Queen Mary
University of London

MTH5100 **Algebraic structures I**

Notes 1 **Spring 2011**

1 Introduction

Lecture 1

This is a course in *abstract* algebra. You have seen some abstract algebra already in the first year course Introduction to Algebra. For example you have seen definitions of structures such as rings, groups, fields. In this course we take this further, and I hope to demonstrate the power and economy of the abstract approach.

Abstraction is introduced into mathematics not to make it hard or inaccessible, but in order to make it useful. As E. T. Bell (a famous historian of mathematics) said "Abstractness, sometimes hurled as a reproach at mathematics, is its chief glory, and its surest title to practical usefulness." Among the important uses of abstraction are

(a) Precision. Once everything is precisely defined, there can be no doubt about what we are talking about.

(b) Generality. When it turns out that lots of apparently different things obey the same mathematical laws, they can be studied simultaneously, thus saving a lot of effort.

This saving of effort is behind the following remark of Matthew Pordage: "One of the endearing things about mathematicians is the extent to which they will go to avoid doing any real work."

The main topic of this course is *rings*. I will remind you of the formal definition in due course. To begin with, just recall the extraordinary variety of mathematical objects which are rings:

(a) Number systems, like the real numbers $\mathbb{R}$, the integers $\mathbb{Z}$, the rational numbers $\mathbb{Q}$, the complex numbers $\mathbb{C}$ (but not the natural numbers $\mathbb{N}$: why not?). The operations which make these things into rings are the arithmetical operations $+$, $-$, $\times$, and the special elements 0 and 1.

(b) Modular arithmetic makes $\mathbb{Z}_n$ into a ring.

(c) Square matrices, with matrix multiplication. The entries can be real numbers, rational numbers, integers—indeed, they can come from any ring at all, even from another matrix ring.

(d) Polynomials. The coefficients can be real numbers, integers, ...: again, they can come from any ring, even a matrix ring, or another polynomial ring. We will need to be quite careful in defining polynomial rings properly, though—they are not as straightforward as you might hope.

The examples of (a) a polynomial ring whose coefficients are matrices, and (b) a matrix ring whose coefficients are polynomials, shows that two rings can look different but actually be mathematically “the same”: this concept is known as *isomorphism*, which will play an important part in this course.

2 Revision of basic mathematical concepts

I will begin by reminding you of those parts of Introduction to Algebra which are most important for this course. All this needs to be at your fingertips before you can proceed. If it is not, then revise it, perhaps from your Introduction to Algebra notes, urgently.

Sets. Set notation. Two sets are equal if (and only if) they have the same elements. Subsets, intersection $A \cap B$, union $A \cup B$, difference $A \setminus B$ and symmetric difference $A \triangle B$. Rules like $(A \cap B) \cap C = A \cap (B \cap C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Cartesian product: $A \times B = \{(a,b) \mid a \in A, b \in B\}$. The ‘ordered pair’ (a,b) will not be defined more formally here, though it can be if needed.

Functions. The word function is used in slightly different senses in different parts of mathematics (like the word ‘polynomial’), and it is therefore important that we are precise about the sense we use. Formally, a function $f : A \to B$ is a subset $\mathcal{F}$ of the Cartesian product $A \times B$, with the property that for every $a \in A$, there is exactly one $b \in B$ such that $(a,b) \in \mathcal{F}$. For example, instead of defining a function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$, we define it as the set $\{(x,x^2) \mid x \in \mathbb{R}\}$, which in other parts of mathematics might be called the *graph* of the function.

Operations. This word is used to describe special types of functions. An operation like ‘plus’ can be thought of as a function from ordered pairs of (say) real numbers, to real numbers, that is

$$\text{plus} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}; (x,y) \mapsto x+y.$$

With our formal definition of a function as a set of ordered pairs, this looks even worse:

$$\text{plus} = \{((x,y),z) \mid x,y,z \in \mathbb{R}, x+y = z\}.$$

Don’t worry, however, we are not going to be using this cumbersome notation very often.

Formally, a *binary operation* on a set S is just a function $f : S \times S \to S$.

Examples of different types of functions: a function $f : A \to B$ is Lecture 2

- *surjective* if every $b \in B$ is of the form $f(a)$ for some $a \in A$;
- *injective* if no $b \in B$ is of the form $f(a_1) = f(a_2)$ for two different elements $a_1, a_2 \in A$;
- *bijective* if both of these conditions hold: i.e. every $b \in B$ is of the form $f(a)$ for a unique element $a \in A$.

Examples of different types of operations: a binary operation $*$ on a set A is

- *commutative* if $a * b = b * a$ for all $a, b \in A$;
- *associative* if $(a * b) * c = a * (b * c)$ for all $a, b, c \in A$; for example, the operation $+$ on $\mathbb{R}$ is associative, but the operation $-$ on $\mathbb{R}$ is not.

Relations. We will not need much on the formal definitions of relations, apart from the important instance of *equivalence relations* (see below). Formally, a *binary relation* on a set A is any subset of $A \times A$. In particular, a function $f : A \to A$ is a special type of relation. Informally, a relation $R \subseteq A \times A$ is more usually written in *infix notation*, that is, instead of writing $(a, b) \in R$ we write aRb, and say 'a is related to b (by R)'.

Examples of different types of relations: to say that a binary relation R on a set A is

- *reflexive* means that: aRa for all elements $a \in A$;
- *symmetric* means that: if aRb, then also bRa;
- *transitive* means that: if aRb and bRc, then also aRc;
- *anti-symmetric* means that: if aRb and bRa, then $a = b$.

To remember which is which, remember that the alphabetical order r, s, t, corresponds to the order 1, 2, 3 of the number of elements of A mentioned in the definition.

Examples of anti-symmetric relations are $\leq$ on $\mathbb{R}$, and $\subseteq$ on $\mathcal{P}(X)$. Also $<$ on $\mathbb{R}$ is anti-symmetric: to prove this we merely need to show that *if* $a < b$ and $b < a$ then something-or-other happens—but the hypothesis ($a < b$ and $b < a$) can never be true and therefore we don't have to prove anything.

3 Equivalence relations

Lecture 3

An *equivalence relation* R is one which is reflexive, symmetric and transitive. If R is an equivalence relation on A, and $a \in A$, then the *equivalence class of a* is the set of elements which are related to a, that is $\{b \in A \mid aRb\}$, written either $R(a)$ or $[a]_R$ or $[a]$, or some other similar notation. The important fact about equivalence relations is that these equivalence classes form a *partition* of A, which means that they (a) are non-empty, since $a \in R(a)$, (b) cover the whole set A, since $\bigcup_{a\in A} R(a) \supseteq \bigcup_{a\in A}\{a\} = A$ implies $\bigcup_{a\in A} R(a) = A$, and (c) do not overlap, that is, if $R(a) \cap R(b) \neq \emptyset$ then in fact $R(a) = R(b)$.

To prove this last statement, suppose $c \in R(a) \cap R(b)$, so that cRa and cRb, so bRc, so bRa. Now if $d \in R(b)$, then bRd, so dRb, so dRa, so aRd, that is $d \in R(a)$. This shows that $R(b) \subseteq R(a)$, and a similar argument gives $R(a) \subseteq R(b)$, and therefore $R(a) = R(b)$ as required.

Conversely, given a partition of A, we can define a relation R on A by saying aRb just when a and b are both in the same part of the partition. It is easy to verify that R is an equivalence relation, and that the equivalence classes are just the parts of the original partition.

Examples of equivalence relations. The easiest example is the relation of equality, on any set A. This is reflexive, since $a = a$ for every $a \in A$. It is symmetric, since *if* $a = b$ then $b = a$. And it is transitive, since *if* $a = b$ and $b = c$, then $a = c$.

The most general example has already been given above: take any partition of A, and construct the equivalence relation corresponding to it. Another way of looking at this is to take any function $f : A \to B$, and define a relation by aRb just when $f(a) = f(b)$. It is easy to check that this is an equivalence relation.

Modular arithmetic. Another example you have already seen is the set of 'integers modulo n', where n is a positive integer ($n > 1$ for some applications). Define a relation R on $\mathbb{Z}$ by aRb whenever $a - b$ is divisible by n, that is $a - b = cn$ for some integer c (depending on a and b, of course). This relation R is (a) reflexive, since $a - a = 0 = 0n$ for every $a \in \mathbb{Z}$, (b) symmetric, since if $a - b = cn$ then $b - a = (-c)n$, and $-c \in \mathbb{Z}$ since $c \in \mathbb{Z}$, and (c) transitive, since if $a - b = cn$ and $b - d = en$ then $a - d = (c+e)n$, with $c + e \in \mathbb{Z}$ because $c, e \in \mathbb{Z}$. It is easy to see that there are just n equivalence classes, namely [0], [1],..., $[n-1]$. These are sometimes written $[0]_n$, $[1]_n$, ..., $[n-1]_n$ for clarity.

The important thing about this construction, however, is that we can do *arithmetic* with these equivalence classes. Notice that if $x \in [a]$ and $y \in [b]$, then $x = a + cn$ and $y = b + dn$ for some integers c, d, and therefore $x + y = (a + b) + (c + d)n \in [a + b]$. This means that it makes sense to *define* $[a] + [b] = [a + b]$, because if you add any element of $[a]$ and any element of $[b]$ you always get an element of $[a + b]$. Similarly you can calculate that $xy = ab + (bc + ad + cdn)n \in [ab]$, so we can define $[a].[b] = [ab]$

without risk of confusion.

This is an example of a construction of a *quotient ring*, which is a very important concept in this course, and in mathematics generally. This type of quotient is often found difficult on first meeting. However, if you really understand how modular arithmetic works, then you should not have any difficulty with quotient rings, which are just a more abstract and general form of the same thing.

