathrm{B}$, $f(2) = \mathrm{C}$, $f(3) = \mathrm{D}$. However, $\mathbb{G}$ and $\mathbb{I}$ are not isomorphic. For example, adding any element to itself in $\mathbb{I}$ gives the neutral element but this is not true in $\mathbb{G}$, so there cannot be any function f from $\mathbb{I}$ to $\mathbb{G}$ that preserves neutral elements and addition.

The notion of isomorphism is very general and powerful in mathematics. The way we gave our definition, there is nothing specific to groups so once we know what a ring or field homomorphism is, we have got a definition of ring and field isomorphisms for free. The same idea could be carried over to other kinds of structure but we would have to give a more abstract definition of what it means for two homomorphisms to be inverses (since they are no longer necessarily functions on sets) which we can happily leave to professional mathematicians.

5.3 Group isomorphisms

There are only two "really different" groups with four elements, both of which we have just seen: one which is "essentially" $(\mathbb{Z}_4, +_4)$ and one in which every element added to itself is zero. The notion of isomorphism allows us to make clear what we mean by this statement: there are only two groups of order 4 "up to isomorphism". To define this formally, let $\sim$ be the equivalence relation "is isomorphic to" for groups. (Exercise: check that this really is an equivalence relation.) Then there are exactly two different equivalence classes that contain groups of order 4. Two groups of different orders cannot be isomorphic so these two classes contain only groups of order 4.

For prime numbers the classification is even simpler:

> **Proposition 5.3.** For any prime number p, there is only one group of order p up to isomorphism.

Another way of putting this is that if we have any two groups of order p where p is a prime then they are automatically isomorphic. Sensibly, one chooses $(\mathbb{Z}_p, +_p)$ as the representative element of this class of group.

We can extend this classification to all finite groups but we need to introduce one more concept to do this.

Exercise. $(\star\star)$ *Isomorphisms preserve group orders.* Show that if $(G, +_G)$ and $(H, +_H)$ are groups and $f : G \to H$ is an isomorphism then for every element $g \in G$, the order of g is the same as the order of $f(g)$.

Exercise. ($\star\star$) *Isomorphisms between additive and multiplicative groups.*

- Start with the group $(\mathbb{Z}_7^\times, \cdot)$ and consider the subgroup $\langle 3 \rangle$. This subgroup has 3 elements; find an isomorphism between this group and $(\mathbb{Z}_3, +)$.
- Find the other isomorphism between the above two groups (there are exactly two).
- There is one isomorphism f from $(\mathbb{Z}_{10}, +)$ to $\langle 2 \rangle \subset (\mathbb{Z}_{11}^\times, \cdot)$ that has $f(1) = 2$. First, find it. Secondly, what is the obvious formula for this isomorphism?

5.4 Products of groups

If $\mathbb{G} = (G, +_G)$ and $\mathbb{H} = (H, +_H)$ are two groups we can form a further group by taking all the pairs of elements (g, h) with $g \in G$ and $h \in H$ and adding component-wise, i.e. the sum of (g, h) and (g', h') is $(g +_G g', h +_H h')$. Since the set of elements of this group is $G \times H$, we call this group $\mathbb{G} \times \mathbb{H}$, the product of groups $\mathbb{G}$ and $\mathbb{H}$.

> **Definition 5.4 (product of groups).** For a finite list of groups $\mathbb{G}_1, \ldots, \mathbb{G}_n$, the product group $\mathbb{G}_1 \times \ldots \times \mathbb{G}_n$ is the group whose elements are tuples of n elements where the i-th element is in $\mathbb{G}_i$, with component-wise addition. For the n-fold product of a group with itself we also write $\mathbb{G}^n$.

By now, we should all be able to find the formulas for the neutral element and the inverse of elements in product groups. Products of rings and fields (and when we introduce them, vector spaces) work in much the same manner.

5.5 Classification of finite Abelian groups

Product groups let us describe all finite Abelian groups:

> **Theorem 5.5 (classification of finite Abelian groups).** Every finite Abelian group $\mathbb{G}$ is isomorphic to exactly one group of the form $\mathbb{Z}_{p_1^{m_1}} \times \ldots \times \mathbb{Z}_{p_k^{m_k}}$ where the p_i are primes and the m_i positive integers.

For a group of order n, we have $n = p_1^{m_1} \cdot \ldots \cdot p_k^{m_k}$ in this decomposition. Groups of the same order can still differ in whether prime powers are inside or outside the subscripts: the two non-isomorphic groups of order 4 are $\mathbb{Z}_{2^2} = \mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2 = (\mathbb{Z}_2)^2$.

By this theorem, a cyclic group of composite order — say, $(\mathbb{Z}_{15}, +)$ — should be isomorphic to a product group $\mathbb{Z}_3 \times \mathbb{Z}_5$. Indeed, if we take any element x of $\mathbb{Z}_{15}$, we can write it as the pair $(x \pmod 3, x \pmod 5)$ to get an element of $\mathbb{Z}_3 \times \mathbb{Z}_5$ and for any such pair, there is exactly one element in $\mathbb{Z}_{15}$ with the given decomposition, as in the table below.

0	(0, 0)	5	(2, 0)	10	(1, 0)
1	(1, 1)	6	(0, 1)	11	(2, 1)
2	(2, 2)	7	(1, 2)	12	(0, 2)
3	(0, 3)	8	(2, 3)	13	(1, 3)
4	(1, 4)	9	(0, 4)	14	(2, 4)

Exercise. $(\star\star)$ *Another decomposition.* Decompose $(\mathbb{Z}_{12}, +)$ into $\mathbb{Z}_4 \times \mathbb{Z}_3$.

5.6 The group of automorphisms

Isomorphism is an equivalence relation. In particular, if f is an isomorphism from $\mathbb{G}$ to $\mathbb{H}$ and g is an isomorphism from $\mathbb{H}$ to $\mathbb{K}$ then the composition gf (as a function on the underlying sets, i.e. for each element x of $\mathbb{G}$ we have $gf(x) := g(f(x))$, an element of $\mathbb{K}$). Since we can compose and invert isomorphisms, can we make them into a group? Not necessarily — we cannot compose any two isomorphisms, only ones with compatible domains. If $f : \mathbb{G} \to \mathbb{H}$ and $k : \mathbb{K} \to \mathbb{L}$ then we cannot compose f and k. We can always compose isomorphisms if they start and end at the same object though:

Definition 5.6 (automorphism). An isomorphism from a group (or ring, field) $\mathbb{G}$ to itself is called an automorphism of $\mathbb{G}$. The automorphisms of any object $\mathbb{G}$ form a group called Aut($\mathbb{G}$) with composition as the operation.

Aut($\mathbb{G}$) is a group because function composition is associative, the identity map that sends every element of $\mathbb{G}$ to itself is an automorphism and forms the neutral element of Aut($\mathbb{G}$) and isomorphisms are invertible by definition.

The definition of an automorphism group applies equally to rings, fields and many other structures; the automorphisms themselves always form a group, whichever structure one looks at.

Exercise. $(\star\star)$ *Group automorphisms.*

- There is exactly one nontrivial automorphism from $(\mathbb{Z}_3, +)$ to itself (that is neither the identity map nor sends everything to the neutral element). Find it.
- Find the automorphism groups of $(\mathbb{Z}_5, +)$ and $(\mathbb{Z}_6, +)$ as well. Hint: automorphisms preserve orders of elements, so they must preserve generators of finite groups as well.
- Find the automorphism group of $\mathbb{Z}_2 \times \mathbb{Z}_2$.

5.7 Ring homomorphisms

To adapt the notion of homomorphism to rings, we just have to take care of multiplication as well.

Definition 5.7 (ring homomorphism). Let $\mathcal{R} = (R, +, \cdot)$ and $\mathcal{S} = (S, \oplus, \odot)$ be two rings. A function $f : R \to S$ is a ring homomorphism if f is a group homomorphism from $(R, +)$ to $(S, \oplus)$ and these two conditions hold:

- For any $a, b \in R$ we have $f(a \cdot b) = f(a) \odot f(b)$.
- We have $f(1_R) = 1_S$ where 1_R is the one (neutral element of multiplication) of $\mathcal{R}$ and 1_S is the one of $\mathcal{S}$.

