Mathematics and Algorithms for Computer Algebra

July 9, 2003

2: Introduction to Abstract Algebra

The notes this week are based on several chapters of the very nice book by Lipson, which now unfortunately appears to be out of print. The first two sections of these notes provide a rapid summary of some of the basic notions of pure mathematics, as a reminder and in order to fix the notation and nomenclature that I will use. Subsequent sections summarize the properties of the computational domains of main interest for computer algebra. We will return in subsequent weeks to consider in more concrete detail several of the topics introduced this week.

1 Sets, relations and functions

1.1 Some fundamental sets

$\mathbb{Z} = \{0, \pm1, \pm2, \ldots\}$ is the set of integers.
$\mathbb{Q} = \{m/n \mid m, n \in \mathbb{Z}; n \neq 0\}$ is the set of rational numbers.
$\mathbb{R}$ is the set of real numbers.
$\mathbb{C} = \{a + ib \mid a, b \in \mathbb{R}\}$ is the set of complex numbers ($i = \sqrt{-1}$).

Some important subsets of these are the following:
$\mathbb{Z}^+ = \{a \in \mathbb{Z} \mid a > 0\}$ is the set of positive integers – the positive rationals and reals are denoted similarly.
$\mathbb{N} = \{a \in \mathbb{Z} \mid a \geq 0\}$ is the set of natural numbers.[1]

Note that $\mathbb{Z}^+ \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ (with strict inclusions).

[1] It could be argued that 0 is not a natural number, and some authors call $\mathbb{Z}^+$ the natural numbers, but in computing it is conventional and convenient to count from 0, and it turns out that in our present context $\mathbb{N}$ is more natural than $\mathbb{Z}^+$.

The *Cartesian product* $A \times B$ of sets A and B is the set of all *ordered pairs* thus

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

Examples: $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ is the Cartesian plane;
$\{0, 1\} \times \{a, b, c\} = \{(0, a), (0, b), (0, c), (1, a), (1, b), (1, c)\}$.

1.2 Equivalence relations

A *relation* "$\equiv$" between a set A and a set B is the subset of $A \times B$ defined by $\{(a, b) \in A \times B \mid a \equiv b\}$. A relation $\equiv$ on a set A (i.e. between A and A) is an *equivalence relation* if $\forall a, b, c \in A$ it is

1. Reflexive: $a \equiv a$
2. Symmetric: $a \equiv b \ \Rightarrow \ b \equiv a$
3. Transitive: $a \equiv b$ and $b \equiv c \ \Rightarrow \ a \equiv c$

Examples: Equality ($=$) is an equivalence (but inequality ($\neq$) is not, because (1) and (3) are not satisfied).

The subset $[a]_{\equiv} = \{b \in A \mid a \equiv b\}$ is called the $\equiv$-equivalence class of a. The set of all $\equiv$-equivalence classes in A is denoted $A/{\equiv} = \{[a] \mid a \in A\}$ and is called either the *quotient set* of A by $\equiv$ or A *modulo* $\equiv$ (abbreviated to $A \bmod \equiv$).

Lemma 1

$$[a] = [b] \ \Rightarrow \ a \equiv b$$

Proof is an exercise! □

1.2.1 Equivalence mod m on $\mathbb{Z}$

For m a positive integer, define

$$a = b \pmod{m} \ \text{ or } \ a \equiv_m b \iff a - b = km \text{ for some } k \in \mathbb{Z}.$$

Thus $[a]_m = \{a + km \mid k \in \mathbb{Z}$. For example, with $m = 3$,

$$\begin{aligned} [0]_3 &= \{\ldots, -6, -3, 0, 3, 6, \ldots\}, \\ [1]_3 &= \{\ldots, -5, -2, 1, 4, 7, \ldots\}, \\ [2]_3 &= \{\ldots, -4, -1, 2, 5, 8, \ldots\}, \end{aligned}$$

and hence $[0]_3 \cup [1]_3 \cup [2]_3 = \mathbb{Z}$. Generally, $\mathbb{Z}/{\equiv_m} = \{[0]_m, [1]_m, \ldots, [m-1]_m\}$.

1.2.2 Rational numbers

Let $X = \mathbb{Z} \times (\mathbb{Z} - \{0\}) = \{(a, b) \mid a, b \in \mathbb{Z}; b \neq 0\}$, and define the relation $\sim \subseteq X$ by $(a, b) \sim (c, d) \iff ad = bc$. Then $[(a, b)]_\sim$ represents all equivalent rational numbers of the form $a/b = (ka)/(kb) \forall k$.

Philosophy: The set $A/\equiv$ is simpler than A because it represents *subsets* of elements of A as single elements of $A/\equiv$. Frequently A is infinite but $A/\equiv$ is finite.

1.2.3 Partial and total orderings

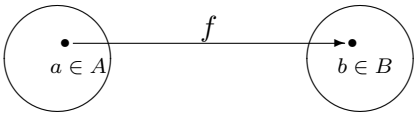
A relation $\leq$ on a set P is a *partial order* if $\forall x, y \in P$ it is

1. Reflexive: $x \leq x$
2. Antisymmetric: $x \leq y$ and $y \leq x \ \Rightarrow \ x = y$
3. Transitive: $x \leq y$ and $y \leq z \ \Rightarrow \ x \leq z$

A set P with a partial order $\leq$ is called a *partially ordered set*. It is *totally ordered* if all elements $x, y \in P$ satisfy *either* $x \leq y$ *or* $y \leq x$, i.e. $x \not\leq y \ \Rightarrow \ y \leq x$.

Examples: $\mathbb{Z}$ is totally ordered by the usual meaning of $\leq$; the positive integers $\mathbb{Z}^+$ are partially ordered by divisibility ($\mid$), but not totally ordered because $2 \nmid 3 \nRightarrow 3 \mid 2$.

1.3 Functions or maps



$A =$ domain of f $\qquad$ $B =$ codomain of f

If $f : A \to B$ is a function or map then it must map $a \mapsto b = f(a)$ for *all* $a \in A$, i.e. f must assign a *unique* image or value $b \in B$ to every $a \in A$. A function $f : A \to B$ defines a relation on $A \times B$, which is its graph. If $A' \subset A$ then $f' : A' \to B$ is a *partial* function on A (e.g. $\surd : \mathbb{R}_+ \to \mathbb{R}$, where $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$ is the set of non-negative real numbers).

If $f, g : A \to B$ then $f = g \iff f(a) = g(a) \forall a \in A$.

The function $f : A \to B$ is:

1. injective or one-to-one if $a \neq b \Rightarrow f(a) \neq f(b)$ (or if $f(a) = f(b) \Rightarrow a = b$);
2. surjective or onto if $\forall b \in B, b = f(a)$ for some $a \in A$;
3. bijective if it is both injective and surjective.

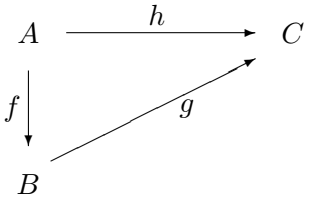
The *image* of $S \subseteq A$ under $f : A \to B$ is

$$f(S) = \{b \in B \mid b = f(s) \text{ for some } s \in S\}.$$

$f(A) \subseteq B$ is called *the image* of f, Im f, or the *range* of f. The *inverse image* of $T \subseteq B$ under $f : A \to B$ is

$$f^{-1}(T) = \{a \in A \mid f(a) \in T\}.$$

The *composition* of $f : A \to B$ and $g : B \to C$ is denoted $g \circ f : A \to C$ and defined by $g \circ f\ (a) = g(f(a))$, which means apply the functions from right to left. Generally function composition does not commute! Function composition is sometimes displayed in a *mapping diagram* of the form



which means that $h : A \to C$ is the composition of $f : A \to B$ with $g : B \to C$, i.e. $h = g \circ f$. (Note the order!)

1.3.1 Inverse functions

$f : A \to B$ is

1. left invertible if $\exists g : B \to A$ such that $g \circ f = 1_A$ (where 1_A is the identity function on A);
2. right invertible if $\exists h : B \to A$ such that $f \circ h = 1_B$;
3. (two-sided) invertible is it is both left and right invertible.