4 Rings

Lecture 4

Definition. A *ring* is a set R with two binary operations $+$ and $.$, and a special element 0, such that:
Rules for addition:

(A0) (closure: not strictly necessary) $a+b \in R$ whenever $a,b \in R$;

(A1) (associativity) $(a+b)+c = a+(b+c)$ whenever $a,b,c \in R$;

(A2) (zero) $a+0 = 0+a = a$ whenever $a \in R$;

(A3) (negatives, or additive inverses) for every $a \in R$, there exists a $b \in R$ such that $a+b = b+a = 0$;

(A4) (commutativity) $a+b = b+a$ whenever $a,b \in R$;

Rules for multiplication:

(M0) (closure: not strictly necessary) $a.b \in R$ whenever $a,b \in R$;

(M1) (associativity) $(a.b).c = a.(b.c)$ whenever $a,b,c \in R$;

Rules involving both:

(D) (distributivity) $a.(b+c) = (a.b)+(a.c)$ and $(b+c).a = (b.a)+(c.a)$ whenever $a,b,c \in R$.

Notice that we have not assumed that there is only one zero, or only one negative of an element. We can prove these facts from the definition:

Uniqueness of zero. Suppose that 0_1 and 0_2 are two different zeros, so that $a+0_1 = 0_1+a = a+0_2 = 0_2+a = a$ for all a. Now substituting $a = 0_1$ into $a+0_2 = a$ gives $0_1+0_2 = 0_1$, while substituting $a = 0_2$ into $0_1+a = a$ gives $0_1+0_2 = 0_2$. Hence $0_1 = 0_2$.

Uniqueness of negatives. Now suppose that a has two negatives, say b and c. Then we have $a+b = b+a = a+c = c+a = 0$ and therefore $b+(a+c) = b+0 = b$ while $(b+a)+c = 0+c = c$ so by the associative law, $b = c$.

Special types of rings. A ring may or may not satisfy the analogues of (A2), (A3), (A4) for multiplication:

(M2) (one) there is an element $1 \in R$ (with $1 \neq 0$) such that $a.1 = 1.a = a$ for all $a \in R$;

(M3) (inverses) for all $a \in R$ except for $a = 0$, there is an element $b \in R$ such that $a.b = b.a = 1$;

(M4) (commutativity) $a.b = b.a$ for all $a, b \in R$.

A ring which satisfies (M2) is called a *ring with one*. A ring which satisfies (M2) and (M3) is called a *division ring*. A ring which satisfies (M4) is called a *commutative ring*. A ring which satisfies all three of (M2), (M3) and (M4) is called a *field*.

Cancellation laws. If $x + y = x + z$ for some elements $x, y, z \in R$, we want to deduce that $y = z$. This follows by the law of negatives: there is an element t such that $t + x = x + t = 0$, so $t + (x + y) = (t + x) + y = y$ and $t + (x + z) = (t + x) + z = z$, and since $t + (x + y) = t + (x + z)$ we have $y = z$.

Multiplication by zero. We want to show that $0.a = a.0 = 0$ for all a. To do this, expand $0 = 0 + 0$ and use the distributive law: $0.a = (0 + 0).a = 0.a + 0.a$ and also $0.a = 0 + 0.a$, so $0 + 0.a = 0.a + 0.a$ and by the cancellation law, $0 = 0.a$. A similar argument gives $a.0 = 0$ for all $a \in R$.

Lecture 5

Negation and subtraction. Unfortunately, a minus sign has two different meanings in ordinary arithmetic, which get carried over to rings in general. The first is the *unary minus*, whereby $-a$ means the unique element of R such that $a + (-a) = (-a) + a = 0$, that is, $-a$ is the *negative* of a. The second is the *binary minus*, whereby $a - b$ is a shorthand notation for $a + (-b)$. We will prove the important properties of the unary minus, from which you can deduce the important properties of the binary minus by using the axioms for addition (A0)–(A4).

Firstly, $(a + b) + ((-b) + (-a)) = ((a + b) + (-b)) + (-a) = (a + (b + (-b))) + (-a) = (a + 0) + (-a) = a + (-a) = 0$, so $(-b) + (-a)$ (or $(-a) + (-b)$) is the negative of $a + b$. In other words $-(a + b) = (-a) + (-b)$.

Next, by the distributive law, $(ab) + (-a)b = (a + (-a))b = 0.b = 0$, so $-(ab) = (-a)b$ and similarly $-(ab) = a.(-b)$. Moreover, $(-a)(-b) + (-a)b = (-a)((-b) + b) = (-a).0 = 0$ so $(-a)(-b)$ is the negative of $(-a)b = -(ab)$. But ab is a negative of $-(ab)$, so $(-a)(-b) = ab$.

Addition of several elements. By repeated use of the associative law of addition we can show that a sum of n elements, that is $a_1 + a_2 + \cdots + a_n = \sum_{i=1}^{n} a_i$ is well-defined, that is, does not depend on the bracketing used to calculate it. The associative law itself is the case $n = 3$ of this. To prove it by induction on n, we can assume it already holds for all sums of at most $n-1$ elements, and it remains to show that

$$(a_1 + \cdots + a_i) + (a_{i+1} + \cdots + a_n) = (a_1 + \cdots + a_j) + (a_{j+1} + \cdots + a_n)$$

where we might as well suppose that $i < j$. We have

$$\begin{aligned}(a_1 + \cdots + a_j) + (a_{j+1} + \cdots + a_n) &= ((a_1 + \cdots + a_i) + (a_{i+1} + \cdots + a_j)) + (a_{j+1} + \cdots + a_n) \\ &= (a_1 + \cdots + a_i) + ((a_{i+1} + \cdots + a_j) + (a_{j+1} + \cdots + a_n)) \\ &= (a_1 + \cdots + a_i) + (a_{i+1} + \cdots + a_n).\end{aligned}$$

A similar inductive argument using the commutative law for addition shows that the *ordering* of the terms $a_1, \ldots, a_n$ does not matter either.

Another consequence (which I did not mention in the lecture, but which is still important) is that we get extended versions of the distributive laws: $(a_1 + \cdots + a_n).b = (a_1.b) + \cdots + (a_n.b)$ and similarly $b.(a_1 + \cdots + a_n) = (b.a_1) + \cdots + (b.a_n)$.

5 Some important examples of rings

Matrix rings. If R is any ring, we can define the ring of $n \times n$ matrices over R, denoted $M_n(R)$, as follows. First we add two matrices by adding the corresponding entries, that is $(A_{ij}) + (B_{ij}) = (A_{ij} + B_{ij})$, where we use (A_{ij}) to denote the matrix whose (i, j) entry is A_{ij}. Then we multiply two matrices by the rule $(A_{ij}).(B_{ij}) = (\sum_{k=1}^{n} A_{ik}.B_{kj})$.

We need to check that $M_n(R)$ satisfies the axioms for a ring. Most of them follow easily from the corresponding axioms for the ring R, with the zero matrix being the one in which every entry is 0. But there is one which is more difficult, namely the associative law for multiplication.

$$\begin{aligned}((A_{ij}).(B_{ij})).C_{ij} &= (\sum_{k=1}^{n}(\sum_{l=1}^{n} A_{il}.B_{lk}).C_{kj}) \\ &= (\sum_{k=1}^{n}\sum_{l=1}^{n}((A_{il}.B_{lk}).C_{kj}) \\ &= (\sum_{l=1}^{n}\sum_{k=1}^{n}(A_{il}.(B_{lk}.C_{kj})) \\ &= (\sum_{l=1}^{n} A_{il}.\sum_{k=1}^{n}(B_{lk}.C_{kj})) \\ &= (A_{ij}).((B_{ij}).(C_{ij})).\end{aligned}$$

In the first line of this argument, we just use the definition of matrix multiplication, twice. In the second, we use the (extended) distributive law to bring C_{kj} inside the summation. In the third line we re-order the sum using the extended commutative law, as well as using the associative law of multiplication. Finally we reverse the argument, using the distributive law again.

It is easy to see that if R has a one, then so does $M_n(R)$: the one is the *identity matrix*, which has the one of R in each diagonal entry, and zero elsewhere. On the other hand, if $n \geq 2$ there are non-zero matrices which do not have inverses, and you can easily find matrices A and B such that $AB \neq BA$. Lecture 6

Polynomials. A polynomial over a ring R is a formal expression $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$, also written $\sum_{i=0}^{n} a_ix^i$. However, we have to be careful about when two such expressions are equal. Essentially, we can add on any number of zero terms without changing the polynomial. That is $\sum_{i=0}^{n} a_ix^i = \sum_{i=0}^{m} b_ix^i$ (for $m \geq n$) if $a_i = b_i$ for all $0 \leq i \leq n$ and $b_i = 0$ for all $n < i \leq m$. For convenience, when talking about a polynomial $\sum_{i=0}^{n} a_ix^i$, we will often talk about coefficients a_i with $i > n$, with the understanding that these are to be taken to be 0.

With this convention we can define addition by

$$(\sum_{i=0}^{n} a_ix^i) + (\sum_{i=0}^{m} b_ix^i) = \sum_{i=0}^{\max\{m,n\}} (a_i + b_i)x^i.$$

It is easy to check all the addition axioms (A0)-(A4) for a ring: they follow directly from the corresponding axioms for R itself.