If we look at a ring homomorphism from a ring $\mathcal{R}$ to itself, there is very little freedom. We know that for such a f, for all r, s we have $f(r+s) = f(r)+f(s)$ and $f(rs) = f(r)f(s)$ along with $f(1) = 1$. So for any $k \in \mathbb{Z}$ and for the ring element

$$r := \underbrace{1 + 1 + \ldots + 1}_{k \text{ times}}$$

we have $f(r) = f(1) + \ldots + f(1) = 1 + \ldots + 1 = r$, so ring homomorphisms from a ring to itself cannot change multiples of 1. For example, in $\mathbb{Z}$, the identity function is the only ring homomorphism.

WARNING: A ring homomorphism is not the same as a linear function (which we will introduce later). A linear function must satisfy $f(ax) = af(x)$ rather than $f(ax) = f(a)f(x)$, which gives it a lot more freedom.

In polynomial rings over finite fields, the interesting part of a ring homomorphism f is therefore what it does to X (which is not a multiple of 1). Since $f(X^2) = f(X) \cdot f(X)$ and so on, the behaviour of a ring homomorphism on such a polynomial ring is determined by its behaviour on X, since we will see in a moment that all non-zero field elements are multiples of 1.

Let's look at multiples of 1 in a ring again. We can construct a function $m : \mathbb{N} \to \mathcal{R}$ for any ring $\mathcal{R}$ that takes n to

$$m(n) := \underbrace{1 + \ldots + 1}_{n \text{ times}}$$

where 1 is the one of the ring. For 0, we set $m(0) := 0$, that is evaluating m at the number zero gives the zero of the ring. We can extend this function to $\mathbb{Z}$ by setting $m(a) := -m(-a)$ for negative a, i.e. to evaulate on a negative integer you invert the integer to get a positive one, compute m on that and then invert back again in the ring. This function m is now a ring homomorphism $\mathbb{Z} \to \mathcal{R}$. If $\mathcal{R} = \mathbb{Z}$ then this m-function is the identity. In fact,

Proposition 5.8. For any ring $\mathcal{R}$, there is exactly one ring homomorphism from $\mathbb{Z}$ to $\mathcal{R}$ and it is the map

$$m(z) := \begin{cases} \underbrace{1_R + \ldots + 1_R}_{z \text{ times}} & z > 0 \\ 0_R & z = 0 \\ -m(-z) & z < 0 \end{cases}$$

where 1_R and 0_R are the one and zero elements of the ring.

With this homomorphism in place, we can define the characteristic of a ring. It is a similarly important number for rings as the order is for groups, although it does not always count elements.

Definition 5.9 (characteristic). The characteristic $\mathsf{char}(\mathcal{R})$ of a ring $\mathcal{R}$ is the smallest positive integer z for which $m(z) = 0$, the zero of the ring. If no such integer exists, the characteristic is 0 (the zero of $\mathbb{Z}$).

For example, $\mathsf{char}(\mathbb{Z}) = 0$ and $\mathsf{char}(\mathbb{Z}_n) = n$. But, $\mathsf{char}(\mathbb{Z}_n[X]) = n$ too, so the characteristic does not "count" elements that are not multiples of 1.

If p is a prime number and $\mathcal{R}$ is a commutative ring of characteristic p, the map $\phi : \mathcal{R} \to \mathcal{R}, x \mapsto x^p$ has the property that $\phi(a)\phi(b) = \phi(ab)$ — this is just the usual formula $a^p b^p = (ab)^p$ that holds in any commutative ring. But if we write out the expansion of $\phi(a + b) = a^p + b^p$, all intermediate terms gain a factor p and vanish: we get $(a + b)^p = a^p + b^p$.

Proposition 5.10 (Frobenius map). In a commutative ring $\mathcal{R}$ with prime characteristic p, the map $\phi : x \mapsto x^p$ is a ring homomorphism, called the Frobenius map.

5.8 Field homomorphisms

A field is a ring and all the information we need to make a field homomorphism is contained in the ring structure already.

> **Definition 5.11 (field homomorphism).** Let $\mathbb{F} = (F, +, \cdot)$ and $\mathbb{K} = (K, \oplus, \odot)$ be fields. A function $f : F \to K$ is a field homomorphism if it is a ring homomorphism.

Don't we have to check that, for example, $f(a/b) = f(a) \oslash f(b)$? This comes for free: since $f(a/b \cdot b) = f(a)$ and know that $f(a/b) \odot f(b) = f(a)$, we can conclude that $f(a/b) = f(a) \oslash f(b)$.

Like for groups, we can use isomorphisms to classify finite fields. We will do this in the next lecture; for now we find a few more properties of field homomorphisms.

> **Proposition 5.12.** In a finite field of characteristic p for a prime p, the Frobenius map $\phi : x \mapsto x^p$ is an isomorphism (and therefore an automorphism).
>
> Further, such a field has p^n elements for some positive integer n and applying ϕ in sequence n times gives the identity map, i.e. for any field element x we have $x^{(p^n)} = x$.

We will see in the next lecture that all finite fields are of this form. A useful formula to remember in finite fields is that the Frobenius map can be used to invert elements: since $x^{p^n-1} = 1$, we must have $x^{p^n-2} = 1/x$. The fact that all finite field elements become one when raised to a certain power is also interesting to study field homomorphisms.

> **Definition 5.13 (root of unity).** A field element x is called a k-th root of unity if $x^k = 1$ in the field, for a positive integer k. If $x^k = 1$ and $x^m \neq 1$ for all $1 \leq m < k$, we say that x is a primitive root of unity. (In this case, k is the order of x in the multiplicative group $(\mathbb{F} \setminus \{0\}, \cdot)$ of the field.)

In finite fields, all elements are roots of unity. But, by the same reasoning that we know over the reals why a non-zero polynomial of degree n cannot have more than n roots, we know that in a finite field there cannot be more than n elements that are n-th roots of unity, i.e. roots of the polynomial $X^n - 1$. This tells us a lot about the orders of elements in finite fields.

The important fact that we need to take away to investigate finite fields is that field homomorphisms preserve roots of unity: if $x^k = 1$ then for any field homomorphism

f we have $f(x^k) = f(x)^k$ so $f(x)$ is still a k-th root of unity. Isomorphisms are even better:

> **Proposition 5.14.** If $\mathbb{F}, \mathbb{K}$ are fields and $f : \mathbb{F} \to \mathbb{K}$ is a field isomorphism then f preserves element orders, specifically $f(x)^k = 1$ if and only if $x^k = 1$.

So if we introduce an equivalence relation on a field where two elements are equivalent if they have the same multiplicative order, any field automorphism can only permute elements around within the classes but never move an element between classes.

Exercise. $(\star\star)$ *Finite field inversion.* Let $p = 1033$, this is a prime number. The exercise is to compute $1/3$ in the finite field $\mathbb{F}_p$ — without a computer (so don't just program a loop that tries all possibilities), though you may use a calculator that supports the modulo operation. Hint: 1031 is 0100 0000 0111 in binary.

5.9 ◇ More on finite Abelian groups

◇ The formula $\mathbb{Z}_{15} \cong \mathbb{Z}_3 \times \mathbb{Z}_5$ is just a special case of a theorem that is supposed to have originated in ancient China.

> **Theorem 5.15 (Chinese remainder theorem).** Let $n_1, \ldots, n_k$ be positive integers such that no two of these have a factor in common. Then for any integers $a_1, \ldots, a_k$ with $0 \leq a_i \leq n_i$ for all i there is exactly one integer x satisfying all the equations $x = a_i \pmod{n_i}$.

◇ The classification theorem for finite Abelian groups can be generalised to finitely generated Abelian groups, where there is a finite set of elements $x_1, \ldots, x_n$ such that $g = \langle x_1, \ldots, x_n \rangle$. In this case, every such group is isomorphic to the direct product of a finite Abelian group as above and a group $\mathbb{Z}^r$ for a unique value of r, which is called the rank of the group.