Theorem 2 *If $f : A \to B$ has a left inverse $g : B \to A$ and a right inverse $h : B \to A$ then $g = h$.*

This *unique* two-sided inverse of f is denoted f^{-1}.

Theorem 3

1. *f is left invertible $\iff$ f is injective.*
2. *f is right invertible $\iff$ f is surjective.*

Corollary 4 *f is (two-sided) invertible $\iff$ f is bijective.*

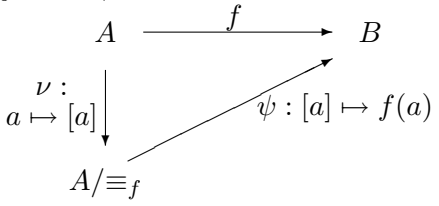
1.3.2 Functions and equivalence relations

An equivalence relation $\equiv$ on a set A naturally induces a function on A, $\nu : A \to A/\equiv, a \mapsto [a]_\equiv$, called the *natural map* of $\equiv$.

Conversely, any function on A naturally induces an equivalence relation on A, $a \equiv_f b \iff f(a) = f(b)$, called the *kernel relation* of f.

1.3.3 Decomposition theorem for functions

Any function $f : A \to B$ can be expressed as the composition $\psi \circ \nu$ of a surjection ν and an injection ψ:



Corollary 5 *f is surjective $\Rightarrow$ ψ is bijective.*

2 The integers

As we have seen, the integers underlie all of computer algebra, and much of abstract algebra is concerned with generalizing properties of the integers. $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ share the same behaviour under the arithmetic operations $+$, $-$ and $\times$, and are totally ordered (by $\leq$), so what distinguishes $\mathbb{Z}$ as fundamental?

Definition 1 *A totally ordered set $[P; \leq]$ is* well-ordered *(by $\leq$) if every nonempty subset of P has a least element.*

Then the natural numbers $\mathbb{N} = \{a \in \mathbb{Z} \mid a \geq 0\}$ are well-ordered by $\leq$. This leads to

Theorem 6 (Induction property of $\mathbb{N}$) *If $S \subseteq \mathbb{N}$ satisfies*

1. $0 \in S$
2. $n \in S \Rightarrow n+1 \in S$

then $S = \mathbb{N}$.

This is the basis of *proof by induction* and *definition by recursion.*

2.1 Integer division

For $a, m \in \mathbb{Z}, m \neq 0$, there exists a unique *quotient* q and *remainder* r such that

$$a = mq + r \quad (0 \leq r < |m|).$$

This relation is important, so here are some examples, with the elements in the same order as in the general statement above in each case:

$$7 = 3.2 + 1 \quad (0 \leq 1 < |3|);$$

$$3 = 7.0 + 3 \quad (0 \leq 3 < |7|);$$

$$-7 = 3.(-2) + 2 \quad (0 \leq 2 < |3|);$$

$$3 = (-7).0 + 3 \quad (0 \leq 3 < |-7|).$$

Remark to computer programmers: The above is related to, but not the same as, integer division in most programming languages, in which integer division usually *truncates toward zero*, so that for example `q = a/m` in FORTRAN is defined so that

$$a = mq + r \quad (0 \leq \text{sign}(a).r < |m|).$$

In the division relation, r is called the *remainder mod m of a* and denoted $a \bmod m$ or $r_m(a)$. If $r_m(a) = 0$ then m *divides* a, written $m \,|\, a$. Thus

$$m \,|\, a \iff a = km \text{ for some } k \in \mathbb{Z}.$$

Examples: $3 \,|\, 6$, $3 \,|\, -3$, $3 \nmid 7$.

2.2 Equivalence and remainders mod m

$$a \equiv_m b \iff m \,|\, (a-b) \iff r_m(a-b) = 0.$$

Lemma 7 *Let $m \in \mathbb{Z}^+$. Then*

1. $a \equiv_m r_m(a)$
2. $a = r_m(a) \iff 0 \leq a < m$
3. $a \equiv_m b \iff r_m(a) = r_m(b)$

Define the "integers mod m" to be $\mathbb{Z}_m = \{0, 1, \ldots, m-1\}$.

Theorem 8

1. *For any $[a] \in \mathbb{Z}/\equiv_m$, $[a] = [r_m(a)]$;*
2. *For any $a, b \in \mathbb{Z}_m$, $a \neq b \Rightarrow [a] \neq [b]$.*

Hence each *equivalence class* $[a]$ has a *unique representative* in $\mathbb{Z}_m$, namely $r_m(a)$.

2.3 The Greatest Common Divisor (GCD)

Divisibility is not affected by sign, so consider only the natural numbers $\mathbb{N} = \{a \in \mathbb{Z} \mid a \geq 0\}$. The integer 12 is divisible by 1, 2, 3, 4, 6 and $12 \in \mathbb{N}$. Similarly, 18 is divisible by 1, 2, 3, 6, 9 and $18 \in \mathbb{N}$. Thus 12 and 18 have the *common* divisors 1, 2, 3, $6 \in \mathbb{N}$, and hence *the greatest common divisor* 6.

Generally, the gcd $g \in \mathbb{N}$ of $a, b \in \mathbb{Z}$ (not both 0) may be defined by

1. $g \,|\, a$ and $g \,|\, b$ (*common* divisor);
2. $c \,|\, a$ and $c \,|\, b \Rightarrow c \,|\, g$ (*greatest* common divisor).

Proposition 9 *The (positive) gcd of $a, b \in \mathbb{Z}$ is unique.*

Proof Suppose g and g' are both gcds of a and b, then $g \,|\, g'$ since g' is a greatest common divisor and g is a common divisor. But symmetrically $g' \,|\, g$. Hence $g' = \pm g$, and since $g, g' > 0$, $g' = g$. □

The unique (positive) gcd of $a, b \in \mathbb{Z}$ is denoted $\gcd(a, b)$.

The following theorem is important.

Theorem 10 *Let $a, b \in \mathbb{N}$ be not both zero. Then*

$$\gcd(a, b) = sa + tb$$

for some $s, t \in \mathbb{Z}$.

In words, a gcd can be expressed as a ($\mathbb{Z}$-)linear combination, which is (I think) not obvious.

Examples:

$$\gcd(3, 7) = 1 = (-2).3 + 1.7$$

$$\gcd(15, 9) = 3 = 2.15 + (-3).9$$

The following proof uses the well-orderedness of $\mathbb{N}$ and the division property of $\mathbb{Z}$.

Proof Let $S(a, b) = \{sa + tb \mid s, t \in \mathbb{Z}\}$. We aim to show that $\gcd(a, b) \in S(a, b)$. Clearly $a = 1a + 0b$ and $b = 0a + 1b$ are both in $S(a, b)$, so $S(a, b)$ contains (at least two) positive integers. Consider the subset of $S(a, b)$ consisting of its positive elements. This subset is also a subset of $\mathbb{N}$ and is non-empty, hence by the well-orderedness of $\mathbb{N}$ it contains a least element. Denote this least *positive* element of $S(a, b)$ as $g = s_g a + t_g b$. Now we prove that $g = \gcd(a, b)$.

Lemma 11 *$S(a, b) = (g) = \{kg \mid k \in \mathbb{Z}\}$.*

Proof of lemma:

1. $(g) \subseteq S(a, b)$, because any $kg \in (g)$ satisfies $kg = k(s_g a + t_g b) = (ks_g)a + (kt_g)b \in S(a, b)$.

2. $S(a, b) \subseteq (g)$, because for any $x = s_x a + t_x b \in S(a, b)$ we can use the division property of $\mathbb{Z}$ to write

   $$x = gq + r, \quad (0 \leq r < g).$$

   Hence

   $$\begin{aligned} r &= x - gq \\ &= (s_x a + t_x b) - (s_g a + t_g b)q \\ &= (s_x - s_g q)a + (t_x - t_g q)b \in S(a, b). \end{aligned}$$

   But then $r = 0$, because g was defined to be the least positive element of $S(a, b)$. Thus $x = gq \in (g)$, so $S(a, b) \subseteq (g)$.