We define multiplication in the way you know, using the law of exponents $x^i.x^j = x^{i+j}$ and collecting together terms with the same power of x. That is

$$(\sum_{i=0}^{n} a_ix^i).(\sum_{j=0}^{m} b_jx^j) = \sum_{k=0}^{m+n} (\sum_{i+j=k} a_ib_j)x^k.$$

Now we can rewite the coefficient of x^k on the right-hand side as

$$\sum_{i+j=k} a_ib_j = \sum_{i=0}^{k} a_ib_{k-i}$$

if we wish. The distributive laws are easy to check, but again the associative law of multiplication is the tricky one.

$$\begin{aligned}((\sum_{i=0}^{n} a_ix^i).(\sum_{j=0}^{m} b_jx^j)).(\sum_{k=0}^{p} c_kx^k) &= (\sum_{r=0}^{m+n} (\sum_{i+j=r} a_ib_j)x^r).(\sum_{k=0}^{p} c_kx^k) \\ &= \sum_{s=0}^{m+n+p} (\sum_{r+k=s} (\sum_{i+j=r} a_ib_j)c_k)x^s \\ &= \sum_{s=0}^{m+n+p} (\sum_{i+j+k=s} (a_ib_j)c_k)x^s\end{aligned}$$

using the extended distributive law in R to re-arrange the sum of products for each coefficient. Now we use the associative law in R to re-write the individual terms $(a_ib_j)c_k = a_i(b_jc_k)$, and a similar argument to the above to show that this is the same as

$$(\sum_{i=0}^{n} a_ix^i).((\sum_{j=0}^{m} b_jx^j).(\sum_{k=0}^{p} c_kx^k)).$$

Lecture 7

Rings of functions. Functions as used in calculus usually form rings under pointwise addition and multiplication. That is, if f and g are functions, say from $\mathbb{R}$ to $\mathbb{R}$, we define $f+g$ as the function that maps x to $f(x)+g(x)$, written $(f+g)(x) = f(x)+g(x)$; and similarly $f.g$ is defined by $(f.g)(x) = f(x).g(x)$. The ring axioms are easy to check, and come directly from the ring axioms for $\mathbb{R}$. The zero function is the function defined by $\text{zero}(x) = 0$, and the negative of f is the function $-f$ defined by $(-f)(x) = -(f(x))$.

More generally, if R is any ring, and X is any set, then we can make the set of functions from X to R into a ring using the same rules.

Rings of sets. If we try to make $\mathcal{P}(X)$, the power set of X, into a ring, we might try to use $\cup$ as addition and $\cap$ as multiplication. But this does not work. (Exercise: which axioms fail?)

Instead, we define $A+B = A\triangle B$ and $A.B = A\cap B$. Then most of the axioms are easy to check. The zero is $\emptyset$, and the negative of A is A itself, since $A\triangle A = \emptyset$. The most difficult one is associativity of $\triangle$, which was in the first set of exercises. The next is distributivity, $A\cap(B\triangle C) = (A\cap B)\triangle(A\cap C)$. To see this, the left-hand-side is $A\cap((B\cap C^c)\cup(C\cap B^c)) = (A\cap B\cap C^c)\cup(A\cap C\cap B^c)$, while the right-hand-side is $((A\cap B)\cap(A\cap C)^c)\cup((A\cap C)\cap(A\cap B)^c)$. Taking one of the two terms on the right-hand-side, we get $(A\cap B)\cap(A\cap C)^c = (A\cap B)\cap(A^c\cup C^c) = (A\cap B\cap A^c)\cup(A\cap B\cap C^c) = (A\cap B\cap C^c)$. Similarly the other term is $(A\cap C\cap B^c)$, and the right-hand-side equals the left-hand-side.

Boolean rings. The above example has the property that $A\cap A = A$ for all $A\in\mathcal{P}(X)$. Any ring with this property, that is $x.x = x$ for all $x\in R$, is called a *Boolean ring*, after George Boole. Note that in any Boolean ring, we have $x+x = 0$ for all $x\in R$. This is because $(x+x).(x+x) = x+x$, so $x.x+x.x+x.x+x.x = x+x$ by the distributive laws, so $x+x+x+x = x+x$, and therefore by the cancellation law, $x+x = 0$. Similarly, $x+y = (x+y).(x+y) = x.x+x.y+y.x+y.y = x+y+x.y+y.x$ so by cancellation $x.y+y.x = 0$. But $x.y+x.y = 0$, so by cancellation again, $x.y = y.x$, that is every Boolean ring is commutative.

Zero rings. If R has an addition defined on it, satisfying the axioms (A0)–(A4), then we can make R into a ring by defining the *zero multiplication*, $x.y = 0$ for all $x,y\in R$. It is easy to check the remaining ring axioms.

6 Subrings

Lecture 8

Definition. We have met several examples of one ring S contained inside another one R. If the ring operations in the small ring S are just the restrictions to S of the corresponding operations in R, we say that S is a *subring* of R.

In order to check that S is a subring of R, it is sufficient to check that

- $0 \in S$;
- if $a,b \in S$ then $a+b \in S$, and $a.b \in S$ and $-a \in S$.

This is because all the other ring axioms follow automatically in S, since they hold in the whole of R. For example, $a+b=b+a$ for all $a,b \in R$, so it certainly follows that $a+b=b+a$ for all $a,b \in S$.

Examples. $\mathbb{Z}$ is a subring of $\mathbb{R}$.

Suppose that S is a subring of R. If R is commutative, then so is S. But if S is commutative, R may not be: for example, if R is the ring of all 2×2 real matrices, and S is the subring of diagonal matrices.

But the situation with ones (identity elements) is rather unexpected. You will not be too surprised to learn that if R has a one, then S need not have a one: for example, $R=\mathbb{Z}$ and $S=2\mathbb{Z}$, the set of even integers. But you may be surprised to learn that if S has a one, it does not follow that R has a one; and even if it does, it *may not be the same as the one in* S. For example, if $R=M_2(\mathbb{Z})$, then it has a one, namely $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, while the subring S consisting of all matrices of the form $\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$ also has a one, namely $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Or keep the same S, and let R be the ring of all diagonal matrices. Now if we change R to be the ring of all diagonal matrices of the form $\begin{pmatrix} a & 0 \\ 0 & 2d \end{pmatrix}$, then R no longer has a one, but its subring S does.

Now suppose that R and S are both rings with one. Clearly R can be a division ring without S being a division ring: for example, take $R=\mathbb{Q}$ and $S=\mathbb{Z}$. It is a little less obvious that S can be a division ring without R being a division ring, even if R and S have the same one. For example, take $R=M_2(\mathbb{R})$, and S to be the subring of scalar matrices, that is matrices of the form $\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$.

A subring test. Suppose that S is a non-empty subset of a ring R. If $a-b \in S$ and $a.b \in S$ for all $a,b \in S$, then S is a subring. To see this, first pick any $a \in S$; then we have $0=a-a \in S$. Next, $-a=0-a \in S$; and if also $b \in S$, then $-b \in S$, so $a+b=a-(-b) \in S$. Lecture 9

7 Quotient rings

Our goal is to do 'modular arithmetic' in as general a context as possible. In that example, we have a ring, $\mathbb{Z}$, and a subring, $n\mathbb{Z}$, so let us see how far we can go if we replace $\mathbb{Z}$ by a general ring R, and replace $n\mathbb{Z}$ by a general subring S.

Congruence. First we define 'congruence modulo S' by analogy with congruence modulo n: just as we say $a \equiv b \bmod n$ if $a - b$ is divisible by n, we say that $a \equiv_S b$ if $a - b \in S$. This relation is

- reflexive, because for any $a \in R$ we have $a - a = 0 \in S$, that is $a \equiv_S a$;
- symmetric, because if $a \equiv_S b$ then $a - b \in S$, so $b - a = -(a - b) \in S$, that is $b \equiv_S a$;
- transitive, because if $a - b \in S$ and $b - c \in S$, then $a - c = (a - b) + (b - c) \in S$.

Thus it is an equivalence relation. Indeed, to prove this we have only used the facts that S is non-empty, and closed under addition and negation. We have not used the multiplication.

The equivalence classes of this relation are called the *cosets* of S. The coset containing a also contains $a + s$ for every $s \in S$. Indeed, it is easy to see that it consists of exactly these elements and no others. We write $S + a = a + S = \{a + s \mid s \in S\}$ for this coset.

When are two cosets equal? Just as in modular arithmetic we have $[a]_n = [b_n]$ if and only if $a - b$ is a multiple of n, so in this more general context we have $S + a = S + b$ if and only if $a - b \in S$.

Addition modulo S. Just as $[a]_n + [b]_n = [a + b]_n$, so we want to define $(S + a) + (S + b) = S + (a + b)$. And we have exactly the same problem as before, to prove that this is well-defined. So pick any $x \in S + a$ and any $y \in S + b$ and see what happens when we add them together. Now $x = s_1 + a$ for some $s_1 \in S$, and $y = s_2 + b$ for some $s_2 \in S$, so $x + y = s_1 + a + s_2 + b = (s_1 + s_2) + (a + b) \in S + (a + b)$. So addition of cosets is well-defined.

It is also commutative, since $(S + a) + (S + b) = S + (a + b) = S + (b + a) = (S + b) + (S + a)$. The zero coset $S = S + 0$ acts as a zero, since $(S + a) + (S + 0) = S + (a + 0) = S + a$. The negative of $S + a$ is $S + (-a)$, since $(S + a) + (S + (-a)) = S + (a + (-a)) = S + 0 = S$. Associativity of addition is also easy to see.

Multiplication modulo S. Just as $[a]_n.[b_n] = [a.b]_n$, we want to define $(S + a).(S + b) = S + (a.b)$. This time we encounter an unexpected problem, however. If we pick as before $x = s_1 + a \in S + a$ and $y = s_2 + b \in S + b$ and calculate $x.y$ we get $x.y = (s_1 + a).(s_2 + b) = s_1.s_2 + a.s_2 + s_1.b + a.b$, and in order to show that this is in $S + a.b$ we need each of the first three terms to lie in S. This is a stronger condition than just that S is a subring. We need the product of any element of R and any element of S to be in S. (This is what we actually use in the case of modular arithmetic: the product of any integer with any multiple of n is again a multiple of n.)

A subring S of a ring R is called an *ideal* of R if it has this extra property: $a.r \in S$ and $r.a \in S$ whenever $r \in R$ and $a \in S$.

If S is an ideal in R, then multiplication of cosets is well-defined, and the other ring axioms (associativity of multiplication, and the two distributive laws) are easy to verify. This means we have turned the set of cosets into a ring, known as the *quotient* of R by S, written R/S. For example, the ring of integers modulo n is the quotient of the ring $\mathbb{Z}$ by the ideal $n\mathbb{Z}$, so is now written $\mathbb{Z}/n\mathbb{Z}$.