◇ One can also give a classification theorem for all finite groups, dropping the Abelian requirement. A theorem by Jordan and Hölder states that all finite groups are composed of finite simple groups (the "primes" of the group world, where every subgroup of a certain form has to be trivial). It remains to classify all finite simple groups — the result is one of the masterpieces of Algebra, completed for the first time (assuming no mistakes) in 2008. A revised version of the proof is being edited for publication and is expected to run to several thousand pages.

Lecture 6 — Finite Fields

Dr. D. Bernhard

In this lecture: finite commutative rings without zero divisors are fields — classification of finite fields — irreducible polynomials — construction of finite fields $GF(p)$ and $GF(p^n)$ — computing in finite fields — isomorphisms between finite fields

Learning outcomes. After this lecture and revision, you should be able to:

- Decide whether or not a finite field of a given size exists.
- Construct a finite field of any size where one exists.
- Compute in finite fields.
- Find isomorphisms between different representations of the same finite field.

6 Finite Fields

A field is a structure in which you can add, subtract, multiply and divide (except by zero). Today we are going to look at finite fields. Recall that $\mathbb{Z}_{256}$ is not a field because of zero divisors. Can we construct a field with 256 elements? We can, but not with the usual addition modulo 256. The first question we have to answer to get our field with 256 elements is, when is a finite ring a field?

6.1 Finite commutative rings

Take any ring $(R, +, \cdot)$. We know that we can classify all elements as zero, unit, zero divisor or neither. But if our ring is finite, the "neither" case cannot happen.

Let's pick an element x in a finite ring that is neither zero nor a zero divisor. This means that if we look at the sequence $x, x \cdot x, x \cdot x \cdot x, \ldots$, we can never hit zero. We can obviously write this as $x, x^2, x^3, \ldots$. But if the ring is finite, at some point an element has to repeat so we get an equation of the form $x^k = x^{k+m}$ for positive integers k, m from which we conclude that $x^m = 1$ (you can cancel x^k since x is not a zero divisor, therefore neither is x^k). Whether or not the ring is commutative, the associative law implies that powers of x commute with each other. Therefore $x \cdot x^{m-1} = 1$ and $x^{m-1} \cdot x = 1$, so x^{m-1} really is the inverse of x under multiplication.

In other words, the only way a nonzero ring element can be neither a unit nor a zero divisor is if the ring is infinite (for example, 3 in $\mathbb{Z}$). In a finite ring such as $\mathbb{Z}_n$, as soon as a nonzero element is not a zero divisor it automatically has an inverse. This proves the following proposition:

> **Proposition 6.1.** A finite commutative ring without zero divisors is a field.

We know that the rings $(\mathbb{Z}_n, +, \cdot)$ satisfy these conditions exactly when n is a prime, except for the special case $(\{0\}, +, \cdot)$. (For a non-prime n, the structure $(\mathbb{Z}_n^\times, \cdot)$ is a group but such a $\mathbb{Z}_n^\times$ is no longer a group under addition so we cannot construct a ring this way, let alone a field.) This means that for each prime p, we can construct a field with p elements.

6.2 Classification of finite fields

Finite fields are relatively "rare" objects. The following theorem describes exactly which ones exist:

> **Theorem 6.2 (classification of finite fields).** For every prime p and every positive integer n, there is exactly one finite field with p^n elements up to isomorphism. These are the only finite fields.

The field with p^n elements can be written either $\mathbb{F}_{p^n}$ or $GF(p^n)$ and pronounced "Galois Field" after the French mathematician E. Galois. The power operator is always left in the description, i.e. one writes $GF(2^8)$ not $GF(256)$.

6.3 Prime fields

The simplest finite fields are those for power $n = 1$: these are just the fields $GF(p) = (\mathbb{Z}_p, +, \cdot)$ with the usual addition and multiplication modulo p that we constructed above. Every other attempt to construct a finite field of order p will produce one isomorphic to the above construction.

6.4 Irreducible polynomials

Here is the general idea to construct a field with p^n elements. Start with the field $GF(p)$ and form the polynomial ring $GF(p)[X]$. This has infinitely many elements. Then, take this ring modulo a polynomial of degree n to get a ring of p^n elements (sequences of n elements from $\mathbb{Z}_p$). As long as we do not end up introducing any zero divisors, since this ring is finite and commutative it is automatically a field.

How do we prevent zero divisors? When going from $\mathbb{Z}$ to $\mathbb{Z}_n$ (viewed as rings), zero divisors are exactly the non-unit elements that divide n so we don't get any if we pick n to be a prime.

I've said before that in Algebra, polynomials are just "another kind of number". So if we pick a polynomial to divide by that is a "prime polynomial", we should not get any zero divisors for exactly the same reason.

We have to quickly get some terminology out of the way. In general algebra, there are two distinct concepts "prime" and "irreducible". What we actually need to avoid zero divisors are irreducible moduli; both the integers and polynomial rings over fields are special cases where prime and irreducible turn out to be the same thing (technically, so-called unique factorisation domains). While "prime number" is the common term for the integers, when talking about polynomials it is more usual to talk about "irreducible polynomials" which we will do from now on. For the purposes of this course, it is not too inaccurate to imagine "irreducible" to mean "something like prime"

> **Definition 6.3 (irreducible polynomial).** A polynomial p over a field is irreducible if it is not a unit and cannot be decomposed into a product of non-units, i.e. if $p = ab$ then either a or b is a unit.

The definition of irreducible applies not only to polynomial rings over fields but to any integral domain, though we will only be using it for polynomials.

Exercise. ($\star$) If $p = ab$ for an irreducible p, why can't both a and b be units?

In $\mathbb{Z}$ (which is an integral domain), the irreducible elements are exactly those numbers p which are not 1 or -1 where $p = ab$ implies that one of a or b is 1 or -1, in which case the other must be p or $-p$ — these are exactly the usual prime numbers, of course.

6.5 Finding irreducible polynomials

Let's take $GF(7)$ as an example field (the rest of this section works equally well with any finite field) and look at polynomials of low degrees in $GF(7)[X]$.

- The polynomial 0 is the zero element, so it is not irreducible (0 is not a unit).

- Polynomials of degree 0 are sequences (c_0) where c_0 is a non-zero field element. All of these are units and therefore not irreducible (the same reason that 1 is not a prime).
- Polynomials of degree 1 can be written $aX + b$ for field elements a, b with $a \neq 0$. In the polynomial ring over a field without any polynomial "modded out", a polynomial of degree 1 or higher cannot be a unit (there is no "$1/X$" to remove the X). Since all non-zero non-units have degree of at least 1, if a, b are two such polynomials then ab has degree at least 2 so all polynomials of degree 1 are irreducible.
- Degree 2 is where it starts to get interesting. Certainly, any polynomial that is the product of two degree-1 polynomials is not irreducible. Since we can always multiply a polynomial through with the inverse of the leading coefficient, let's consider only polynomials of the form $X^2 + bX + c$. If such a polynomial factors, it will be without loss of generality into the form $(X + u)(X + v)$ which gives $b = u + v$ and $c = uv$. In a finite field, we could in principle make a table with columns $u, v, u + v, uv$ for all values of u and v. To check if a polynomial is irreducible, we see if the pair (b, c) appears in the $(u + v, uv)$ columns anywhere — if so, we have factored the polynomial, otherwise it is irreducible.

  A simple counting argument now shows that there must always be an irreducible polynomial of degree 2 in a field of the form $GF(p)$. There are p possible values each for b, c so there are p^2 quadratic polynomials with leading coefficient 1. Similarly, since our table runs through p values each of u and v, there are p^2 rows in the table. The argument is that if any two rows repeat an $(u + v, uv)$ pair then at least one of the possible (b, c) pairs cannot appear in the table at all. And indeed, for any distinct values of u and v, the rows starting (u, v) and (v, u) will have the same sum and product. For example $(0, 1, 1, 0)$ and $(1, 0, 1, 0)$ both have the same sum of 1 and product of 0.

	u	**v**	$\mathbf{b = u + v}$	$\mathbf{c = uv}$
	0	0	0	0
⇒	0	1	1	0
	...			
	0	$p-1$	$p-1$	0
⇒	1	0	1	0
	...			