$(g) \subseteq S(a,b)$ and $S(a,b) \subseteq (g) \;\Rightarrow\; S(a,b) = (g)$. □

To complete the proof of the theorem, note that $a, b \in S(a,b)$ and so $a, b \in (g)$ by the lemma, so $g \,|\, a$ and $g \,|\, b$ and therefore g is a *common* divisor of a and b. Moreover, if $c \,|\, a$ and $c \,|\, b$ then (trivially) $c \,|\, (g = s_g a + t_g b)$, so g is a *greatest* common divisor, i.e. $g = \gcd(a,b)$. □

Remark 1: Later we will consider algorithms to compute $\gcd(a,b)$, s and t.

Remark 2: (g) is an ideal, actually a principal ideal, of the ring $\mathbb{Z}$, to which we will also return later.

2.4 Prime factorization

Definition 2 *An integer $n \geq 2$ is called* prime *if it has no proper divisors, i.e. its only (positive) divisors are 1 and n. An integer $n \geq 2$ that is not prime is called* composite*.*

Theorem 12 (Existence of prime factorization) *Any integer $n \geq 2$ can be expressed as a product of primes.*

Lemma 13 *Let p be a prime. Then*

$$p \,|\, ab \;\Rightarrow\; p \,|\, a \quad \textit{or} \quad p \,|\, b.$$

Corollary 14 $p \,|\, \prod_i a_i \;\Rightarrow\; p \,|\, a_i$ *for some* i.

Definition 3 *If $\gcd(a,b) = 1$ then a and b are called* relatively prime*.*

Theorem 15 (Uniqueness of prime factorization) *Every integer $n \geq 2$ can be expressed as a product of primes uniquely apart from reordering of factors.*

Corollary 16 *Any integer $n \geq 2$ has a unique "prime decomposition" of the form*

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r} \quad (e_i \geq 1)$$

where the p_i are distinct prime factors of n.

Proof The two main theorems above can be proved by induction on n. □

3 Groups

I want to emphasize the way that complex mathematical structures are built from simpler ones, i.e. their *hierarchical* nature. In particular, group properties underlie the properties of the rings and fields that are the structures of most direct importance in CA.

An *algebra* is a very general structure consisting of a set A together with operations on the set. If an operation maps $A^n \to A$ then it is said to have *arity* n. A *binary* algebra has a single binary (arity 2) operation, perhaps together with operations of lower arity.

Definition 4 *A* groupoid $[G; \cdot]$ *is an algebra with a binary operation* $\cdot : (x, y) \mapsto x \cdot y$ *or* xy*. It is called* commutative *if* $xy = yx$*.*

An *additive* groupoid, in which the binary operation is written as $+$, is always commutative.

Definition 5 *A* semigroup $[S; \cdot]$ *is a groupoid that is* associative*:* $x(yz) = (xy)z$ *(or* $x + (y + z) = (x + y) + z$*).*

Definition 6 *A* monoid $[M; \cdot, 1]$ *is an algebra with a binary operation* $\cdot$ *and a nullary (arity 0) operation 1 (whose value is always 1) such that*

1. $[M; \cdot]$ *is a semigroup (associativity),*
2. $x1 = 1x = x$ *(identity).*

For a multiplicative monoid $[M; \cdot, 1]$ *the identity is called a* unit element*; for an additive monoid* $[M; +, 0]$ *the identity is called a* zero element*.*

Definition 7 *Finally, a* group $[G; \cdot, ^{-1}, 1]$ *is an algebra with one binary operation* $\cdot$*, one unary (arity 1) operation* $^{-1}$ *(inversion) and one nullary operation 1, such that*

1. $[G; \cdot, 1]$ *is a monoid (associativity, identity),*
2. $xx^{-1} = x^{-1}x = 1$ *(inverse).*

An additive (which implies commutative or "Abelian") group is written $[G; +, -, 0]$ where $-x$ is the (additive) inverse of x.

In a multiplicative group $[G; \cdot, ^{-1}, 1]$, *division* is defined by $a/b = ab^{-1} (\neq b^{-1}a$ generally) and, in an additive group $[G; +, -, 0]$, *subtraction* is defined by $a - b = a + (-b)$ $[= (-b) + a]$.

Examples:

1. $[\mathbb{N}; +, -, 0]$: the natural numbers form an additive group. However, $[\mathbb{N}; -]$ is not even a semigroup because subtraction is not associative – it is *only* a groupoid.

2. Similarly, $[\mathbb{Z}; +, -, 0]$ and $[\mathbb{R}; +, -, 0]$ are additive groups.

3. $[\mathbb{Q}^+; \cdot, ^{-1}, 1]$: the *positive* rationals form a multiplicative group, as does $[\mathbb{R}^*; \cdot, ^{-1}, 1]$, the *non-zero* reals.

4. The multiplicative monoids $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}_m$ are *not* groups because 0 does not have a multiplicative inverse.

3.1 Group properties of $\mathbb{Z}_m$ (the integers mod m)

The following are *semigroups* (i.e. associative):

1. $[\mathbb{Z}_m; \oplus]$ where $a \oplus b = r_m(a + b)$ or $(a + b) \bmod m$;

2. $[\mathbb{Z}_m; \odot]$ where $a \odot b = r_m(ab)$ or $ab \bmod m$.

Moreover, $[\mathbb{Z}_m; \oplus, \ominus, 0]$ is an additive *group*, where

$$\ominus a = \begin{cases} m - a & \text{if } a \neq 0, \\ 0 & \text{if } a = 0. \end{cases}$$

However, as remarked earlier, the multiplicative monoid $[\mathbb{Z}_m; \odot, 1]$ is not a group because 0 has no multiplicative inverse.

So is $\mathbb{Z}_m^* = \mathbb{Z}_m - \{0\}$ a group? Consider $\mathbb{Z}_4^* = \{1, 2, 3\}$. Then $2 \odot 2 = 4 \bmod 4 = 0 \notin \mathbb{Z}_4^*$, so $[\mathbb{Z}_4^*; \odot]$ is not *even* a groupoid because $\odot$ is not defined on all elements of $\mathbb{Z}_4^*$, i.e. the algebra is not *closed*. However:-

Theorem 17 *Under mod m multiplication, $\mathbb{Z}_m^*$ is a multiplicative group $\iff$ m is prime.*

This result is important, and we should have a few proofs in this course, so let us prove this theorem.

Proof

($\Rightarrow$): m composite $\Rightarrow$ $m = st$, $1 < s, t < m$. Then $s, t \in \mathbb{Z}_m^*$ but $s \odot t = st \bmod m = m \bmod m = 0 \notin \mathbb{Z}_m^*$. Thus m not prime $\Rightarrow$ $\mathbb{Z}_m^*$ not a group, so $\mathbb{Z}_m^*$ a group $\Rightarrow$ m prime.

($\Leftarrow$): This is harder, but illustrates well the use of the hierarchical structure. Let m be a prime p.

Closure under $\odot$. Let $a, b \in \mathbb{Z}_p^*$. We must show that $a \odot b \in \mathbb{Z}_p^*$, i.e. $a \odot b \neq 0$.

$$\begin{aligned} a \odot b = 0 \quad &\Rightarrow \quad ab \bmod p = 0 \\ &\Rightarrow \quad p \,|\, ab \\ &\Rightarrow \quad p \,|\, a \text{ or } p \,|\, b \end{aligned}$$

But $1 \leq a, b < p$ so this is impossible. Hence $a \odot b \neq 0$ and $[\mathbb{Z}_p^*; \odot]$ is closed and hence a *groupoid*.