This construction of a general quotient ring is exactly the same as the construction of modular arithmetic, just done a little more abstractly, with a slightly different notation. So if you understand modular arithmetic, you should understand quotient rings. But now we can do 'modular arithmetic' in a very general context, and we can make a very wide range of examples, which turn out to be very useful in a huge range of genuine applications, from encryption of internet commercial transactions, to error-correcting codes in virtually every electronic device in use today.

Summary. Here we summarise the construction of modular arithmetic and quotient rings, to show the connections clearly. Lecture 10

	Modular arithmetic $\mathbb{Z}/n\mathbb{Z}$	Quotient rings R/I
Equivalence relation:	$a \equiv b \bmod n$ if $a-b \in n\mathbb{Z}$	$a \equiv_I b$ if $a-b \in I$
Equivalence classes:	$[a]_n$	$I+a$
Addition:	$[a]_n + [b]_n = [a+b]_n$	$(I+a)+(I+b) = I+(a+b)$
is well-defined:	$x \in [a]_n, y \in [b]_n$	$x \in I+a, y \in I+b$
	$\Rightarrow x+y \in [a+b]_n$	$\Rightarrow x+y \in I+(a+b)$
Multiplication:	$[a]_n.[b]_n = [a.b]_n$	$(I+a).(I+b) = I+(a.b)$
is well-defined:	$x \in [a]_n, y \in [b]_n$	$x \in I+a, y \in I+b$
	$\Rightarrow x.y = (a+kn).(b+ln)$	$\Rightarrow x.y = (a+i_1).(b+i_2)$
	$= a.b+(kb+al+kln)n \in [a.b]_n$	$= a.b+a.i_2+i_1.b+i_1.i_2 \in I+(a.b)$
Zero:	$[0]_n+[a]_n = [a]_n$	$I+(I+a) = I+a$
Negatives:	$[a]_n+[-a]_n = [0]_n$	$(I+a)+(I+(-a)) = I+0 = I$

Unfortunately, there is plenty of scope for confusion with the notation here. For example, the symbol $+$ is used for at least three different things in this table: first, it is the addition in $\mathbb{Z}$ or R, as in $a+b$; second, it is used in the notation $I+a$ for the coset containing a, which we should perhaps called $[a]$ or $[a]_I$ instead; and third, it is used for the addition in $\mathbb{Z}/n\mathbb{Z}$ or R/I, as in the middle of $(I+a)+(I+b)$.

Examples of ideals. In any commutative ring R, and for any fixed $r \in R$, the set of multiples of r, that is $I = \{rx \mid x \in R\}$, is an ideal. This is because (it is non-empty and) for any $rx, ry \in I$ we have $rx - ry = r(x-y) \in I$, and for any $y \in R$ we have $(rx)y = r(xy) \in I$.

In a non-commutative ring, you need to multiply on both sides, and $I = \{xry \mid x, y \in R\}$ is an ideal. But this often gives the whole ring. For example, it can be shown that in $M_2(\mathbb{R})$ there are only two ideals, namely $\{0\}$ and the whole ring.

If $R = \mathcal{P}(A)$ with addition given by symmetric difference, and multiplication by intersection, the set of multiples of B (for a fixed $B \subseteq A$) is just the set $I = \mathcal{P}(B)$ of subsets of B. Since $\mathcal{P}(B)$ is non-empty, and for any $X \in R$ and $Y, Z \in I$ we have $Y + Z = Y \triangle Z \subseteq B$ and $X.Y = X \cap Y \subseteq B$, we see that I is an ideal in R. Lecture 11

What does the quotient ring R/I look like in this case? First look at the cosets: for any $X \in R$, the coset of X is $X + S = \{X \triangle Y \mid Y \subseteq B\}$. Now putting $Y = X \cap B$ we have $X \triangle Y = X \triangle (X \cap B) = X \setminus (X \cap B) = X \cap (A \setminus B) \subseteq A \setminus B$. So every coset contains an element which is a subset of $A \setminus B$. Indeed, it is quite easy to see that it contains *exactly* on such element. So there is a bijection between the set of cosets of $\mathcal{P}(B)$ in $\mathcal{P}(A)$, and the set of subsets of $A \setminus B$. That is, a bijection between $\mathcal{P}(A)/\mathcal{P}(B)$ and $\mathcal{P}(A \setminus B)$.

Now what about the ring operations? Look first at multiplication, that is, intersection. If $X, Y \subseteq A \setminus B$, then by definition $(\mathcal{P}(B) + X).(\mathcal{P}(B) + Y) = \mathcal{P}(B) + (X \cap Y)$. Similarly, $(\mathcal{P}(B) + X) + (\mathcal{P}(B) + Y) = \mathcal{P}(B) + (X \triangle Y)$. Thus the ring operations in the quotient $\mathcal{P}(A)/\mathcal{P}(B)$ *obviously* correspond exactly to the ring operations in $\mathcal{P}(A \setminus B)$.

Hence the quotient ring $\mathcal{P}(A)/\mathcal{P}(B)$ looks exactly the same as $\mathcal{P}(A \setminus B)$. We say that $\mathcal{P}(A)/\mathcal{P}(B)$ is isomorphic to $\mathcal{P}(A \setminus B)$.

A good example of a ring isomorphism is the map f between $\mathcal{P}(\{x\})$ under $\triangle$ and $\cap$ on the one hand, and $\mathbb{Z}/2\mathbb{Z}$ under $+$ and . on the other, defined by mapping $\emptyset$ to $[0]_2$ and $\{x\}$ to $[1]_2$.

Isomorphism. More formally, a map $f : R \to S$ from a ring R to a ring S is an *isomorphism* if it is a bijection, and it preserves the ring operations, in the sense that for all $a, b \in R$, Lecture 12

- $f(a+b) = f(a) + f(b)$;
- $f(a.b) = f(a).f(b)$.

Notice that the ring operations $+$ and . on the left are the operations in R, while those on the right are the operations in S.

If f is a ring isomorphism, then $f(a) + 0_S = f(a) = f(a + 0_R) = f(a) + f(0_R)$ so by cancellation, $f(0_R) = 0_S$. Similarly, $0_S = f(0_R) = f(a + (-a)) = f(a) + f(-a)$, so $f(-a) = -f(a)$.

Homomorphism. If we generalise this definition to any function f, not necessarily a bijection, we get a *homomorphism.* It still satisfies the rules $f(a+b) = f(a) + f(b)$ and $f(a.b) = f(a).f(b)$; and it still follows that $f(0) = 0$ and $f(-a) = -f(a)$.

A good example is the map $f : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ defined by $f(a) = [a]_n$. This is a homomorphism which is surjective but not injective. Another example is $g : \mathbb{Z} \to \mathbb{R}$ defined by $f(x) = x$. This homomorphism is injective but not surjective.

The *image* of a homomorphism $f : R \to S$ is the subset $\mathrm{im}(f) = f(R) = \{f(r) \mid r \in R\}$ of S. It is a subring, since it is non-empty ($f(0) = 0 \in f(R)$), and for all $f(a), f(b) \in f(R)$ we have $f(a) - f(b) = f(a-b) \in f(R)$ and $f(a).f(b) = f(a.b) \in f(R)$.

The *kernel* of a homomorphism $f : R \to S$ is the subset $\ker(f) = \{r \in R \mid f(r) = 0\}$ of R, that is the subset of elements which get mapped to zero. It is an ideal in R, since it is non-empty, and if $f(a) = f(b) = 0$ then $f(a-b) = f(a) - f(b) = 0 - 0 = 0$, and if also $r \in R$, then $f(a.r) = f(a).f(r) = 0.f(r) = 0$ and $f(r.a) = f(r).f(a) = f(r).0 = 0$.

The first isomorphism theorem. In the above example of $f : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$, the kernel of f is $\{k \in \mathbb{Z} \mid [k]_n = [0]_n\} = n\mathbb{Z}$, and you will notice that $\mathrm{im}(f) = \mathbb{Z}/\ker(f)$. In fact the same is true in general: if $f : R \to S$ is any ring homomorphism, then $\mathrm{im}(f)$ is isomorphic to $R/\ker(f)$. We write $\mathrm{im}(f) \cong R/\ker(f)$.

To prove this, we need to show there is an isomorphism $\phi : \mathrm{im}(f) \to R/\ker(f)$. Clearly we must define $\phi(f(r)) = (\ker(f)) + r$. But does this even make sense? If $f(r) = f(s)$ then $f(r-s) = f(r) - f(s) = 0$, so $r - s \in \ker(f)$, and therefore $(\ker(f)) + r = (\ker(f)) + s$. Similarly, reversing this argument, if $(\ker(f)) + r = (\ker(f)) + s$ then $f(r) = f(s)$, which means that ϕ is injective. Since ϕ is obviously surjective, this implies ϕ is a bijection. This is the hard part of the proof done. To show that ϕ preserves the ring operations is easy by comparison: $\phi(f(r) + f(s)) = \phi(f(r+s)) = (\ker(f)) + (r+s) = (\ker(f) + r) + (\ker(f) + s) = \phi(f(r)) + \phi(f(s)$, and similarly for multiplication.

To summarise: if $f : R \to S$ is a ring homomorphism, then (a) $\mathrm{im}(f)$ is a subring of S; (b) $\ker(f)$ is an ideal in R; and (c) $\mathrm{im}(f) \cong R/\ker(f)$. Lecture 13

As a canonical example, you should think of $R = \mathbb{Z}$, and $S = \mathbb{Z}_n$, that is the ring of integers modulo n. The map $f : \mathbb{Z} \to \mathbb{Z}_n$ defined by $f(a) = [a]_n$ is a ring homomorphism, which is surjective, and therefore we have $\mathbb{Z}_n \cong \mathbb{Z}/n\mathbb{Z}$. The latter isomorphism is given by $[a]_n \mapsto n\mathbb{Z} + a$.