  Table of all possible factorisations of quadratic polynomials, showing repeated b, c entries in two different rows.

  As an example, the polynomial $X^2 + X + 6$ is irreducible over $GF(7)$.
- The situation with polynomials of degree 3 or higher is more complex and we don't treat it here. Suffice it to say that irreducible polynomials of any degree > 0

always exist.

Exercise. ($\star\star$) *Irreducible polynomials.* Find the the following irreducible polynomials:

1. Over $GF(2)$, all irreducible polynomials of degrees 2 and 3.
2. Over $GF(2)$, one irreducible polynomial of degree 4.
3. Over $GF(3)$, all irreducible polynomials of degree 2.
4. Over $GF(5)$, one irreducible polynomial of degree 3.

6.6 Example: $GF(7^2)$

Let's look at some example finite fields. To construct $GF(7^2)$ we take $\mathbb{F}_7[X]/(X^2+X+6)$, giving 49 field elements which we can represent as pairs (a,b) or equivalently, as linear polynomials $(a+bX)$. Addition in this field is just component-wise addition in $\mathbb{F}_7$. To multiply two elements (a,b) and (c,d), viewing them as polynomials we get $bdX^2+(bc+ad)X+ac$ which we have to reduce modulo X^2+X+6. So we factor out bd and rewrite the product as

$$bd\cdot(X^2+X+6)+(bc+ad-bd)\,X+(ac-6bd)$$

for the linear and constant terms, we have "telescoped" out the required factor bd. This gives us the following multiplication formula for this particular representation of the field, using $-6=1$:

$$(a,b)\cdot(c,d)=(ac+bd,bc+ad-bd)$$

6.7 Automorphisms of $GF(7^2)$

Let's find the automorphisms of $GF(7)[X]/(X^2+X+6)$, that is the functions f on this domain with $f(x+y)=f(x)+f(y)$, $f(0)=0$, $f(1)=1$ and $f(xy)=f(x)f(y)$. All that we need to determine an automorphism is $f(X)$, since for any element (a,b) of the field we have $f(a+bX)=a+b\cdot f(X)$.

We start with setting $f(X)=u+vX$ for variables u,v. Then we have $f(X)\cdot f(X)=(u+vX)(u+vX)=u^2+2uvX+v^2(X^2+X+6)-v^2(X+6)=(u^2+v^2)+(2uv-v^2)X$. However, we also have $f(X)\cdot f(X)=f(X^2)=f(-X-6)=(1-u)-vX$, giving us the equations $1-u=u^2+v^2$ and $-v=2uv-v^2$. The last equation gives us two cases: either $v=0$, which is definitely not an automorphism, or $v\neq0$ in which case we divide by v to get $-1=2u-v$ and substitute to get the quadratic equation $1-u=u^2+(2u+1)^2$ which gives us the solutions $u=0,v=1$ and $u=6,v=6$. Our automorphisms are $f_1(X) = X$ and $f_2(X) = 6 + 6X$, from which we find $f_1(a + bX) = a+bX$ — the identity function, which is not surprising — and $f_2(a+bX) = (a+6b)+6bX$.

Exercise. ($\star$) Why can $v = 0$ in the above calculation not yield an automorphism?

6.8 Isomorphisms in $GF(7^2)$

We said that there is only one finite field $GF(p^n)$ up to isomorphism for each prime p and positive integer n. Let's look at some examples of this too.

For $GF(7^2)$, another representation of the same field comes from choosing a different irreducible polynomial, such as Y^2+1 which gives the multiplication $(a, b)\odot(c, d) = (ac - bd, ad + bc)$.

Let's try and compute the isomorphisms

$$f : GF(7)[X]/(X^2 + X + 6) \to GF(7)[Y]/(Y^2 + 1)$$

We will use the symbol X for elements in the first representation and the symbol Y for elements in the second; this way the symbol name tells us which polynomial we have to use when reducing elements after multiplication.

We know that $f(1, 0) = (1, 0)$ and thus that $f(a, 0) = (a, 0)$ for any field element $a \in GF(7)$ since 1 is a generator of $(\mathbb{Z}_7, +)$. So all we need to find is $f(0, 1) = f(X)$ since $f(a + bX) = a + b \cdot f(X)$. Writing $f(X) = u + vX$ for variables u, v ranging over $GF(7)$, for any element (a, b) we have $f(a, b) = (a + ub, vb)$. Now look at the equation $f(a, b) \odot f(c, d) = f((a, b) \cdot (c, d))$ that any isomorphism must satisfy. Writing this out and combining terms gives the conditions $2uv = -v$ and $u^2 - v^2 = 1 - u$ which give $u = 3$ and $v = 2 \lor v = 5$. So we have two isomorphisms

$$\begin{aligned} f_1 &: (a, b) \mapsto (a + 3b, 2b) \\ f_2 &: (a, b) \mapsto (a + 3b, 5b) \end{aligned}$$

We can describe both these isomorphisms by their action on the polynomial $(0, 1)$ that represents the monomial X: $Y_1 = f_1(X) = 3 + 2X$ and $Y_2 = f_2(X) = 3 + 5X$.

Exercise. ($\star\star$) *The field* $GF(5^2)$.

1. Find the explicit multiplication formula for the representation of $GF(5^2)$ modulo the irreducible polynomial $X^2 + 2X + 3$.
2. Do the same for the irreducible polynomial $2Y^2 + 4Y + 1$.
3. Find the isomorphisms from the first representation to the second.

6.9 Example: $GF(2^8)$

For $GF(2^8)$, our field with 256 elements, we take the irreducible polynomial

$$p(X) = X^8 + X^4 + X^3 + X + 1$$

as an example. In addition to tuples and expressions with a variable X, we have a third representation of $GF(2^8)$ as 8-bit strings with the lowest coefficient rightmost, i.e. the polynomial $X^3 + X + 1$ which is $(1, 1, 0, 1)$ as a tuple can be written `00001011`. The operations on the individual bits, as elements of $GF(2)$, are:

+	0	1
0	0	1
1	1	0

·	0	1
0	0	0
1	0	1

These are, of course, the binary exclusive-or (XOR) and AND operations. Addition of tuples is component-wise; for multiplication we could write out the formula as for $GF(7^2)$ but this becomes cumbersome. Instead, we give an example. First, we consider the multiplication of binary polynomials without any polynomial modulus. The special thing about working over $GF(2)$ is that $1 + 1 = 0$ so all coefficients in our polynomials are either present or absent but we don't have to worry about field multiplication too much.

Suppose we want to compute the product of the polynomials represented by the bytes `10001010` and `00101101` in $\mathbb{F}_2[X]$. Just like "normal" multiplication of bytes (as perfomed by the x86 `MUL` operation), the result will be a 2-byte value. Writing these out,

$$\begin{array}{rcl} \texttt{10001010} & = & X^7 + X^3 + X \\ \texttt{00101101} & = & X^5 + X^3 + X^2 + 1 \end{array}$$

we can factor out the second operand and write this multiplication in the form

$$X^7(X^5 + X^3 + X^2 + 1) + X^3(X^5 + X^3 + X^2 + 1) + X(X^5 + X^3 + X^2 + 1)$$

Each left-hand side of a product in this term is a monomial with coefficient 1 (the only nonzero element of the base field). But multiplying with X^k like this is just shifting the right-hand factor to the left by k bits. So we can do polynomial multiplication by repeated addition in the usual longhand way:

$$\begin{array}{rcr} \texttt{10001010} \cdot \texttt{00101101} & = & \texttt{0 0101101.} \\ & + & \texttt{001 01101...} \\ & + & \texttt{0010110 1.......} \\ \hline & & \texttt{00010111 10110010} \end{array}$$

which is the polynomial $X^{12} + X^{10} + X^9 + X^8 + X^7 + X^5 + X^4 + X$.