Associativity of $\odot$.

$$\begin{aligned} r_m(a) r_m(b) \quad &= \quad (a + k_a m)(b + k_b m), \quad k_b, k_b \in \mathbb{Z} \\ &= \quad ab + (k_a b + k_b a + k_a k_b m) m \\ &\equiv_m \quad ab \end{aligned}$$

Thus $r_m(ab) = r_m(r_m(a) r_m(b))$ (*)

Then

$$\begin{aligned} a \odot (b \odot c) &= r_m(a r_m(bc)) && \text{defn. of } \odot \\ &= r_m(r_m(a) r_m(bc)) && a \in \mathbb{Z}_m \\ &= r_m(abc) && \text{by (*)} \\ &= r_m(r_m(ab) r_m(c)) && \text{assoc. of } \cdot \text{ and (*)} \\ &= (a \odot b) \odot c && \text{defn. of } \odot \end{aligned}$$

Associativity of $[\mathbb{Z}_m; \odot] \Rightarrow$ associativity of $[\mathbb{Z}_p^*; \odot]$, i.e. both are *semigroups*.

Multiplicative identity. $1 \in \mathbb{Z}_p^*$ is an identity element with respect to $\odot$, hence $[\mathbb{Z}_p^*; \odot, 1]$ is a *monoid*.

Multiplicative inverses. Let $a \in \mathbb{Z}_p^*$. Since p is prime and $1 \leq a < p$ then

$$1 = \gcd(a, p) = sa + tp$$

for some $s, t \in \mathbb{Z}$ (from the previous theory of integer gcds). Taking remainders mod p gives

$$\begin{aligned} r_p(1) = 1 &= r_p(sa + tp) \\ &= r_p(sa) && \text{defn. of } r_p \\ &= r_p(r_p(s) r_p(a)) && \text{by (*) above} \\ &= r_p(s) \odot r_p(a) && \text{defn. of } \odot \text{ above} \\ &= r_p(s) \odot a && a \in \mathbb{Z}_p \end{aligned}$$

Hence $a^{-1} = r_p(s) \in \mathbb{Z}_p^*$ and so $[\mathbb{Z}_p^*; \odot, ^{-1} \bmod p, 1]$ is a *group*. $\square$

3.2 Order of a group element

Definition 8 *The* order*, $o(a)$, of an element $a \neq 1$ in a multiplicative group $[G; \cdot, ^{-1}, 1]$ is defined by*

$$o(a) = \begin{cases} n & \text{if } a^n = 1 \text{ and } a^k \neq 1 \text{ for } 1 \leq k < n, \\ \infty & \text{if } a^n \neq 1 \text{ for any } n \in \mathbb{Z}^+. \end{cases}$$

Equivalently, the order of an element $a \neq 0$ in an additive group $[G; +, -, 0]$ is defined by

$$o(a) = \begin{cases} n & \text{if } na = 0 \text{ and } ka \neq 0 \text{ for } 1 \leq k < n, \\ \infty & \text{if } na \neq 0 \text{ for any } n \in \mathbb{Z}^+. \end{cases}$$

3.3 Subalgebras

A subset B of an Ω-algebra A is called a subalgebra of A if B is closed under Ω (Ω-closed), meaning that for all $\omega \in \Omega$

$$x_1, \ldots, x_n \in B \;\Rightarrow\; \omega(x_1, \ldots, x_n) \in B,$$

where n is the arity of ω. A subalgebra is denoted $B \leq A$ if $B \subseteq A$ and $B < A$ if $B \subset A$.

Let A be an Ω-algebra and $H \subseteq A$. Denote by $[H]$ the smallest subalgebra of A that includes H. If $[H] = A$ then A is said to be *generated by* H. If H is finite then A is said to be *finitely generated by* H.

3.4 Subgroups

A subset S of an (additive) group $[G; +, -, 0]$ is a subgroup of G if:

1. $x, y \in S \;\Rightarrow\; x + y \in S$;
2. $x \in S \;\Rightarrow\; -x \in S$;
3. $0 \in S$.

Proposition 18 *Let S be a nonempty subset of $[G; +, -, 0]$. Then S is a subgroup of G if*

$$x, y \in S \;\Rightarrow\; x - y \in S. \qquad (*)$$

Proof Since $S \neq \varnothing$ $\exists x \in G$ in S, so by (*) $x - x = 0 \in S$, and hence $0 - x = -x \in S$. If $x, y \in S$, then $-y \in S \;\Rightarrow\; x - (-y) = x + y \in S$. $\square$

4 Morphisms

A *morphism* is a structure-preserving map from one algebraic system (algebra) to another, and morphisms underlie many of the more sophisticated algorithms of CA.

Let $[A;\Omega]$ be an algebra consisting of a set A (the "carrier") together with a collection Ω of operations.

Definition 9 *Two algebras* $[A;\Omega]$ *and* $[A';\Omega']$ *are* similar *if there is a bijection between* Ω *and* Ω' *such that corresponding operations* $\omega \in \Omega$ *and* $\omega' \in \Omega'$ *have the same arity.*

Note that this bijection relates the operations and not the carrier sets. Hence *any* two groupoids are similar *by definition.* (In practice, it is often convenient to use the same collection of operator symbols Ω for both algebras, only distinguishing the carrier sets, and so to speak of similar Ω-algebras A and A'.)

Definition 10 *Let* $[A;\Omega]$ *and* $[A';\Omega']$ *be similar algebras. A map* $\phi : A \to A'$ *is called a* morphism *(or* homomorphism*) from* $[A;\Omega]$ *to* $[A';\Omega']$ *if, for every* $\omega \in \Omega$ *and* $a_1, \ldots, a_n \in A$*,*

$$\phi(\omega(a_1, \ldots, a_n)) = \omega'(\phi(a_1), \ldots, \phi(a_n)).$$

This means that one can perform an operation in A and then map the result to A', or one can map the arguments to A' and then apply the corresponding operation in A', and either way the result is the same. Diagrammatically, if ω has arity n,

$$\begin{array}{ccc} A^n & \xrightarrow{\omega} & A \\ {\scriptstyle \phi^n}\downarrow & & \downarrow{\scriptstyle \phi} \\ (A')^n & \xrightarrow[\omega']{} & A' \end{array}$$

where

$$\phi^n : A^n \to (A')^n, \quad (a_1, \ldots, a_n) \mapsto (\phi(a_1), \ldots, \phi(a_n)).$$

Example: A morphism ϕ from a (multiplicative) group $[G;\cdot,{}^{-1},1]$ to an (additive) group $[G';+,-,0]$ must be such that for all $x, y \in G$

1. $\phi(x \cdot y) = \phi(x) + \phi(y)$,
2. $\phi(x^{-1}) = -\phi(x)$,
3. $\phi(1) = 0$

(although in fact requirements 2 and 3 are consequences of 1 and the definition of a group).

4.1 Special morphisms

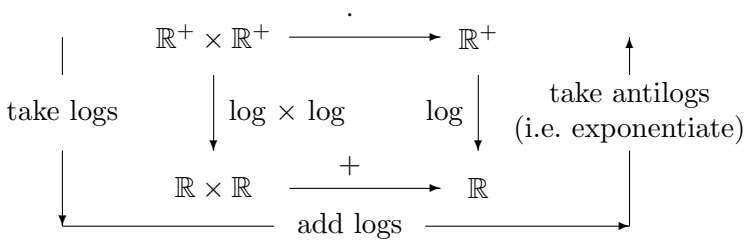
A morphism $\phi : A \to A'$ is called:

1. an *isomorphism* if ϕ is bijective (invertible);
2. an *epimorphism* if ϕ is surjective (onto);
3. an *monomorphism* if ϕ is injective (1-to-1).

4.1.1 Isomorphisms

If A and A' are *isomorphic*, denoted $A \cong A'$, then they are identical except for the names (symbols) used.

Example: Let $\phi : [\mathbb{R}^+; \cdot] \to [\mathbb{R}; +], x \mapsto \log x$. Then ϕ is an isomorphism, which underlies the standard technique of using logarithms to multiply, thus:



4.1.2 Epimorphisms

These are probably the most important morphisms in practice for CA.