Another example, considered above, is $R = \mathcal{P}(A)$ and $S = \mathcal{P}(B)$, where $B \subset A$, and f is defined by $f(X) = (X \cap B)$ for every $X \subseteq A$. Using properties of the set operations we get $(X \cap B) \triangle (Y \cap B) = (X \triangle Y) \cap B$ and (more obviously) $(X \cap B) \cap (Y \cap B)$, and therefore f is a ring homomorphism. The kernel of f is $\{X \subseteq A \mid X \cap B = \emptyset\} = \mathcal{P}(A \setminus B)$. Thus $\mathcal{P}(A \setminus B) \cong \mathcal{P}(A)/\mathcal{P}(B)$. In particular, we see that $\mathcal{P}(B)$ is an ideal in $\mathcal{P}(A)$.

The second isomorphism theorem. This is about quotient rings of quotient rings, and tells you that really they are just the same as ordinary quotient rings (which should be a relief). Suppose R is a ring, and $I \subseteq J$ are two ideals in R. Then we can define a ring homomorphism $f : R/I \to R/J$ by $f(I + r) = J + r$. This is well-defined because if $I + r = I + s$ then $r - s \in I \subseteq J$, so $J + r = J + s$. It is a homomorphism because $f((I+r)+(I+s)) = f(I+(r+s)) = J+(r+s) = (J+r)+(J+s) = f(I+r)+f(I+s)$ and $f((I+r).(I+s)) = f(I+(r.s)) = J+(r.s) = (J+r).(J+s) = f(I+r).f(I+s)$. Its kernel is $\{I + r \mid J + r = J\} = \{I + r \mid r \in J\} = J/I$. Hence by the first isomorphism theorem, $R/J \cong (R/I)/(J/I)$.

As an example, take $R = \mathbb{Z}$, $I = 6\mathbb{Z}$ and $J = 3\mathbb{Z}$. Then R/I is just the integers modulo 6, which as a set is $\{6\mathbb{Z}, 6\mathbb{Z}+1, 6\mathbb{Z}+2, 6\mathbb{Z}+3, 6\mathbb{Z}+4, 6\mathbb{Z}+5\}$. Inside this, $J/I = \{6\mathbb{Z}, 6\mathbb{Z}+3\}$, which has three cosets in R/I, namely itself and $\{6\mathbb{Z}+1, 6\mathbb{Z}+4\}$ and$\{6\mathbb{Z}+2, 6\mathbb{Z}+5\}$. These three cosets form the quotient ring $(R/I)/(J/I)$ and you can see that these three cosets just look like $0,1,2$ respectively in the ring of integers modulo 3. Indeed, the three cosets $3\mathbb{Z}$, $\mathbb{Z}+1$ and $\mathbb{Z}+2$ of $3\mathbb{Z}$ in $\mathbb{Z}$ are just formed by taking the appropriate union of cosets of $6\mathbb{Z}$.

One obvious consequence of the second isomorphism theorem is that (with the same notation) J/I is an ideal in R/I. Indeed, *every* ideal in R/I is of this form, for some ideal J in R with $I \subseteq J$. Similarly, every subring of R/I is of the form S/I for some subring S of R with $I \subseteq S$. This is known as the correspondence theorem. Lecture 14

The third isomorphism theorem. This is about quotient rings of subrings. If S is a subring of the ring R, and I is an ideal in R, then let us define $S+I = \{s+i \mid s \in S, i \in I\}$. Then we get a homomorphism $f : S \to R/I$ by defining $f(s) = I+s$. The image of f is then $(S+I)/I$. The kernel of f is $\{s \in S \mid I+s = I\} = s \in S \mid s \in I\} = S \cap I$. Hence $(S+I)/I \cong S/(S \cap I)$. In particular, $S+I$ is a subring of R, and $S \cap I$ is an ideal in S. (However, not every ideal of S need be of this form.)

8 Factorisation

You saw in Introduction to Algebra that $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime. When we generalize to quotients of other rings than $\mathbb{Z}$, we are going to need appropriate generalizations of ‘prime’, and appropriate generalizations of factorizing elements as products of primes. To avoid complications, let us assume our ring is commutative, and has a one. Lecture 15

Zero divisors. In $\mathbb{Z}_6$ we have the annoying fact that $[2].[3] = [0]$. That is, a product of two non-zero elements can be zero. This has the even more annoying consequence that $[2].[4] = [2]$, which plays havoc with any idea of factorising elements as products of primes. More generally, if $a,b \in R$ are non-zero elements with $a.b = 0$, we call a and b *zero-divisors*. In this situation, we have $a.(b+1) = a$, which again means that there is no sensible meaning to ‘factorising’ an element.

Clearly we shall want to restrict attention to rings which do not have zero-divisors. A commutative ring with a one which has no zero-divisors is called an *integral domain* (or sometimes just a domain: but do not confuse this with the domain of a function).

Example: In $\mathbb{Z}/n\mathbb{Z}$, the element $[a]$ is a zero-divisor if and only if $1 < \gcd(a,n) < n$. For if $n > \gcd(a,n) = d > 1$ then $a = dx$ and $n = dy$ for some $x,y, \in \mathbb{Z}$ so n divides $nx = dyx = ay$ but n does not divide y or a. In other words $[a].[y] = [0]$ but $[a] \neq [0] \neq [y]$, so $[a]$ is a zero-divisor. On the other hand, if $\gcd(a,n) = n$ then $[a] = [0]$, so $[a]$ is not a zero-divisor. Similarly, if $\gcd(a,n) = 1$, and $[a].[y] = [0]$, then n divides ay but has no common factors with a, so must divide y.

This tells us that $\mathbb{Z}/n\mathbb{Z}$ is an integral domain if and only if n is prime.

Example: If R is an integral domain, then so is $R[x]$ (the ring of polynomials over R). For if the leading terms of two polynomials are a_nx^n and b_mx^m, with (by definition) $a_n \neq 0 \neq b_m$, then the leading term of their product is $a_nb_mx^{n+m}$, since $a_nb_m \neq 0$. In particular, the product of two non-zero polynomials is non-zero.

Units. In a ring with a one, some elements may have multiplicative inverses: an *inverse* of $a \in R$ is an element b such that $ab = ba = 1$. If a has an inverse, then that inverse is unique (the proof is very similar to the proof that $-a$ is well-defined), and is written a^{-1}. An element which has an inverse is called a *unit*. Now 1 is a unit, since $1.1 = 1$, and if u is a unit, then so is u^{-1}. Moreover, if u and v are units, then so is uv, since $(uv)(v^{-1}u^{-1}) = 1$. Lecture 16

This means that the units in a ring R form a *group*, written $U(R)$, under multiplication.

In $\mathbb{Z}$, the units are 1 and -1. In $\mathbb{Z}/n\mathbb{Z}$ they are all $[a]$ such that $\gcd(a,n) = 1$. For if $\gcd(a,n) = 1$ then by Euclid's algorithm $1 = as + nt$ for some $s,t \in \mathbb{Z}$, so $[a].[s] = [1]$. Conversely, if $\gcd(a,n) > 1$ then either $[a] = [0]$ or $[a]$ is a zero-divisor. In either case, $[a]$ is not a unit.

(In general, an element cannot be both a unit and a zero-divisor: for if a is a unit, then there exists b such that $ab = 1$, while if a is a zero-divisor, there exists $c \neq 0$ such that $ac = 0$, and putting these together gives $c = 1.c = ab.c = ac.b = 0.b = 0$, which is a contradiction.)

Two elements a, b in R, a commutative ring with one, are called *associates* if $a = bu$ for some unit u. The properties just proved for units mean that the relation of being associates is reflexive, symmetric and transitive, so is an equivalence relation. The equivalence classes are called *associate classes*. The associate classes in $\mathbb{Z}/15\mathbb{Z}$ are $\{0\}$, the units $\{1,2,4,7,8,11,13,14\}$, $\{3,6,9,12\}$ and $\{5,10\}$.

More generally, the associate classes in $\mathbb{Z}/n\mathbb{Z}$ are $\{a \mid \gcd(a,n) = d\}$ for each divisor d of n.

Divisibility in integral domains. If we want to define greatest common divisors (g.c.d.s) in general, we must use only divisibility, and not the ordering which we used on $\mathbb{Z}$. Even then, g.c.d.s may not exist. In any integral domain R, we say that *a divides b*, and write $a|b$, if there exists $c \in R$ such that $b = ac$. Lecture 17

Notice that if $a|b$ and $b|a$ then a and b are associates: for if $b = ac$ and $a = bd$ we can substitute one in the other to get $b = bcd$, and then $b(1-cd) = 0$. Since R is an integral domain, we deduce that either $b = 0$ (in which case also $a = 0$) or $1 - cd = 0$ (in which case $cd = 1$ so c and d are units). The converse is also true, and easy to see: if a and b are associates, then $a = bu$ and $b = av$ for some units u, v, so $a|b$ and $b|a$.

Now a *greatest common divisor* of a and b is any element d which satisfies:

- $d|a$ and $d|b$;

- if $e|a$ and $e|b$ then $e|d$.

Suppose that d' is another g.c.d. of a and b. Then $d'|a$ and $d'|b$, so by the second part of the definition $d'|d$. Similarly, since d' is a g.c.d. of a and b, and $d|a$ and $d|b$, we deduce that $d|d'$. Hence d and d' are associates.

Conversely, if d' is an associate of d, say $d' = du$ and $d = d'v$ for some units u, v, then $d'|d|a$ and $d'|d|b$, and if $e|a$ and $e|b$ then $e|d|d'$, so d' is another g.c.d. of a and b.

Euclidean domains. So far, the only method we know for finding g.c.d.s is Euclid's algorithm. This relies on repeated application of the *division algorithm*, which allows us to divide one (positive) integer by another to get a quotient and a remainder, with the crucial property that the remainder is *smaller* than the divisor. You can also easily generalise to negative integers. That is, given a, $b \neq 0$ we can find q and r such that $a = bq + r$, and $0 \leq r < b$. Lecture 18

In order to generalise this we need an appropriate notion of 'smaller than'. A *Euclidean function* on an integral domain R is a function $d : R \to \mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$ with the following properties:

- $d(ab) \geq d(a)$, whenever $b \neq 0$;
- if $b \neq 0$, and $a \in R$, then there exist $q, r \in R$ such that

$$\begin{aligned} &a = bq + r \text{ and} \\ &d(r) < d(b), \text{ or } r = 0. \end{aligned}$$

If R has a Euclidean function, then it is called a *Euclidean domain*.