To take such a polynomial modulo $X^8+X^4+X^3+X+1$, we write the two bytes that make up the product as (hi, lo) so the polynomial is actually $hi \cdot X^8 + lo$. Since lo is of degree at most 7 it does not need to be reduced any further. For hi, we write out the division with remainder by the modulus polynomial $p(X)$ of all higher powers. For example, $X^8 = 1 \cdot (X^8+X^4+X^3+X+1) + (X^4+X^3+X+1)$.

power	**q**	**r**	binary **r**
X^8	1	X^4+X^3+X+1	`00011011`
X^9	X	$X^5+X^4+X^2+X$	`00110110`
X^{10}	X^2	$X^6+X^5+X^3+X^2$	`01101100`
X^{11}	X^3	$X^7+X^6+X^4+X^3$	`11011000`
X^{12}	X^4+1	$X^7+X^5+X^3+X+1$	`10101011`
X^{13}	X^5+X+1	$X^6+X^3+X^2+1$	`01001101`
X^{14}	X^6+X^2+X	$X^7+X^4+X^3+X$	`10011010`
X^{15}	$X^7+X^3+X^2+1$	$X^5+X^3+X^2+X+1$	`00101111`

With this table, we can just add the lo component of our product to the remainders of all the powers present in the hi component:

$$\begin{array}{rlr} & \texttt{10110010} & (lo) \\ + & \texttt{00011011} & X^8 \\ + & \texttt{00110110} & X^9 \\ + & \texttt{01101100} & X^{10} \\ + & \texttt{10101011} & X^{12} \\ \hline & \texttt{01011000} & \end{array}$$

This gives us our result in $GF(2^8)$, represented with the irreducible polynomial $X^8 + X^4+X^3+X+1$, of $\texttt{10001010}\cdot\texttt{00101101} = \texttt{01011000}$ or $(X^7+X^3+X)\cdot(X^5+X^3+X^2+1) = (X^6+X^4+X^3)$.

6.10 Implementation of $GF(2^8)$ multiplication

Here is binary multiplication in $\mathbb{F}_2[X]/p(X)$ written as `C` code, where `u8` is an unsigned 8-bit integer datatype and `bool` is a boolean datatype (`int` would do fine as well):

```
/*
Multiply two values a, b in GF(2^8) represented by
p(X) = X^8 + X^4 + X^3 + X + 1  (0x1b).
*/
u8 mul(u8 a, u8 b)
{
    u8 x, y, r;
    bool carry;
    x = a;
    y = b;
    r = 0x00;
    while (x)
    {
        if (x & 0x01) { r ^= y; }
        carry = y & 0x80;
        y <<= 1;
        x >>= 1;
        if (carry) { y ^= 0x1b; }
    }
    return r;
}
```

The result accumulates in `r`. Each pass through the loop looks at the low-order bit of `x` with `x & 0x01` and if set, adds the current multiple of `b` (which is stored in `y`) to `r`. Afterwards, we shift `x` one position to the right to get the next bit. After each loop iteration, we shift `y` one position to the left, representing a multiplication by the polynomial X. If this overflows, we have to reduce `y` modulo $p(X)$ which has the binary representation `0x1b = 00011011`.

◇ The `C` language does not offer a way to check for carries except with an explicit variable (`carry = y & 0x80` checks if the high bit of `y` is set). In an assembler implementation, this could be handled much better by a branch-if-carry instruction using the processor's carry flag. The small number of constants and local variables involved would also suggest implementing the entire algorithm in the processor's registers.

6.11 Automorphisms of finite fields

The group of automorphisms of a finite field can be found with the following theorem:

> **Theorem 6.4.** The group of automorphisms of $GF(p^n)$ is isomorphic to the group $(\mathbb{Z}_n, +)$ and the Frobenius map $X \mapsto X^p$ is a generator of the automorphism group.

In a finite field represented as $GF(p)[X]/q(X)$, we can compute $f(X)$ for all the automorphisms by repeatedly applying the Frobenius map giving $X^p, X^{p^2}, X^{p^3}, \dots$ and reducing modulo $q(X)$.

In our case, $p = 2$ and we compute the powers of the Frobenius map for representations of $GF(2^8)$ modulo $p(X) = X^8 + X^4 + X^3 + X + 1$ and $q(Y) = Y^8 + Y^4 + Y^3 + Y^2 + 1$.

n	mod $p(X)$	mod $q(Y)$
0	X	Y
1	X^2	Y^2
2	X^4	Y^4
3	X^4+X^3+X+1	$Y^4+Y^3+Y^2+1$
4	$X^6+X^4+X^3+X^2+X$	$Y^6+Y^3+Y^2$
5	$X^7+X^6+X^5+X^2$	$Y^7+Y^4+Y^3+Y^2+1$
6	$X^6+X^3+X^2+1$	$Y^6+Y^4+Y^3+Y^2+Y+1$
7	$X^7+X^6+X^5+X^4+X^3+X$	Y^7+Y^2+1

6.12 Isomorphisms of $GF(2^8)$

Next, let's look for the isomorphisms of $GF(2^8)$ from the representation modulo $p(X) = X^8+X^4+X^3+X+1$ to another representation, for example $q(Y) = Y^8+Y^4+Y^3+Y^2+1$ (this is another irreducible polynomial; note the Y^2 in place of the X). That is, we're looking for an isomorphism f that maps field elements to other field elements such that $f(a \cdot b) = f(a) \odot f(b)$ where $\odot$ is multiplication modulo $q(Y)$ and $\cdot$ is multiplication modulo $p(X)$. Again, the value $f(X)$ will determine an isomorphism f from the representation modulo $p(X)$ to the representation modulo $q(Y)$.

If we find any one isomorphism from $GF(2)[X]/p(X)$ to $GF(2)[X]/q(Y)$ then we can get the whole set of isomorphisms by composing our one isomorphism with these automorphisms.

To get an isomorphism f, it is enough to find $f(X)$ which completely determines the isomorphism. To find this, we have to briefly work with p as a polynomial over $GF(2^8)$. So far, we have considered polynomials over $GF(2)$ to represent elements of $GF(2^8)$, i.e. our polynomials had coefficients in $GF(2)$. Now, we consider polynomials with coefficients in $GF(2^8)$. This can seem confusing at first because we will need two variable symbols: one to represent the "variable" of the polynomial and one to represent elements of $GF(2^8)$. For example, if a and b are elements of $GF(2^8)$ then $f(X) = aX + b$ is a polynomial over $GF(2^8)$ with coeffcients a, b. Suppose that $a = Y^2 + 1$ and $b = 2Y$, so we are using the letter Y to represent elements, then $f(X) = (Y^2 + 1)X + 2Y$. f is still a degree-one polynomial in one variable X; we just needed another variable symbol Y to write some of the coefficients which are elements of $GF(2^8)$.

If we pick any element a of $GF(2^8)$ represented modulo p, we have $p(a) = 0$. So for an isomorphism f from this representation to any other representation, we must have $f(p(a)) = f(0) = 0$. On the other hand, $f(p(a)) = p(f(a))$ since f is an isomorphism. Suppose we take $a = X$ (this X represents a field element) and $b = f(X)$, the value we are interested in. We have $p(b) = 0$ in the representation modulo the target of f, that is modulo $q(Y)$. This means that the polynomial $p(X)$ — now using X as a variable — has a zero at the value b we are interested in, which is the same as saying that $(X - b)$

divides $p(X)$ (modulo $q(Y)$). If we can find these zeros, for example by factoring $p(X)$, we get all the isomorphisms.

> **Theorem 6.5.** A function f is an isomorphism from $GF(z^n)$ represented modulo $p(X)$ to $GF(z^n)$ represented modulo $q(Y)$ if and only if f commutes with addition and multiplication, $f(1) = 1$ and for $b = f(X)$, $(X - b)$ divides $p(X)$ as polynomials modulo $q(Y)$.