If $\phi : A \to A'$ is an epimorphism then A' is called a *homomorphic image* of A, and is regarded as an *abstraction* or *model* of A. The implication is that ϕ is not also an injection, and so A' is smaller than A, but nevertheless captures some of the structure of A.

If $\phi : b \mapsto b'$ then by computing $b' = \omega(\phi(a_1), \ldots, \phi(a_n)) \in A'$ we do not determine $b \in A$ completely, but we know that $b \in \phi^{-1}(b')$. Thus we have found not b but an equivalence class in A that contains b, i.e. we know b modulo the kernel relation of ϕ. [See the earlier notes on equivalence relations and functions.]

Examples:

1. The "remainder mod m" map, $a \mapsto r_m(a)$, is an epimorphism (of additive groups) from $[\mathbb{Z}; +, -, 0]$ to $[\mathbb{Z}_m; \oplus, \ominus, 0]$. This is one of the main uses of epimorphisms in CA; note that whereas $\mathbb{Z}$ is infinite, $\mathbb{Z}_m$ is finite and hence (much) simpler!

2. The projection maps $p_1 : (a, b) \mapsto a,\ p_2 : (a, b) \mapsto b$ are epimorphisms (of additive groups) from points in the plane $\mathbb{R}^2$ to points on the line $\mathbb{R}$.

Projection always implies a loss of information, so that epimorphisms always correspond to projections in some general sense. One normally has the mental image of projection downwards (perhaps by shining a light from above and observing a shadow), and therefore any attempt to undo a projection or an epimorphism is normally called *lifting*. We will return to lifting later.

4.1.3 Monomorphisms

If $\phi : A \to A'$ is a monomorphism then $\phi : A \to \phi(A) \subseteq A'$ is clearly an isomorphism, so A' includes what is essentially a copy of A and $\phi : A \to A'$ is sometimes called an *embedding* of A into A'.

4.1.4 Endomorphisms and automorphisms

A morphism $\phi : A \to A$ that maps an algebra to itself is called an *endomorphism* of A. If ϕ is also a bijection and hence an isomorphism it is called an *automorphism* of A.

4.1.5 Structure-preserving properties of morphisms

Epimorphisms preserve semigroup, monoid and group structure, so that for example a homomorphic image of a group is a group. Morphisms in general preserve subalgebras and their generators.

5 Rings

Groups have a single binary operation, whereas rings (and fields) have two: addition *and* multiplication.

Definition 11 *A* ring $[R; +, -, 0, \cdot, 1]$ *is an algebra such that*

1. $[R; +, -, 0]$ *is an abelian (commutative) group (the "additive group" of* R*);*
2. $[R; \cdot, 1]$ *is a monoid (the "multiplicative monoid" of* R*);*
3. $a(b + c) = ab + ac$, $(b + c)a = ba + ca$ *(multiplication* $(\cdot)$ *is left and right distributive over addition* $(+)$*, which is the only link between* $+$ *and* $\cdot$*).*

A ring is called *commutative* if its multiplication commutes. (The addition *must* be commutative.) The multiplicative identity element 1 is called the *unity, unit element* or *one*, but the term *unit* alone has a special meaning, and usually 1 is just one of several units in a ring. Do not confuse R for ring with $\mathbb{R}$ for reals!

Examples: $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all rings under the usual operations $+, -, 0, \cdot, 1$, and form a hierarchy of (proper) subrings $\mathbb{Z} < \mathbb{Q} < \mathbb{R} < \mathbb{C}$. $\mathbb{Z}_m$ is a ring under mod m versions of $+, -, \cdot$, where the "mod m" is usually implied by working in $\mathbb{Z}_m$.

A *derived ring* is a ring built on, or defined in terms of, an existing "ground" ring, as in the following examples.

5.1 Formal power series

Let R be a commutative (ground) ring (e.g. $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{Z}_m$). Define $R[[x]]$ to be the set of all expressions of the form

$$a(x) = a_0 + a_1 x + a_2 x^2 + \cdots = \sum_{i=0}^{\infty} a_i x^i,$$

where the *coefficients* a_i $(i \in \mathbb{N})$ lie in the ground ring R. Since x is a symbol with no value the summation cannot actually be performed and convergence is not an issue. Hence each such $a(x)$ is called a *formal power series over* R *in the indeterminate* x.

$R[[x]]$ is endowed with a ring structure by defining operations $+, -, \cdot, 0\,[= 0(x)], 1\,[= 1(x)]$ such that for all $i \in \mathbb{N}$:

$$\begin{array}{lcl}
\text{in } R[[x]] & & \text{in } R \text{ (the ground ring)} \\
c(x) = a(x) + b(x) & \iff & c_i = a_i + b_i \\
c(x) = a(x)b(x) & \iff & c_i = \sum_{j=0}^{i} a_j b_{i-j} \\
c(x) = -a(x) & \iff & c_i = -a_i \\
c(x) = 0 & \iff & c_i = 0 \\
c(x) = 1 & \iff & c_i = \begin{cases} 1 \text{ if } i = 0 \\ 0 \text{ otherwise} \end{cases}
\end{array}$$

5.2 Univariate polynomials

Define $R[x]$, the set of all polynomials over R in the indeterminate x, to be the subset of $R[[x]]$ consisting of *finite* sums ot terms. This is a subring of $R[[x]]$.

5.3 Bivariate polynomials

Since $R' = R[x]$ defined above is a ring, we can use it as the ground ring of a new polynomial ring in y, $R'[y] = (R[x])[y] = R[x][y]$. Similarly, we could construct $R[y][x]$, and in fact the two rings are identical (assuming all multiplication commutes) and are therefore often written more symmetrically as

$$R[x][y] = R[y][x] = R[x, y].$$

However, the distinction can be useful to indicate how the elements are regarded, and whether the degree of terms with respect to one particular variable or the total degree is considered important. For example,

$$\begin{array}{rcll}
a(x, y) & = & (5x^2 - 3x + 4)y + (2x^2 + 1) & \in \mathbb{Z}[x][y], \\
& = & (5y + 2)x^2 + (-3y)x + (4y + 1) & \in \mathbb{Z}[y][x], \\
& = & 5x^2y + (2x^2 - 3xy) + (4y) + 1 & \in \mathbb{Z}[x, y].
\end{array}$$

Note that the forms $R[x][y]$ and $R[y][x]$ correspond to the nested multinomial representation the we considered last week.

5.4 Matrix rings

The set $M_n(R)$ of all $n \times n$ matrices over a commutative ring R (i.e. with elements in R) is a ring under normal matrix arithmetic, but it is *always*

non-commutative (if $n > 1$) even though R is commutative, as trivially shown by many simple examples.

5.5 Integral multiples; characteristic of a ring

Definition 12 *An* integral multiple $n \cdot a$, $n \in \mathbb{Z}$ *of a ring element* $a \in R$ *is defined in terms of the additive group of* R *(alone) by*

$$\begin{aligned} n > 0: &\quad n \cdot a = a + a + \cdots + a \ (n \text{ terms}), \\ n < 0: &\quad n \cdot a = -(|n| \cdot a), \\ n = 0: &\quad 0 \cdot a = 0 \in R. \end{aligned}$$

It is independent of the multiplicative monoid structure of R ($ab \in R$ if $a, b \in R$).

Definition 13 *The* characteristic *of a ring* R, char R, *is the order of the unity element 1 in the additive group of* R *if that order is finite, and zero otherwise.*

Thus if R has *finite characteristic* m then $m \cdot 1 = 0$ and $n \cdot 1 \neq 0$ for $1 \leq n < m$. $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ all have characteristic zero; $\mathbb{Z}_m$ has characteristic m. A finite ring must have finite characteristic (otherwise it would not be an additive group), but the converse is false, e.g. $\mathbb{Z}_m[x]$ is an infinite ring with finite characteristic m.

5.6 Subrings and extension rings

If R' is a subring of R ($R' \leq R$) then also R is an *extension ring* of R' ($R \geq R'$). A subring $R' \leq R$ is a subset of R that is closed under all the ring operations $+, -, 0, \cdot, 1$ of R. But in fact a sufficient condition is that a subset S of a ring R is a subring of R if, for all $a, b \in S$,

$$a - b \in S, \quad ab \in S, \quad 1 \in S,$$

by using the earlier proposition on subgroups to express all the additive group conditions within the condition $a - b \in S$.