For example, $\mathbb{Z}$ is a Euclidean domain: there are various possible choices for the Euclidean function, for example $d(a) = |a|$, or $d(a) = |a| - 1$, or $d(a) = a^2$. Another example is the ring $\mathbb{R}[X]$ of polynomials with real coefficients, with $d(f) = \deg(f)$.

In any of these examples, we now have a method of computing g.c.d.s. Given a and b, put $a_0 = a$ and $a_1 = b$, and successively compute $a_0 = a_1q_1 + a_2, \ldots, a_{n-1} = a_nq_n + a_{n+1}$, $a_n = a_{n+1}q_{n+1}$, where we suppose that a_{n+1} is the last non-zero remainder. (The algorithm does terminate, because the remainders at each stage are getting 'smaller', in the sense that $d(a_{i+1}) < d(a_i)$, and because the degrees are non-negative integers, they must eventually stop.) The last equation shows that $\gcd(a_n, a_{n+1}) = a_{n+1}$, and the general equation $a_i = a_{i+1}q_{i+1} + a_{i+2}$ shows that $\gcd(a_i, a_{i+1}) = \gcd(a_{i+1}, a_{i+2})$. Hence by induction, $\gcd(a, b) = a_{n+1}$. Working backwards, substituting each equation in the previous one, enables us to write $a_{n+1} = ax + by$ for some $x, y \in R$.

Principal ideal domains. It is clear that if R is any commutative ring, and $a \in R$, then the set $aR = \{ar \mid r \in R\}$ of multiples of a is an ideal. An ideal of this form is called a *principal ideal*. In a Euclidean domain, it turns out that *all* ideals are of this form. Lecture 19

For suppose that R is a Euclidean domain, and I is a non-zero ideal in R. Then we can pick $a \in I \setminus \{0\}$ of minimal degree, that is $d(a) \leq d(x)$ for all $x \in I \setminus \{0\}$. I claim that in this case $I = aR$. One direction is obvious: if $ar \in aR$, then $ar \in I$, by definition of an ideal. For the other direction, suppose $x \in I$. Then by Euclid's algorithm, there exist $q, r \in R$ such that $x = aq + r$, and either $r = 0$ or $d(r) < d(a)$. The latter possibility leads to $r = x - aq \in I$ (since $x, a \in I$ and I is an ideal), which contradicts the choice of a as a 'smallest' element in I. Hence $r = 0$, and so $x = aq \in aR$.

A *principal ideal domain* (abbreviated PID) is an integral domain in which every ideal is principal. Thus we have shown that every Euclidean domain is a principal ideal domain.

The converse is actually false: that is, there are PIDs which are not Euclidean. However, it is not easy to prove this.

The real point of introducing PIDs is that they give us a slightly more general context in which we can talk about greatest common divisors. To see this, let R be a PID, and let $a, b \in R$. Then it is easy to see that if $I = aR + bR = \{ax + by \mid x, y \in R\}$ then I is an ideal (just check the defining conditions for an ideal), so by assumption $I = dR$ for some $d \in R$. Now it is quite easy to check that d is a gcd of a and b: first we have $a \in I = dR$ so $a = dx$ for some $x \in R$, that is $d|a$. Similarly, $d|b$. Moreover, if $e|a$ and $e|b$ then e divides every element $ax + by$ of I. In particular, $e|d$.

Notice that not only does a $\gcd(a, b)$ exist, but it can also be written in the form $ax + by$ with $x, y \in R$. This is what Euclid's algorithm really does for us, but as just mentioned, there are some PIDs which are not Euclidean domains, so we get the same result in a more general context.

Unique factorisation domains. You have probably seen how Euclid's algorithm can be used to show that factorisation of (positive) integers into primes is "essentially unique". The essential point is to prove that if a prime p divides $a.b$ then either it divides a or it divides b (or both). Then by induction, if p divides a product of primes $q_1.q_2.\cdots.q_r$ then p is equal to one of the primes q_i. The same idea works more generally in PIDs, although there are a few extra technicalities. Lecture 20

First we need a definition: in any integral domain R, an element $p \in R$ is called *irreducible* if it cannot be factorised in a non-trivial way, that is, if $p = a.b$ then either a or b (but not both!) is a unit. (In particular, note that 0 is not irreducible, and units are not irreducible.)

Now suppose that R is a PID, and $p \in R$ is irreducible. If $p|ab$, then either $p|a$, or (because p is irreducible) $\gcd(a, p) = 1$. In the latter case, since R is a PID, we have that $1 = ax + py$ for some $x, y \in R$. Hence, $b = abx + pyb$, and since p divides both terms on the RHS, we have $p|b$. This result can be extended, by an easy induction, to the following: if R is a PID, and $p \in R$ is irreducible, and if $p|(q_1.q_2.\cdots.q_s)$, then $p|q_i$ for some i.

Now we can prove that factorisations, if they exist, are unique, in the following sense: if $p_1.p_2.\cdots.p_r = q_1.q_2.\cdots.q_s$, where the p_i and q_j are all irreducible, then

$r = s$, and after re-ordering if necessary, each p_i is an associate of the corresponding q_i. For p_1 divides $q_1.q_2.\cdots.q_s$, so p_1 divides one of the q_i. Let us re-order the q_i so that p_1 divides q_1. Thus $q_1 = p_1u_1$, and because q_1 is irreducible, u_1 is a unit. Now because we are in an integral domain, the cancellation law for multiplication holds, and we have $u_1p_2.\cdots.p_r = q_2.\cdots.q_s$. So by induction we can strip off the irreducibles from both sides, one at a time, until we run out of irreducibles on one side or the other. At this stage we have a product of irreducibles equal to a unit, which cannot happen, unless we have run out of irreducibles on the other side as well. Hence $r = s$ as required.

The only problem that remains is that we have not shown that factorisations actually exist in a PID! In other words, we have not shown that the process of refining factorisations of an element ever stops. In fact it does, but the proof is not easy, and does not work in every integral domain. Suppose we have an element $a = a_0$, and that we can keep on factorising $a_0 = a_1b_1, \ldots, a_{n-1} = a_nb_n, \ldots$, indefinitely. Then we have a sequence of ideals each contained in the next:

$$a_0R \subset a_1R \subset a_2R \subset \cdots$$

It is easy to see that the union $\bigcup_{i\geq 0} a_iR$ is also an ideal (just check the definition), so, since R is a PID, it is of the form dR for some d. Since d is in this ideal, it is in one of the a_iR. But then $a_{i+1}R \subseteq dR \subseteq a_iR$, which is a contradiction. So the factorising cannot continue for ever.

This is worth recording in a formal definition: a *unique factorisation domain* (abbreviated UFD) is an integral domain in which

- every element (except 0, units, and irreducibles) can be written as a product of irreducibles, and
- whenever $p_1.p_2.\cdots.p_r = q_1.q_2.\cdots.q_s$, where the p_i and q_j are all irreducible, then $r = s$, and after re-ordering if necessary, each p_i is an associate of the corresponding q_i.

We have shown that every PID is a UFD.

Examples. We started with examples like $\mathbb{Z}$ and $\mathbb{R}[x]$, and distilled out of them the definition of a *Euclidean domain*. We proved that every Euclidean domain is a PID, and that every PID is a UFD. In particular, $\mathbb{Z}$ has unique factorisation, and so does $\mathbb{R}[x]$. Lecture 21

More generally, $F[x]$, the ring of polynomials over F, is a Euclidean domain, provided F is a field. However, if F is not a field, it need not be. For example, $\mathbb{Z}[x]$ is not even a PID, since we already showed that the ideal of poylnomials with constant term in $2\mathbb{Z}$ is not a principal ideal. In fact, however, $\mathbb{Z}[x]$ is a UFD, though this is somewhat technical to prove: the idea is to work with factorisations in $\mathbb{Q}[x]$, where they uniqueness condition holds, and then to shuffle the denominators around until they disappear.

We showed that the Gaussian integers $\mathbb{Z}[i] = \{a+bi \mid a,b \in \mathbb{Z}\}$, where $i^2 = -1$, is a Euclidean domain. Similarly, $\mathbb{Z}[\sqrt{-2}] = \{a+b\sqrt{-2} \mid a,b \in \mathbb{Z}\}$ is a Euclidean domain. However, $\mathbb{Z}[\sqrt{-3}]$ is not. For $4 = 2.2 = (1+\sqrt{-3})(1-\sqrt{-3})$, and we can show that all the factors 2, $1 \pm \sqrt{-3}$ are irreducible, but are not associates of each other. Thus it is not even a UFD, let alone a PID or a Euclidean domain. Similarly in $\mathbb{Z}[\sqrt{-5}]$ we have $6 = 2.3 = (1+\sqrt{-5})(1-\sqrt{-5})$.

Now consider the ring $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-3})]$. (First check that it is a ring, by checking it is a subring of the complex numbers.) Now we find 6 units, being ± 1 and $\frac{1}{2}(\pm 1 \pm \sqrt{-3})$. This ring is a Euclidean domain, with the usual Euclidean function $d(z) = |z|$, or $|z|^2$. Trying a few factorisations, we find 2 is irreducible, $3 = \sqrt{-3}.(-\sqrt{-3})$, and 5 is irreducible. Then $7 = (2+\sqrt{-3}).(2-\sqrt{-3})$.

9 Fields

Lecture 22

Recall that a field is a commutative ring with a one, in which every non-zero element has a (multiplicative) inverse. That is, for any $x \in F$ with $x \neq 0$ there exists $y \in F$ with $xy = 1$. In other words, every non-zero element is a unit. Since units cannot be zero-divisors, it follows that every field is an integral domain.

Notation for principal ideals: we have written aR for the ideal $\{ar \mid r \in R\}$ in any commutative ring. Sometimes we shall write $\langle a \rangle$ or (a) for this (principal) ideal.