One of the isomorphisms has $b = Y + 1$. We can check this by computing $p(X)/(X - (Y + 1))$ modulo $q(Y)$ and find

$$\begin{array}{lrl} (X^8 + X^4 + X^3 + X + 1) = (X - (Y + 1)) \cdot (& & \\ & 1 & X^7 \\ & +(Y + 1) & X^6 \\ & +(Y^2 + 1) & X^5 \\ & +(Y^3 + Y^2 + Y + 1) & X^4 \\ & +Y^4 & X^3 \\ & +(Y^5 + Y^4 + 1) & X^2 \\ & +(Y^6 + Y^4 + Y + 1) & X \\ & +(Y^7 + Y^6 + Y^5 + Y^4 + Y^2) & \\ ) \quad (\text{mod } q(Y)) & & \end{array}$$

If the mixture of X and Y variables is confusing, we can also represent elements of $GF(2)[Y]/(Y + 1)$ as two-digit hexadecimal numbers. In this case, $b = \texttt{0x03}$ and the above equation is

$$(X^8 + X^4 + X^3 + X + 1) = (X - \texttt{0x03})(X^7 + \texttt{0x03}\,X^6 + \texttt{0x05}\,X^5 + \texttt{0x0f}\,X^4 + \texttt{0x10}\,X^3 + \texttt{0x31}\,X^2 + \texttt{0x53}\,X + \texttt{0xf4}) \pmod{q(Y)}$$

6.13 Factoring polynomials in SAGE

Factoring polynomials to find isomorphisms, like factoring integers, is a job best delegated to computers. This is how one can use SAGE to factor polynomials.

```
F=GF(2)
R.<x>=F[x]
U.<y>=GF(2^8,modulus=x^8+x^4+x^3+x^2+1)
S.<x>=U[x]
S(x^8+x^4+x^3+x+1).factor()
```

```
(x + y + 1) * (x + y^2 + 1) * (x + y^4 + 1) *
(x + y^4 + y^3 + y^2) * (x + y^6 + y^3 + y^2 + 1) *
(x + y^6 + y^4 + y^3 + y^2 + y) * (x + y^7 + y^2) *
(x + y^7 + y^4 + y^3 + y^2)
```

In line 3 we set up the finite field U with a modulus of our choice; lines 1 and 2 prepare this (we can only use a custom modulus if we have bound the variable to the correct ring). Line 4 constructs S as the polynomial ring over our finite field, (re-)using the variable x. In line 5 we finally take the polynomial that we want to factor, inject it into the ring S and then call the factor operation. The factors are presented as field elements (of U) using the variable y.

So the 8 isomorphisms of $GF(2^8)$ from the representation modulo $p(X) = X^8 + X^4 + X^3 + X + 1$ to the representation modulo $q(Y) = Y^8 + Y^4 + Y^3 + Y^2 + 1$ are the functions $f_1, \ldots, f_8$ with

$$\begin{array}{ll}
f_1(X) = Y + 1 & f_2(X) = Y^2 + 1 \\
f_3(X) = Y^4 + 1 & f_4(X) = Y^4 + Y^3 + Y^2 \\
f_5(X) = Y^6 + Y^3 + Y^2 + 1 & f_6(X) = Y^6 + Y^4 + Y^3 + Y^2 + 1 \\
f_7(X) = Y^7 + Y^2 & f_8(X) = Y^7 + Y^4 + Y^3 + Y^2
\end{array}$$

Exercise. ($\star\star$) *Computation in $GF(2^8)$.* Let $p(X) = X^8 + X^4 + X^3 + X + 1$.

- Compute $(X^7 + X + 1)(X^6 + X^3 + X) + (X^7 + X^2 + 1)$ in $GF(2^8)$ using the representiation modulo $p(X)$.
- Solve the equation $X^3 + X + 1 = W \cdot (X^5 + 1)$ for W in $GF(2^8)$ represented modulo $p(X)$. Note: W is a polynomial, not an integer.
- Compute the powers of the Frobenius map in $GF(2^8)$ modulo $r(Z) = Z^8 + Z^7 + Z^2 + Z + 1$ (this is irreducible).
- Find an isomorphism from $GF(2)[Z]/r(Z)$ to $GF(2)[X]/p(X)$.

Exercise. ($\star$) *Computation in $GF(2^3)$.* Once you have mastered finite fields, you will be expected to solve exercises like this one for small enough fields almost as easily as arithmetic on integers. This exercise contains a lot of computations, each of which should be quick and easy.

We consider the field $GF(2^3)$ in the representations modulo the irreducible polynomials $p(X) = X^3 + X + 1$ and $q(Y) = Y^3 + Y^2 + 1$.

1. Reduce X^3 and X^4 modulo $p(X)$ and Y^3 and Y^4 modulo $q(Y)$.

2. How many elements does the group of automorphisms of $GF(2^3)$ have? What are these elements "called"?
3. Compute the Frobenius map for the representations modulo $p(X)$ and $q(Y)$. The quickest way to do this is to start with $\phi(X) = X^2$ and find $\phi(X^2)$, then express ϕ for an arbitrary field element as $\phi(aX^2 + bX + c) = uX^2 + vX + w$, i.e. find u, v, w in terms of a, b, c.
4. Do the same for all powers of the Frobenius map, in both representations (hint: there aren't too many.)
5. Find all the isomorphisms from the representation modulo $p(X)$ to the representation modulo $q(Y)$. Hint: how many are there? Find one isomorphism, then derive the others as follows: if f is one isomorphism and a is an automorphism of $GF(2^3)$ represented modulo $p(X)$, then $g = f \circ a$ is an isomorphism too. So compute $f(a(X))$ to get the value of $g(X)$.
6. ($\star\star$) Find the explicit multiplication formulas in both representations. That is, for $(aX^2 + bX + c)(dX^2 + eX + f) = (uX^2 + vX + w)$ find u, v, w in terms of a to f. Repeat the same for Y.

Lecture 7 — Vector spaces

Dr. D. Bernhard

In this lecture: vector spaces — linear independence and bases — linear maps — application to finite fields

Learning outcomes. After this lecture and revision, you should be able to:

- Define vector spaces.
- Multiply and (where possible) invert matrices, especially over finite fields.
- Interpret polynomial spaces over fields and finite fields as vector spaces and operations on them as linear maps.

7 Vector spaces

After the last lecture's very intense computation in finite fields, we close this section of the course with a slightly easier topic that we've actually been using implicitly already, namely vector spaces.

7.1 Definitions

The new thing about vector spaces is that we start with a given structure, a field, and build a vector space over this field. Vector spaces are groups in which you can add vectors but also "scale" vectors by multiplying them with field elements; in general you cannot multiply vectors with each other. The multiplication operation that takes a field element and a vector is sometimes called scalar multiplication.

> **Definition 7.1.** Start with any field $\mathbb{F}$. A vector space over $\mathbb{F}$ is a structure $(V, +, \cdot)$ where $+ : V \times V \to V$ and $\cdot : \mathbb{F} \times V \to V$ satisfying the following laws:
>
> 1. $(V, +)$ is an Abelian group.
> 2. Field and scalar multiplication associate: for any field elements f, g and any vector $\overline{a}$ we have $(fg) \cdot \overline{a} = f \cdot (g \cdot \overline{a})$. Here fg is multiplication in $\mathbb{F}$.

3. Field multiplication distributes over vector addition: for any vectors $\overline{a}, \overline{b} \in V$ and any field elements $f, g \in \mathbb{F}$ we have $f \cdot (\overline{a} + \overline{b}) = f \cdot \overline{a} + f \cdot \overline{b}$ and $(f +_{\mathbb{F}} g) \cdot \overline{a} = f \cdot \overline{a} + g \cdot \overline{a}$. We marked the field addition with $+_{\mathbb{F}}$ to distinguish it from vector addition here.

We use the convention that we write vectors with a line over them, e.g. $\overline{a}$ to distinguish them from field elements.

Examples. We have encountered many vector spaces already without mentioning it.