Example: Let $\alpha = \sqrt[3]{2}$, then

$$S = \{a + b\alpha + c\alpha^2 \mid a, b, c \in \mathbb{Q}\}$$

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

Proof $1 = 1 + 0\alpha + 0\alpha^2 \in S$. If $x = a + b\alpha + c\alpha^2$ and $y = d + e\alpha + f\alpha^2$ then by direct computation $x - y$ and xy are in S. □

The *subring generated* by any subset $S \subseteq R$ of a ring R is the smallest subring of R that includes S, and is denoted $[S]$.

The *unital subring* [1] is the smallest subring of any ring R, since every ring *must* have a unity element. In fact,

$$[1_R] = \{n \cdot 1_R \mid n \in \mathbb{Z}\},$$

where the subscript distinguishes the ring of which 1 is considered to be the unity element when necessary.

Proof By closure under $+, -, 0$ (additive group of R) it follows that $\{n \cdot 1_R \mid n \in \mathbb{Z}\} \subseteq [1_R]$. But $1_R = 1 \cdot 1_R \in \{n \cdot 1_R \mid n \in \mathbb{Z}\}$, which also satisfies the other subring conditions, and $[1_R]$ is the *smallest* subring containing 1_R, hence $[1_R] \subseteq \{n \cdot 1_R \mid n \in \mathbb{Z}\}$. □

Therefore $[1_R] = [0_R]$ and $[1] = \mathbb{Z}$.

5.7 Morphisms of rings

A homomorphic image of a (commutative) ring is a (commutative) ring.

Let E and R be commutative rings with $R \leq E$. For $\alpha \in E$, define the (α-)evaluation map $\phi_\alpha : R[x] \to E$ by

$$\phi_\alpha : a(x) \mapsto a(\alpha).$$

Then ϕ_α is a morphism of rings – the "evaluation morphism". If $E = R$ then ϕ_α is an *epi*morphism $R[x] \to R$ (i.e. onto) since any $a \in R$ is the image under ϕ_α of the constant polynomial $a \in R[x]$.

Note that the derived rings $R[x]$ and $R[[x]]$ include the ground ring R as a subring: $R < R[x] < R[[x]]$.

5.8 Ring adjunction and extension rings

Let R be a ring and E an extension ring of R. For some $\alpha \in E$, the smallest extension ring of R that includes α is the subring $[R \cup \{\alpha\}]$ (i.e. the subring generated by $R \cup \{\alpha\} \subseteq E$). It is denoted simply $R[\alpha]$ and called "R adjoined by α". This is the same notation as used for polynomials over R, for the following reason.

Proposition 19 *Let E and R be* commutative *rings with $E \geq R$, and let $\alpha \in E$. Then*

$$R[\alpha] = \{a(\alpha) \mid a(x) \in R[x]\}.$$

[That is, $R[\alpha]$ is the set of *all* polynomials over R evaluated at α or equivalently, regarding α as a symbol, all polynomials in α.]

Proof Let $S = \{a(\alpha) \mid a(x) \in R[x]\}$. Any subring of E that includes $R \cup \{\alpha\}$ must contain any $a(\alpha) \in S$ by closure. (Polynomials only involve ring operations.) Hence $S \subseteq R[\alpha]$.

Now consider the evaluation morphism $\phi_\alpha : R[x] \to E,\ a(x) \mapsto a(\alpha)$. Because morphisms preserve subalgebras, $\phi_\alpha(R[x]) = S$ is a subring of E. Moreover, S contains each $a = \phi_\alpha(a)$ in R and $\alpha = \phi_\alpha(x)$. Therefore $R[\alpha] \subseteq S$, because $R[\alpha]$ by definition is the *smallest* subalgebra that contains each $a \in R$ and α. □

If $\alpha \in E$ satisfies a polynomial (i.e. is a root of a polynomial equation) over R then some simplification is possible in $R[\alpha]$. For example, if $\alpha = \sqrt[3]{2}$ then

$$\mathbb{Q}[\alpha] = \{a + b\alpha + c\alpha^2 \mid a, b, c \in \mathbb{Q}\}.$$

[That is, the ring of rationals adjoined by $\alpha = \sqrt[3]{2}$ is generated by all *quadratic* polynomials in α, and it is not necessary to use *all* polynomials in α. Thus, as a vector space over $\mathbb{Q}$, $\mathbb{Q}[\alpha]$ has dimension 3 (rather than infinity!), with a suitable basis being $\{1, \alpha, \alpha^2\}$.]

Proof Let $S = \{a + b\alpha + c\alpha^2 \mid a, b, c \in \mathbb{Q}\}$. Then $S \subseteq \mathbb{Q}[\alpha]$ by closure of the ring $\mathbb{Q}[\alpha]$. But S is a subring of $\mathbb{R}$ (proved above) that contains each $a = a + 0\alpha + 0\alpha^2 \in \mathbb{Q}$ and $\alpha = 0 + 1\alpha + 0\alpha^2$. Hence $\mathbb{Q}[\alpha] \subseteq S$, because $\mathbb{Q}[\alpha]$ by definition is the smallest subring of $\mathbb{R}$ that contains each $a \in \mathbb{Q}$ and α. □

6 Integral domains and fields

6.1 Zerodivisors and units

Let R be a commutative ring.

Definition 14 *If $a \neq 0, b \neq 0$ in R satisfy $ab = 0$ then both a and b are called* zerodivisors.

Definition 15 *If $u \neq 0, v \neq 0$ in R satisfy $uv = 1$ then both u and v are called* units*, and $u = v^{-1}, v = u^{-1}$.*

Example: In $\mathbb{Z}_8$, $2 \times 4 = 0$ so 2 and 4 are zerodivisors, and $3 \times 3 = 1$ so 3 is a unit.

The set of all units of a ring R is denoted $U(R)$, and forms a multiplicative group – the "group of units of R".

Proposition 20 *Let R be a commutative ring and $u \in R$ be a unit. Then u is not a zerodivisor.*

Proof Assume $uv = 0$ for some v. Then $v = u^{-1}uv = u^{-1}0 = 0$. □

Definition 16 *An* integral domain *is a nontrivial commutative ring with no zerodivisors, so that $ab = 0 \Rightarrow a = 0$ or $b = 0$.*

Theorem 21 *The ring D is an integral domain if and only if the following* cancellation law *holds:*

$$(ab = ac \text{ and } a \neq 0) \Rightarrow b = c.$$

Definition 17 *A* field *is a nontrivial commutative ring in which* every *nonzero element is a unit.*

Hence a field is an integral domain, because a unit is not a zerodivisor.

Fields are important because they allow *division* by all elements except 0, and moreover $ab^{-1} = ba^{-1}$ can be unambiguously written as $\frac{a}{b}$ because fields are commutative.

Examples: The prototypical integral domain is the ring of integers $\mathbb{Z}$. The only units are 1 and -1, hence $\mathbb{Z}$ is not a field. However, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all fields.

6.2 Prime fields

We proved earlier that for prime p, $\mathbb{Z}_p^*(= \mathbb{Z}_p - \{0\})$ is a multiplicative group, and hence every element has a multiplicative inverse and so is a unit in $\mathbb{Z}_p$ regarded as a ring. However, if m is composite and $m = ab$ $(1 < a, b < m)$ then $ab = 0$ in $\mathbb{Z}_m$. Thus we have proved

Theorem 22

1. *If p is prime then $\mathbb{Z}_p$ is a (finite) field.*

2. *If m is composite then $\mathbb{Z}_m$ is a ring* with *zerodivisors.*

There exist other finite fields. The ring of integers is an infinite integral domain that is not a field. However:

Theorem 23 *A finite integral domain is a field.*

Theorem 24 *If an integral domain has finite characteristic m then m is prime.*

6.3 Univariate polynomials and formal power series (again)

Definition 18 *If a polynomial has the form $a(x) = \sum_{i=0}^{n} a_i x^i$, $a_n \neq 0$, then its* degree *is defined to be* $\deg a(x) = n$*; otherwise* $\deg 0 = -\infty$.