Recall that $\mathbb{Z}/n\mathbb{Z}$ is a field whenever n is prime, but is not even an integral domain if n is composite. The same is true for arbitrary PIDs in place of $\mathbb{Z}$, and 'irreducible' in place of 'prime'. That is, if R is a PID, and $a \in R$, then $R/\langle a \rangle$ is a field if and only if a is irreducible. To prove this, we first see that if $a = bc$, with neither b nor c being a unit, then $(\langle a \rangle + b).(\langle a \rangle + c) = \langle a \rangle$; moreover if $\langle a \rangle + b = \langle a \rangle$ then $b \in \langle a \rangle$, so $a|b$, but we already have $b|a$, so a and b are associates, which means c is a unit. This is a contradiction. Hence we have zero-divisors in $R/\langle a \rangle$, so it is not a field.

Conversely, if a is irreducible, and $\langle a \rangle + b \neq \langle a \rangle$, then a does not divide b, so $\gcd(a,b) = 1$. Hence by the PID property, $\gcd(a,b) = 1 = ax+by$ for some $x,y \in R$, and in the quotient ring we have $(\langle a \rangle + b).(\langle a \rangle + y) = \langle a \rangle + 1$, which means that $\langle a \rangle + b$ is invertible. Hence $R/\langle a \rangle$ is a field.

Examples. We know that $\mathbb{R}[X]$ is a field. Also, X^2+1 is irreducible, since it cannot be factorised into real linear factors. So what does the quotient $\mathbb{R}[X]/\langle X^2+1 \rangle$ look like? Its elements are the cosets $\langle X^2+1 \rangle + aX + b$, where $a,b \in \mathbb{R}$, and these add together just like linear polynomials. They multiply together by the rule $(\langle X^2+1 \rangle + aX+b).(\langle X^2+1 \rangle + cX+d) = \langle X^2+1 \rangle + acX^2 + (ad+bc)X + bd = \langle X^2+1 \rangle + (ad+bc)X + (bd-ac)$ which is just the same as multiplying the complex numbers $ai+b$ and $ci+d$. In other words $\mathbb{R}[X]/\langle X^2+1 \rangle \cong \mathbb{C}$. The coset $\langle X^2+1 \rangle + X$ plays the role of $i = \sqrt{-1}$.

Similarly in other cases: the effect is to 'adjoin a root of the irreducible polynomial'. So for example $\mathbb{Q}[X]/\langle X^2-2 \rangle \cong \mathbb{Q}[\sqrt{2}]$, because the coset $\langle X^2-2 \rangle + X$

behaves exactly like a square root of 2.

In general, we know that $F[X]$ is a PID whenever F is a field. We can even take F to be a finite field, for example $F = \mathbb{Z}/2\mathbb{Z}$: let's use the simplified notation $\{0,1\}$ for this field, with the understanding that $1+1=0$. Now X^2+X+1 is irreducible in $F[X]$, and the quotient $F[X]/\langle X^2+X+1\rangle$ consists of just four cosets, $\langle X^2+X+1\rangle + aX+b$, which we might as well simplify to $0,1,\alpha,\alpha+1$, where $\alpha = \langle X^2+X+1\rangle + X$. Thus we get a field of order 4.

In lecture 23 I think I talked about adjoining a root of an irreducible polynomial in general (Proposition 2.37 from Prof. Cameron's notes), and a little bit about finite fields: the fact that every finite field has order p^n, where p is prime and $n \geq 1$, and that there is exactly one field of each such order (without proof!). As an example, I showed that adjoining a root of X^2+X-1 to $\mathbb{Z}_3$ gives the same field of order 9 as adjoining a root of X^2+1.

Field of fractions. We generalise the construction of the field of rational numbers, $\mathbb{Q}$, from the integral domain of the integers, $\mathbb{Z}$. There is no analogue in general of the idea of putting a fraction into its lowest terms, so we have to consider all possible ways of writing a fraction to be equivalent. Of course, we want two fractions a/b and c/d to be equal if and only $ad = bc$. Lecture 24

Thus we start by taking an integral domain R, and making the set X of all ordered pairs (a,b), with $a,b \in R$ and $b \neq 0$. Since (a,b) is going to represent the fraction a/b, we define an equivalence relation $\equiv$ on X by $(a,b) \equiv (c,d)$ whenever $ad = bc$. (It is quite easy to check that this is an equivalence relation: $ab = ba$ implies $(a,b) \equiv (a,b)$, so the relation is reflexive; if $ad = bc$ then $cb = da$ so the relation is symmetric; and if $ad = bc$ and $cf = de$ then $ade = bce$ so $acf = bce$, so $c(af - be) = 0$, and therefore *either* $af - be = 0$, so $af = be$, *or* $c = 0$, in which case $ad = 0$ but $d \neq 0$ so $a = 0$, and similarly $e = 0$, so $af = 0 = be$, so again $af = be$, and the relation is transitive.)

Now we *define* the fraction a/b to be the equivalence class containing (a,b). In order to make this set of fractions into a ring, we must define addition and multiplication by the usual formulas: $a/b + c/d = (ad+bc)/bd$ and $a/b.c/d = ac/bd$. In each case we must check that it is well-defined: that is, if we replace the name a/b of a fraction by another name a'/b' of the same fraction (that is, with $ab' = a'b$), then we just get another name for the same fraction in the sum, or product. In other words, we check that $(a'd + b'c)bd = (ad+bc)b'd$ and $a'c.bd = ac.b'd$.

With these definitions, we find that $0/1$ is the zero, and $(-a)/b$ is the negative of a/b. Moreover, if $a/b \neq 0$, then $a \neq 0$, and b/a is a multiplicative inverse to a/b, since $a/b.b/a = ab/ab = 1/1 = 1$. Hence we have constructed a field.

The original integral domain is a subring of this field, if we identify $r \in R$ with the fraction $r/1$.

10 Groups

Lecture 25 can be found elsewhere on the course web-page.

Recall the definition of a group, as a set G with a binary operation $\star$, satisfying the following: Lecture 26

(G0) Closure: for any $g,h \in G$, we have $g \star h \in G$;

(G1) Associativity: $(g \star h) \star k = g \star (h \star k)$ for any $g,h,k \in G$;

(G2) Identity: there exists $e \in G$ with $g \star e = e \star g = g$ for every $g \in G$;

(G3) Inverses: for every $g \in G$ there exists $g^{-1} \in G$ such that $g \star g^{-1} = g^{-1} \star g = e$.

If also (G4) $g \star h = h \star g$ for all $g,h \in G$, then G is called *commutative* or (more usually) *abelian*.

Examples. If R is a ring, then R with the operation $+$ is an abelian group: the group axioms are just the same as the axioms for addition in the ring. The identity element e is just the 0 of the ring, and the inverse of an element g is the negative $-g$, since $g + (-g) = 0$.

If R is a ring with a one, and $U(R)$ is the set of units of R (that is, the elements which have a multiplicative inverse), then $U(R)$ forms a group with the operation of multiplication. The identity element of the group is just the one in the ring, and the group inverse is the same as the multiplicative inverse in the ring.

To take a special case, suppose F is a field, and let $R = M_2(F)$ be the ring of 2×2 matrices with entries from F. Then the units of R are just the matrices with non-zero determinant. These form a group under matrix multiplication, called the *general linear group* (of degree 2, over F). Write $GL_2(F)$ for this group.

To take an even more special case, let $F = \mathbb{Z}_2 = \{0,1\}$ be the field of order 2. Then there are just 6 invertible 2×2 matrices, and we have

$$GL_2(\mathbb{Z}_2) = \{\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}, \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}, \begin{pmatrix}0 & 1\\1 & 0\end{pmatrix}, \begin{pmatrix}0 & 1\\1 & 1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 0\end{pmatrix}, \begin{pmatrix}1 & 1\\0 & 1\end{pmatrix}\}.$$

You can check that all these elements are their own inverses, except that

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

so that these two matrices are inverses of each other.

Properties of groups. Many of these are proved in exactly the same way as for rings. For example, there is only one identity element: if e and e' are identity elements, then $e \star e' = e$ because e' is an identity element, and $e \star e' = e'$ because e is an identity element, so $e = e'$.

Similarly, if g has two inverses, say h and k, then $g \star h = h \star g = e$ and $g \star k = k \star g = e$ so $h \star (g \star k) = h \star e = h$ and $(h \star g) \star k = e \star k = k$, so by the associative law, $h = k$. The cancellation laws are similar: if $g \star h = g \star k$, then $h = g^{-1} \star g \star h = g^{-1} \star g \star k = k$, and similary for the other cancellation law.

The inverse of $g \star h$ is $h^{-1} \star g^{-1}$, as we already saw in the case of units in rings.

Lecture 27

Notation. Various different notations for groups are used. General groups are usually written with the operation written $g.h$ or just gh, and the identity element is then usually written 1. This is in conformity with the notation for multiplication in rings and fields. Abelian groups are often written with the operation written $g + h$, in conformity with the notation for the additive group of a ring: in this case, the identity element is always written 0, and the additive inverse of g is written $-g$ instead of g^{-1}.

The *order* of a group G is just the number of elements it has, that is $|G|$.

The *order* of an *element* $g \in G$ is defined in a completely different way: it is the smallest positive integer n such that $g^n = 1$. Here $g^n = g.g.g. \cdots .g$ with n factors. If we define $g^{-n} = (g^{-1})^n$ then it should be clear that $g^{-n}.g^n = 1$. More generally $g^n.g^k = g^{n+k}$, where $n, k \in \mathbb{Z}$, and can be negative.

We will see in a moment that these two concepts of 'order' are closely related. First we need:

Subgroups. A subgroup of a group G is a subset H which is a group in its own right (with the same operation). That is, it must contain 1, and whenever $g, h \in H$, we must have the product $gh \in H$ and the inverse $g^{-1} \in H$.