- Any field is automatically a vector space over itself; vector addition and scalar multiplication are just field addition and multiplication.
- The most common notion of a vector is a tuple or sequence of elements. For any field $\mathbb{F}$, the vector space $\mathbb{F}^n$ consists of vectors of length n with componentwise addition and the scalar multiplication $f \cdot (v_1, \ldots, v_n) := (f v_1, \ldots, f v_n)$.
- By the same logic, the polynomials over a field form a vector space, as do the polynomials over a field modulo some fixed polynomial. It is thus possible to interpret $GF(p^n)$ as the n-dimensional vector space $GF(p)^n$, "forgetting" about the multiplication of polynomials.

Linear (in)dependence and bases. Where a vector space is, a basis (plural: bases) is not far away. A basis plays a similar role to a set of generators of a group, but the additional field multiplication that turns a group into a vector space gives us much more to work with. Specifically, linear combinations:

Definition 7.2 (linear combination). For a finite set $\{\overline{v}_i\}_i$ of vectors, a linear combination is a sum $\sum_i c_i \cdot \overline{v}_i$ with coefficients c_i in the field $\mathbb{F}$.

If the index set is something like $I = \{1, 2, \ldots, n\}$ then we can write a linear combination as $c_1 \cdot \overline{v}_1 + \ldots + c_n \cdot \overline{v}_n$.

◇ The basic definitions of linear algebra (linear independence, basis etc.) can also be defined for infinite sets, but the exact definition is a bit subtle. We will not need to worry about this too much in this course. A linear combination for an infinite set V of vectors is a sum where only a finite number of coefficients are non-zero.

A linear combination of vectors where all coefficients are zero (the neutral element of field addition) is automatically the zero vector (the neutral element of vector addition). A set of vectors is linearly independent if this is the only linear combination that is zero; another way of saying this is that no vector in the set can be written as a linear combination of the others.

> **Definition 7.3 (linear (in)dependence).** A set $\{\overline{v}_i\}_i$ of vectors is linearly independent if no linear combination of the vectors $\sum_i c_i \cdot \overline{v}_i$ with coefficients in $\mathbb{F}$ gives the zero vector (neutral element of vector addition), unless all coefficients are already zero (the neutral element of the field's addition). A set of vectors that is not linearly independent is called linearly dependent.

◇ An infinite set V is linearly independent if no finite sum of elements in V with coefficients in $\mathbb{F}$ gives the zero vector, unless all coefficients are zero. This is equivalent to saying that every finite subset of V is linearly independent.

And finally, a basis is a finite set of linearly independent vectors (in a particular order) that generates the entire space.

> **Definition 7.4 (basis).** A basis of a vector space V is a finite list $(\overline{v}_1, \ldots, \overline{v}_n)$ of vectors that is linearly independent and generates V, i.e. $V = \langle \overline{v}_1, \ldots, \overline{v}_n \rangle$.

In other words, every vector $\overline{w}$ in the space can be written as a linear combination of the basis vectors: $w = w_1\overline{v}_1 + \ldots + w_n\overline{v}_n$. In fact, if a vector space has a basis then any two bases of the space have the same number of elements (which we call the dimension of the space) and for any vector v and any basis in a fixed order, there is exactly one way (one tuple of coefficients) to write W as a linear combination of the basis vectors.

◇ An infinite set W of vectors is a basis of a vector space V if (1) it is linearly independent — that is, every finite subset of W is linearly independent in the usual sense and (2) every element in $v \in V$ can be written as a *finite* linear combination of elements in W. This definition is required to make the theorem "every vector space has a basis" true even in the infinite-dimensional case, assuming the Axiom of Choice.

> **Proposition 7.5.** Any two bases of a vector space have the same number of elements. If a vector space has a basis with n elements, we say that the space has dimension n.

And finally, every vector space has a basis. This proposition is only really mathematically interesting to discuss in the infinite case but it is the start of most constructions in linear algebra: given any vector space V, we can simply assume that a basis B is given as well.

> **Proposition 7.6.** Every vector space has a basis.

Linear maps. A vector space homomorphism is a function $f : V \to W$ between two vector spaces over the same field $\mathbb{F}$ that preserves vector addition and scalar multiplication. We call such a function a linear map.

> **Definition 7.7 (linear).** If V and W are two vector spaces over a field $\mathbb{F}$, we call a function $f : V \to W$ linear if for any $\overline{x}, \overline{y}$ in V and any $a \in \mathbb{F}$ we have
>
> - $f(\overline{v} + \overline{w}) = f(\overline{v}) + f(\overline{w})$.
> - $f(a \cdot \overline{v}) = a \cdot f(\overline{v})$.

The reader should understand by now which operation symbols refer to V-operations and which ones refer to W-operations.

If V is a vector space with basis $B = (\overline{b}_1, \dots, \overline{b}_n)$ then a linear map $f : V \to W$ can be computed on any vector from its values on the basis alone. Namely, if you know $f(\overline{b}_1), \dots, f(\overline{b}_n)$ and want to compute $f(\overline{v})$ then you can write $\overline{v}$ in exactly one way as $v = c_1 \cdot \overline{b}_1 + \dots + c_n \cdot \overline{b}_n$, giving $f(\overline{v}) = c_1 \cdot f(\overline{b}_1) + \dots + c_n \cdot f(\overline{b}_n)$.

If we have a basis $P = (\overline{p}_1, \dots, \overline{p}_m)$ of W as well, you can compute the coefficients of the images of the basis elements under f: there are unique coefficients $(a_{1,1}, \dots, a_{1,m})$ such that $f(\overline{b}_1) = a_{1,1} \cdot \overline{p}_1 + \dots + a_{1,m} \cdot \overline{p}_m$ and the same for the other basis elements. In other words, a linear map between a n-dimensional vector space V and a m-dimensional vector space W can be specified as a $n \cdot m$ rectangle of coefficients in the field $\mathbb{F}$:

$$f : V \to W \quad \leftrightarrow \quad \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{pmatrix}$$

We call such a rectangle of coefficients a matrix. Matrices form a vector space which we write $\mathbb{F}^{m \times n}$ for the space of matrices with m rows and n columns as in the example above. Matrix addition is component-wise; scalar multiplication with a field element just multiplies all matrix components with the field element.

Applying a linear map to a vector becomes matrix-vector multiplication. If we consider linear maps from a space to itself, we can compose them: for $f, g : V \to V$ we can from the map fg that takes $\overline{v}$ to $f(g(\overline{v}))$. If both f and g are linear, so is fg (exercise). We can use this to define a multiplication operation on square matrices, turning $\mathbb{F}^{n \times n}$ into a ring (for every positive integer n) — this is just the usual matrix multiplication. Of course we can define matrix multiplication between compatible non-square matrices too but we don't get a ring that way.

7.2 Polynomial spaces as vector spaces

We look at the space $V = GF(p)^n$ constructed by taking a finite field $GF(p^n) = GF(p)[X]/q(X)$ modulo an irreducible polynomial q of degree $n \geq 1$ and interpreting it as a vector space. The elements of this space are of the form $(c_0, c_1, \ldots, c_{n-1})$ which could be written as $c_0 + c_1X + \ldots + c_{n-1}X^{n-1}$. Interpreting it as a vector space means forgetting about multiplication of $GF(p^n)$ elements but adding multiplication with $GF(p)$ elements: $a \cdot (c_0, \ldots, c_{n-1}) := (a \cdot c_0, \ldots, a \cdot c_{n-1})$.

One basis of this vector space consists of vectors $\overline{b}_i$ for $i = 0$ to $n - 1$ where $\overline{b}_i$ is 1 at position r and 0 elsewhere. Written as polynomials, the i-th basis vector $\overline{b}_i$ is the monomial X^i.