Definition 19 *If a formal power series has the form $a(x) = \sum_{i=n}^{\infty} a_i x^i$, $a_n \neq 0$, then its* order *is defined to be* $\operatorname{ord} a(x) = n$*; otherwise* $\operatorname{ord} 0 = +\infty$.

Proposition 25

1. *In $R[x]$:*
$$\deg a(x)b(x) \leq \deg a(x) + \deg b(x),$$
$$\deg[a(x) + b(x)] \leq \max[\deg a(x), \deg b(x)].$$

2. *In $R[[x]]$:*
$$\operatorname{ord} a(x)b(x) \geq \operatorname{ord} a(x) + \operatorname{ord} b(x),$$
$$\operatorname{ord}[a(x) + b(x)] \geq \min[\operatorname{ord} a(x), \operatorname{ord} b(x)].$$

The definitions of $\deg 0$ and $\operatorname{ord} 0$ are designed to make the above proposition apply to zero polynomials.

Henceforth, let D denote an integral domain and F denote a field.

Theorem 26 *$D[x]$ and $D[[x]]$ are integral domains.*

Corollary 27 *In $D[x]$ and $D[[x]]$ the* equality *holds in the degree and order relations for products and sums (because there are no zerodivisors).*

Theorem 28 *$a(x)$ is a unit of $D[x]$ $\iff$ $a(x)$ is a unit of D.*

Proof $\Leftarrow$ is trivial. To prove $\Rightarrow$ let $a(x) \in D[x]$ be a unit. Then $a(x)b(x) = 1$ for some $b(x) \in D[x]$, so taking degrees gives $0 = \deg 1 = \deg a(x)b(x) = \deg a(x) + \deg b(x) \Rightarrow \deg a(x) = \deg b(x) = 0$, so $a(x) = a \in D$. Similarly for b, and $ab = 1 \Rightarrow a$ is a unit of D. □

Corollary 29 *The units of $F[x]$ are the nonzero constant polynomials of F.*

6.4 Ring morphisms

Ring morphisms preserve units and inverses. *Iso*morphisms preserve integral domains and fields, but *epi*morphisms may not: for example, if the integral domain $\mathbb{Z}$ is mapped to $\mathbb{Z}_m$ then this homomorphic image is not an integral domain if m is composite.

6.5 Field of quotients

Given an integral domain D it is useful to be able to construct a field F that includes D, i.e. so that every element of F except 0 is invertible. If D is finite then it is already a field, so assume D is infinite (and not a field). The construction generalizes the construction of $\mathbb{Q}$ from $\mathbb{R}$.

Theorem 30 (Field of quotients of an integral domain) *Let D be an integral domain, and let $D^* = D - \{0\}$. Define a relation $\sim$ on $D \times D^*$ by*

$$(a, b) \sim (c, d) \iff ad = bc.$$

Then $\sim$ is an equivalence relation on $D \times D^*$.

Now let $Q(D) = (D \times D^*)/\sim$ and denote the equivalence class $[(a, b)]$ by the "fraction" or "quotient" a/b. Then

$$a/b = c/d \iff ad = bc.$$

Define $+$ and $\cdot$ on $Q(D)$ by

$$a/b + c/d = (ad + bc)/(bd),$$

$$(a/b) \cdot (c/d) = (ac)/(bd);$$

then $+$ and $\cdot$ are well defined.

$Q(D)$ is a field, in which

$$\begin{array}{rrcl} \textit{zero}: & 0 & = & 0/1, \\ \textit{unity}: & 1 & = & 1/1, \\ \textit{additive inverse}: & -(a/b) & = & (-a)/b, \\ \textit{multiplicative inverse}: & (a/b)^{-1} & = & b/a \ (a \neq 0). \end{array}$$

$\tilde{D} = \{a/1 \mid a \in D\}$ is a subdomain of $Q(D)$ that is isomorphic to D under the map $a \mapsto a/1$. Identifying $a \in D$ with $a/1 \in \tilde{D}$ gives the result that if F is a field that includes D then F includes $Q(D)$. Thus $Q(D)$ is the smallest *field that includes D.*

Example: $\mathbb{Q} = Q(\mathbb{Z})$.

Proposition 31 *Let D, D' be isomorphic integral domains $D \cong D'$, then $Q(D) \cong Q(D')$.*

6.5.1 Rational functions

Just as D has a field of quotients, so does $D[x]$, denoted by

$$D(x) = \{a(x)/b(x) \mid a(x), b(x) \in D[x]; b(x) \neq 0\},$$

and called the *field of rational functions over D*. [Note the important distinction between the notations $D[x]$ – a ring of polynomials – and $D(x)$ – a field of rational functions!]

7 Divisibility in integral domains

As usual, D denotes an integral domain and F denotes a field (a special case of an integral domain).

Definition 20 *The divisibility relation $a \,|\, b$ in D* means *that $b = ac$ for some $c \in D$.*

Examples:

1. In $\mathbb{Z}$: $3 \,|\, 6$, $3 \nmid 7$.

2. In $\mathbb{Z}[x]$: if $a(x) = 2x + 4$, $b(x) = 2x^2 + 3x - 2$, then

$$\frac{b(x)}{a(x)} = \frac{(2x-1)(x+2)}{2(x+2)} = \frac{2x-1}{2}$$

in the quotient field $Q(\mathbb{Z}[x])$ ($\cong \mathbb{Q}(x)$), or $x - \frac{1}{2}$ in $\mathbb{Q}[x]$, $\notin \mathbb{Z}[x]$. Hence $a(x) \mid b(x)$ in $\mathbb{Q}[x]$, but $a(x) \nmid b(x)$ in $\mathbb{Z}[x]$.

3. In any field: $a \mid b$ provided $a \neq 0$, because $b = a(a^{-1}b)$, and hence divisibility is trivial in a field.

Definition 21 *$a, b \in D^*$ ($= D - \{0\}$) are called* associates*, denoted $a \sim b$ ("a tilde b"), if $a = ub$ where u is a unit in D.*

Proposition 32

$$a \sim b \iff a \mid b \text{ and } b \mid a.$$

Proof

($\Rightarrow$): If $a \sim b$ then $a = ub$, so $b \mid a$. Also $b = u^{-1}a$ since u is a unit, so $a \mid b$.

($\Leftarrow$): If $a \mid b$ and $b \mid a$ then $b = xa$ and $a = yb$ for some $x, y \in D$. Therefore $b = xyb$, so $xy = 1$ by cancellation of $b \neq 0$. Thus y is a unit (and so is x) and hence $a \sim b$. □

A divisor a of b is called *proper* if it is neither a unit nor an associate of b. (This therefore excludes the unity element 1, which is always a unit, and b itself.)

Proposition 33 *If $a \sim b$ and $c \sim d$ then $a \mid c \Rightarrow b \mid d$.*

Thus divisibility is interesting only modulo associates. But this ambiguity can be removed as follows.

Proposition 34 *The relation $\sim$ is an equivalence relation on D^*.*

Thus D^* is partitioned by $\sim$ into equivalence classes of the form

$$[a] = \{ua \mid u \in U(D)\},$$

where $U(D)$ is the group of units of D. A *distinguished associate* of $a \in D^*$ is a distinguished representative of $[a]_\sim$. The following are conventional choices:

1. In $\mathbb{Z}$, $U(\mathbb{Z}) = \{1, -1\}$. Hence $[m] = \{m, -m\}$ and the distinguished associate is chosen to be *positive* (i.e. $|m|$).

2. In $F[x]$, $U(F[x]) = F^*$. Hence $[a(x)] = \{ca(x) \mid c \in F^*\}$ and the distinguished associate is chosen to be monic (by dividing by the leading coefficient).