A possibly slightly simpler test is the following: $H \subseteq G$ is a subgroup if (and only if) H is non-empty, and for all $g, h \in H$, we have $gh^{-1} \in H$. The 'only if' part is not really used, but is easy: if H is a subgroup, then $1 \in H$ so H is non-empty; and if $g, h \in H$ then $h^{-1} \in H$, and then $g.h^{-1} \in H$. The 'if' part is what we shall use:

If H is non-empty, say $g \in H$, then by assumption (taking $h = g$) we have $1 = gg^{-1} \in H$. Next, for any $x \in H$, we have (taking $g = 1$, $h = x$) that $x^{-1} = 1.x^{-1} \in H$. Finally, if $x, y \in H$, then $y^{-1} \in H$, and therefore (taking $g = x, h = y^{-1}$) we have $xy = x(y^{-1})^{-1}) \in H$. (Note that this last step requires the uniqueness of inverses: both y and $(y^{-1})^{-1}$ are inverses of y^{-1}, so are equal.)

Cyclic groups. If $g \in G$ is an element of order n, then I claim that $\{1 = g^0, g = g^1, g^2, \ldots, g^{n-1}\}$ is a subgroup of G, of order n. To check this, first observe that all its elements are distinct: for if $g^k = g^l$ with $0 \leq k < l \leq n-1$ then $g^{l-k} = 1$ and $1 \leq k - l \leq n-1$, contradicting the assumption that the order of g is n.

Now check the subgroup conditions: certainly the set is non-empty, and if g^k, g^l are elements of this set, then either $k \geq l$, in which case $g^k.(g^l)^{-1} = g^{k-l}$, and $0 \leq k-l \leq n-1$; or $k < l$, in which case $g^k.(g^l)^{-1} = g^{k-l} = g^{k-l+n}$, and $1-n \leq k-l \leq -1$ so $1 \leq k-l+n \leq n-1$. Hence it is a subgroup, as claimed.

Isomorphism. Notice that in a cyclic group, all that matters is how the exponents k in g^k behave: when we *multiply* group elements we *add* the exponents, *modulo n*. So the exponents behave like $\mathbb{Z}_n$ under addition modulo n. Thus the cyclic group is really the same as (we say, isomorphic to) the additive group of integers modulo n.

More formally, an *isomorphism* between two groups G and H is a bijection $f : G \to H$ such that $f(gh) = f(g)f(h)$ for all $g,h \in G$. Notice that this implies that $f(1) = f(1.1) = f(1).f(1)$ so $f(1) = 1$; and then $1 = f(1) = f(g.g^{-1}) = f(g).f(g^{-1})$, so by uniqueness of inverses, we have $f(g^{-1}) = (f(g))^{-1}$. In other words, all the group structure is preserved by the function f.

Subgroups of cyclic groups. Since the cyclic group C_n of order n is isomorphic to the additive group of integers modulo n, we may as well use this more familiar example in order to illustrate the subgroup structure. All the basic theorems depend on Euclid's algorithm (or at least the division algorithm), just as they do in the ring theory part of this course. Indeed, we are almost proving the same theorems again. Lecture 28

So let $\mathbb{Z}_n = \{0,1,2,\dots,n-1\}$ be the additive group of integers modulo n, so that $2 = 1+1$, $3 = 1+1+1$ etc., and $-1 = n-1$ and so on. If $k|n$, say $n = kl$, then let $H = \{0,k,2k,\dots,(l-1)k\}$. It is easy to check that H is a subgroup of $\mathbb{Z}_n$, and that it has order l. In this way we get a subgroup of order l, for each divisor l of n.

Conversely, suppose that H is any subgroup of $\mathbb{Z}_n$, and suppose that k is the smallest positive integer such that $k \in H$. We show that $H = \{0,k,2k,\dots,(l-1)k\}$, where $kl = n$. First we need to show that $k|n$: if not, then $n = kq+r$ with $0 \leq r < k$, and $r = -kq \in H$ (remember we are working modulo n), which is a contradiction. Certainly H contains $\{0,k,2k,\dots,(l-1)k\}$, so suppose that $b \in H$ is some other element. Then $b = kq+r$ for some $0 \leq r < k$, and $r = b-kq \in H$, so $r = 0$. Hence $b = kq$, and therefore $H = \{0,k,2k,\dots,(l-1)k\}$ as claimed.

Right cosets. We generalise the idea of cosets in rings: in the situation where we had a subring S of a ring R (in fact, we did not even need S to be closed under multiplication) we defined an equivalence relation $\equiv_S$ on R by $a \equiv_S b$ if and only if $a-b \in S$. We do the same thing here, except that we replace the additive notation by the multiplicative notation:

Suppose H is a subgroup of a group G, and define a relation $\equiv_H$ on G by $g \equiv_H h$ if and only if $gh^{-1} \in H$. We check that this is an equivalence relation:

- $1 = gg^{-1} \in H$ for all $g \in G$, so the relation is reflexive;

- if $gh^{-1} \in H$ then $hg^{-1} = (h^{-1})^{-1}g^{-1} = (gh^{-1})^{-1} \in H$, so the relation is symmetric;
- if $gh^{-1} \in H$ and $hk^{-1} \in H$, then $gk^{-1} = (gh^{-1}).(hk^{-1}) \in H$, so the relation is transitive.

The equivalence classes are called the *right cosets* of H in G. The equivalence class containing g is the set of all x such that $xg^{-1} \in H$, that is the set of $x = hg$, where $h \in H$. Thus we write $Hg = \{hg \mid h \in H\}$ for this right coset.

Left cosets. In a similar way we can define *left cosets* $gH = \{gh \mid h \in H\}$, using the equivalence relation defined by $g^{-1}h \in H$. Sometimes the left cosets are the same as the right cosets, but often they are not.

Lagrange's Theorem. It is fairly obvious that every coset of H in G has the same Lecture 29
number of elements, because to each element $h \in H$ there is a corresponding element $hg \in Hg$. More formally, the map $\phi : H \to Hg$ defined by $\phi(h) = hg$ is a bijection, because it has an inverse map $\psi : Hg \to H$ defined by $\psi(x) = xg^{-1}$. (Proof: $\phi(\psi(x)) = \phi(xg^{-1}) = (xg^{-1})g = x$ and $\psi(\phi(h)) = \psi(hg) = (hg)g^{-1} = h$.)

Now G is partitioned into the right cosets of H, since these are the equivalence classes of an equivalence relation. Hence the total number of elements in G is equal to the number of right cosets times the number of elements ($|H|$) in each. In other words, the number of right cosets of H in G is $|G|/|H|$, and since this must be a positive integer, we have that $|H|$ divides $|G|$.

As an immediate corollary, we see that the order of any element $g \in G$ also divides $|G|$, since it is equal to the order of the subgroup $\langle g \rangle = \{1, g, g^2, \ldots\}$ of G.

Of course, in the proof we could equally well have used left cosets. In fact, there is a bijection ϕ between the set of left cosets and the set of right cosets defined by $\phi(gH) = Hg^{-1}$. To prove this formally, we first need to check that this map is well-defined: if $xH = yH$ then $x \in yH$ so $x = yh$ for some $h \in H$, and therefore $x^{-1} = h^{-1}y^{-1} \in Hy^{-1}$, so $Hx^{-1} = Hy^{-1}$. This argument in reverse shows that ϕ is injective, and it is obvious that ϕ is surjective. Hence ϕ is a bijection, as required.

Homomorphisms. We already saw the definition of an isomorphism of groups: a bijection θ which preserves the group multiplication, in the sense that $\theta(xy) = \theta(x)\theta(y)$ for all x, y. Similarly, a *homomorphism of groups* is a function $\theta : G \to H$, where G and H are groups, such that $\theta(xy) = \theta(x)\theta(y)$ for all $x, y \in G$.

It follows easily that θ preserves the identity element, and inverses: that is, $\theta(1_G) = 1_H$ and $\theta(g^{-1}) = \theta(g)^{-1}$. Proof: $1_H.\theta(1_G) = \theta(1_G) = \theta(1_G.1_G) = \theta(1_G).\theta(1_G)$ so by cancellation, $1_H = \theta(1_G)$. Also $1_H = \theta(1_G) = \theta(g.g^{-1}) = \theta(g).\theta(g^{-1})$, so $\theta(g^{-1}) = \theta(g)^{-1}$.

Now if $\theta : G \to H$ is a group homomorphism, then the *image* of θ is $\in \theta = \{\theta(g) \mid$ Lecture 30
$g \in G\}$, and the *kernel* of θ is $\ker(\theta) = \{g \in G \mid \theta(g) = 1\}$. Notice that this definition

of kernel is subtly different from the definition of the kernel of a *ring* homomorphism! One way of explaining the reason for this is that we want the image of the kernel (which consists of a single element) to be a *subgroup* of H, which forces it to be $\{1_H\}$. Similarly, in the case of rings, we want the image of the kernel (of a ring homomorphim $\phi : R \to S$) to be a *subring* of S, which forces it to be $\{0_S\}$.

Indeed, if $\theta : G \to H$, then $\mathrm{im}(\theta)$ is always a subgroup of H, since if $a, b \in \mathrm{im}(\phi)$, then $a = \theta(x)$ and $b = \theta(y)$ for some $x, y \in G$, and we have $ab^{-1} = \theta(x)\theta(y)^{-1} = \theta(x)\theta(y^{-1}) = \theta(xy^{-1}) \in \mathrm{im}(\theta)$. Similarly, $\ker(\theta)$ is a subgroup of G, since if $x, y \in \ker(\theta)$ then $\theta(x) = \theta(y) = 1$, so $\theta(y^{-1}) = \theta(y)^{-1} = 1^{-1} = 1$, and then $\theta(xy^{-1}) = \theta(x)\theta(y^{-1}) = 1.1 = 1$, so $xy^{-1} \in \ker(\theta)$.

Normal subgroups. A subgroup H of a group G is called *normal* if the left and right cosets of H in G are the same, that is, if $Hg = gH$ for every $g \in G$. The normal subgroups play the same role in group theory that the ideals play in ring theory.