The map $GF(p) \to GF(p^n), a \mapsto (a, 0, \ldots, 0)$ is a field homomorphism. It is sometimes called the embedding of the base field $GF(p)$ into the extension field $GF(p^n)$. This map commutes with field multiplication in the following way: for any element a of $GF(p)$ and any element $\overline{v}$ of $GF(p^n)$, you get the same if you perform the scalar multiplication $a \cdot \overline{v}$ or if you embed a in $GF(p^n)$ and then do field multiplication there. We can express this in a diagram.

$$\begin{tikzcd}
\begin{array}{c} a \in GF(p) \\ \overline{v} \in GF(p^n) \end{array} \arrow[r, "\text{embedding}"] \arrow[d, "\begin{array}{l}\text{scalar}\\ \text{mult.}\end{array}"'] & \begin{array}{l} (a, 0, \ldots, 0) \in GF(p^n) \\ \overline{v} = (v_0, \ldots, v_{n-1}) \in GF(p^n) \end{array} \arrow[dl, "\begin{array}{l}GF(p^n)\\ \text{mult.}\end{array}"] \\
(av_0, \ldots av_{n-1}) &
\end{tikzcd}$$

7.3 Automorphisms revisited

An automorphism f of $GF(p^n)$ can be represented as a $n \times n$ matrix, since such a f must be linear over the field $\mathbb{F}$: if $a \in \mathbb{F}$ and $\overline{v} \in V$ then $f(a\overline{v}) = a \cdot f(\overline{v})$. However, we know that field automorphisms cannot change degree-0 polynomials: $f(1, 0, \ldots, 0) = (1, 0, \ldots, 0)$. Writing f out as a matrix, we see

$$\begin{pmatrix} f_{0,0} & f_{0,1} & \cdots & f_{0,n-1} \\ f_{1,0} & f_{1,1} & \cdots & f_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n-1,0} & f_{n-1,1} & \cdots & f_{n-1,n-1} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} f_{0,0} \\ f_{1,0} \\ \vdots \\ f_{n-1,0} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

so the zeroth column of the matrix of a must start with a 1 and be zero everywhere else. Since we also know how to compute $f(X^2)$ from $f(X)$, this means we know how to compute $f(0,0,1,0,\ldots,0)$ from $f(0,1,0,\ldots,0)$ and so on — so all the information about f is contained in the first column of the matrix of f and we can always compute the other columns from it.

We look at two examples. The first is $GF(7^2)$ where for $p(X) = X^2+X+6$. we found two automorphisms id, ϕ with $id(X) = X$ and $\phi(X) = 6+6X$ (the Frobenius map). As matrices, these are

$$id = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \phi = \begin{pmatrix} 1 & 6 \\ 0 & 6 \end{pmatrix}$$

It is now obvious how to calculate ϕ on an arbitrary field element $(a+bX)$:

$$\phi(a+bX) = \begin{pmatrix} 1 & 6 \\ 0 & 6 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a+6b \\ 6b \end{pmatrix}$$

If we introduce another irreducible polynomial $q(Y) = Y^2+1$, the isomorphisms we found between these representations last time were $f_1(a,b) = (a+3b, 2b)$ and $f_2(a,b) = (a+3b, 5b)$. If we know f_1, we can calculate f_2 by multiplying from the right with the Frobenius map:

$$f_1\phi = \begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 6 \\ 0 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 0 & 5 \end{pmatrix} = f_2$$

Matrices give us an easy way to find the inverses of isomorphisms, that is the isomorphisms of $GF(7^2)$ from the representation modulo $q(Y)$ back to the representation modulo $p(X)$. All we need to do is invert the matrices of f_1, f_2:

$$\begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix}^{(-1)} = \begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix} \qquad \begin{pmatrix} 1 & 3 \\ 0 & 5 \end{pmatrix}^{(-1)} = \begin{pmatrix} 1 & 5 \\ 0 & 3 \end{pmatrix}$$

Giving $f_1^{(-1)}(a+bX) = (a+2b)+4bX$ and $f_2^{(-1)}(a+bX) = (a+5b)+3bX$.

Our second example is $GF(2^3)$, this time with the irreducible polynomial $p(X) = X^3+X+1$. The Frobenius map sends $X \mapsto X^2$ and $X^2 \mapsto [X^4] = X^2+X$. As a matrix, we get

$$\phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \qquad \phi \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} a \\ c \\ b+c \end{pmatrix}$$

from which we read off $\phi(a+bX+cX^2) = a+cX+(b+c)X^2$. Finding the other automorphisms is easy too:

$$\phi^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

which maps $X \mapsto X + X^2$ (middle column) and $X^2 \mapsto X$ (right column). Multiplying with the column vector $(a; b; c)$ we get $\phi^2(a + bX + cX^2) = a + (b + c)X + bX^2$. If we look at the third power

$$\phi \cdot \phi^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

we get the identity map back, as expected. If we take the isomorphism $f(X) = Y^2 + 1$ into the representation modulo $q(Y) = Y^3 + Y^2 + 1$, we find $f(X^2) = Y^2 + Y$. The other isomorphisms are

$$f_1\phi = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} = f_2$$

$$f_1\phi^2 = f_3$$

from which we read off the columns $f_2(X) = Y^2 + Y$ and $f_2(X^2) = 1 + Y$, giving $f_2(a + bX + cX^2) = (a + c) + (b + c)Y + bY^2$. We could read this last formula off the rows of the matrix directly, since we get the last formula by multiplying the matrix of f_2 with the column vector $(a; b; c)$. To invert the isomorphisms, one again just needs to invert the matrices.

Exercise. *The rest of the example.*

- ($\star$) Compute $f_3 = f_1\phi^2$, then write out the expressions for $f_3(X)$, $f_3(X^2)$ and $f_3(a + bX + cX^2)$.
- ($\star\star$) Invert the matrix for f_1 to get the inverse isomorphism. Note: matrix inversion is a lot easier in $GF(2)$ as $1 + 1 = 0$ so you never need to multiply rows through to cancel constants!

Exercise. ($\star$) *More finite fields.* Consider the field $GF(3^3)$ with the irreducible polynomials $p(X) = X^3 + 2X + 1$ and $q(Y) = Y^3 + 2Y^2 + Y + 1$.

1. Find the Frobenius map ϕ as a 3×3 matrix modulo $p(X)$.
2. Find the powers of the Frobenius map.
3. One isomorphism f between the representations $p(X)$ and $q(Y)$ has $f(X) = 2X^2 + 2X$. Find the matrix of f.
4. Find the inverse of f by inverting the matrix of f.
5. How many isomorphisms are there between the two representations?

6. Find the other isomorphisms by matrix multiplication using the matrix of f and the Frobenius map ϕ.
7. ($\star\star$) Here is another way to find the Frobenius map in the representation modulo $q(Y)$. We have the following situation with $V = GF(3)[X]/p(X)$ and $W = GF(3)[Y]/q(Y)$:

$$\begin{tikzcd}
V \arrow[r, shift left, "f"] \arrow[d, "\phi"'] & W \arrow[l, shift left, "f^{(-1)}"] \arrow[d, "\hat{\phi}"] \\
V \arrow[r, shift left, "f"] & W \arrow[l, shift left, "f^{(-1)}"]
\end{tikzcd}$$

   From this we see that the Frobenius map in W has matrix $\hat{\phi} = f \cdot \phi \cdot f^{(-1)}$. Compute $\hat{\phi}$ this way.

◇ The conditions for field automorphisms say that $f(\overline{v} + \overline{w}) = f(\overline{v}) + f(\overline{w})$, $f(\overline{v} \cdot \overline{w}) = f(\overline{v}) \cdot f(\overline{w})$ and $f(1, 0, \ldots, 0) = (1, 0, \ldots, 0)$ since this element is the one of the field. Since the scalar multiplication $a \cdot \overline{v}$ we get for V as a vector space is equivalent to the vector multiplication $(a, 0, \ldots, 0) \cdot \overline{v}$, a field automorphism must satisfy $f(a \cdot \overline{v}) = f((a, 0, \ldots, 0) \cdot \overline{v}) = f(a, 0, \ldots, 0) \cdot f(\overline{v}) = (a, 0, \ldots, 0) \cdot f(\overline{v}) = a \cdot f(\overline{v})$. This explains why in field multiplication we get $f(a \cdot \overline{v}) = a \cdot f(\overline{v})$ with the a appearing "outside f" whereas the automorphism rule says $f(\overline{v} \cdot \overline{w}) = f(\overline{v}) \cdot f(\overline{w})$.