3. In *any* integral domain D, $[u] = U(D)$ if u is a unit. The distinguished unit is chosen to be the unity element 1.

7.1 Greatest common divisors

Definition 22 *A* greatest common divisor *(gcd) of $a, b \in D$ (a, b not both zero) is an element $g \in D$ such that*

1. *$g \mid a$ and $g \mid b$ (*common*);*

2. *$c \mid a$ and $c \mid a \Rightarrow c \mid g$ (*greatest*).*

Proposition 35 *If g is a gcd of $a, b \in D$ then so is any associate of g. Conversely, if g, h are gcds of $a, b \in D$ then $g \sim h$.*

Hence a gcd is determined only up to associates. A unique gcd can be obtained by choosing a distinguished associate, as above. Thus:

1. in $\mathbb{Z}$ choose $\gcd(a, b) > 0$;

2. in $F[x]$ choose $\gcd(a(x), b(x))$ monic;

3. in any D, if a gcd of a and b is a unit then take $\gcd(a, b) = 1$: a and b are then called *relatively prime*.

7.2 Primes and factorization

Definition 23 *A nonunit $p \in D^*$ is called* prime *if $p = ab\,(a, b \in D^*) \Rightarrow$ either a or b is a unit.*

Thus a prime has no proper divisors. A nonunit that is not prime is called *composite*. Thus a composite element $c \in D^*$ has a factorization of the form $c = ab\,(a, b \in D^*)$ where neither a nor b is a unit.

Proposition 36 *If p is prime then so is any associate of p.*

Thus only distinguished primes need be considered, e.g. positive primes in $\mathbb{Z}$.

Polynomial domains depend subtly on their underlying number (coefficient) domains. If $a(x)$ is prime (resp. composite) in $D[x]$ then $a(x)$ is called *irreducible* (resp. *reducible*) *over* D.

In $F[x]$ (F a field) the units are the nonzero constant (degree zero) polynomials in F^*. Hence $a[x] \in F[x]$ is reducible over F if $a(x) = b(x)c(x)$ for polynomials $b(x)c(x) \in F[x]$ having degree ≥ 1. In $D[x]$ (D an integral domain) the latter condition is sufficient but not necessary for reducibility over D, as shown by the following.

Examples:

1. $2x+2$ is *irreducible* over $\mathbb{Q}$ (because 2 is a unit in $\mathbb{Q}[x]$), but reducible over $\mathbb{Z}$ as $2(x+1)$ (because neither 2 nor $x+1$ is a unit in $\mathbb{Z}[x]$).

2. x^2-2 is reducible over $\mathbb{R}$ as $(x+\sqrt{2})(x-\sqrt{2})$, but irreducible over $\mathbb{Q}$ (because $\sqrt{2} \notin \mathbb{Q}$).

3. x^2+1 is reducible over $\mathbb{C}$ as $(x+i)(x-i)$, but irreducible over $\mathbb{R}$ (or $\mathbb{Q}$ or $\mathbb{Z}$).

An application to factorization: Let $a(x) \in F[x]$ be a quadratic or cubic. If $a(x)$ is reducible then $a(x) = b(x)c(x)$ where $b(x), c(x)$ have degree ≥ 1. Hence one of $b(x)$ and $c(x)$ must be linear (in both the quadratic and cubic cases), and so must have a *root* in F. This therefore proves the following.

Proposition 37 *A quadratic or cubic polynomial $a(x) \in F[x]$ is reducible over F if and only if $a(x)$ has a root in F.*

Example: Let $a(x) = x^3 + 2x + 2$. Then:

1. $a(x)$ is reducible over $\mathbb{Z}_5$ since $a(1) = 1+2+2$ ($= 5$ in $\mathbb{Z}$) $= 0$ in $\mathbb{Z}_5$.

2. $a(x)$ is irreducible over $\mathbb{Z}_3$ since $a(0) = 2$, $a(1)$ ($= 5$ in $\mathbb{Z}$) $= 2$ in $\mathbb{Z}_3$, $a(2)$ ($= 8+4+2$ in $\mathbb{Z}$) $= 2+1+2 = 2$ in $\mathbb{Z}_3$, and the only possible roots in $\mathbb{Z}_3$ are 0, 1, 2.

8 Euclidean domains

Polynomials over a field have essentially the same division properties as $\mathbb{Z}$, but polynomials over an integral domain that is not a field do not. The notion of Euclidean domain explains this difference, and generalizes many important properties of $\mathbb{Z}$.

Definition 24 *A* Euclidean domain *is an integral domain D together with a "degree" function (sometimes called an "integral norm") $d : D^* \to \mathbb{N}$ such that*

1. *$d(ab) \geq d(a)$ $(a, b \neq 0)$;*
2. Division property: *for every $a, b \in D$ $(b \neq 0)$ there exist a "quotient" q and "remainder" r in D such that*

$$a = bq + r, \quad d(r) < d(b) \text{ or } r = 0.$$

Examples:

1. $\mathbb{Z}$ with $d(a) = |a|$ is a Euclidean domain, because by the division property of $\mathbb{Z}$

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

and hence also

$$a = b(q + 1) + (r - b), \quad |r - b| < |b| \text{ if } r > 0.$$

Thus in $\mathbb{Z}$ there are generally two possible quotient-remainder pairs, so the axioms of a Euclidean domain do not guarantee a unique division. In $\mathbb{Z}$, uniqueness is obtained by requiring a non-negative remainder. The unique or preferred remainder of a divided by b is denoted $r_b(a)$.

2. $F[x]$ with $d(a) = \deg a(x)$ is a Euclidean domain with a unique quotient and remainder. However, $\mathbb{Z}[x]$ with $d(a) = \deg a(x)$ is *not* a Euclidean domain. As a proof by counterexample, an attempt to divide $3x^5$ by $2x^3$ in $\mathbb{Z}[x]$ gives

$$3x^5 = 2x^3.0 + 3x^5$$

because $2 \nmid 3$ in $\mathbb{Z}$. But $(d(3x^5) = 5) \not< (d(2x^3) = 3)$ so the division axiom cannot be satisfied.

...

The rest of this section is a fairly straightforward (but important) generalization of properties of the integers, and it is instructive to compare the proofs of results here with those given for the integers earlier.

8.1 Greatest common divisors in Euclidean domains

A "gcd domain" is an integral domain in which every pair of elements has a gcd that can be expressed as a linear combination of the pair, as follows.

Theorem 38 (A Euclidean domain is a gcd domain) *Let D be a Euclidean domain, and let $a, b \in D$ (not both zero). Then a and b have a gcd g expressible in the form*

$$g = sa + tb \quad (s, t \in D).$$

The main tool to prove this theorem is the following

Lemma 39 *With $a, b \in D$ (not both zero), define*

$$S(a, b) = \{sa + tb \mid s, t \in D\}.$$

Choose $g \in S(a, b)$ ($g \neq 0$) to have minimum degree: $d(g) \leq d(x)$ for all $x \in S(a, b)$. Then

$$S(a, b) = (g),$$

where $(g) = \{rg \mid r \in D\}$ is the set of all ring multiples of g.

Proof $S(a, b)$ contains nonzero elements because at least one of $a = 1a + 0b$ and $b = 0a + 1b$ in $S(a, b)$ is nonzero by assumption. Therefore it is possible to choose a $g \neq 0$.

$(g) \subseteq S(a, b)$ is obviously true from the definitions of (g) and $S(a, b)$.

To prove $S(a, b) \subseteq (g)$, perform a Euclidean division of $x \in S(a, b)$ by g to obtain

$$x = gq + r \quad \text{where } d(r) < d(g) \text{ or } r = 0.$$

Now show that in fact $r = 0$. Since $x, g \in S(a, b)$, we have

$$x = s_x a + t_x b, \quad g = s_g a + t_g b,$$

and hence

$$r = x - gq = (s_x - s_g q)a + (t_x - t_g q)b$$

is in $S(a, b)$. Hence $r = 0$, otherwise we would have $d(r) < d(g)$ in contradiction to g having minimum degree in $S(a, b)$. Thus $S(a, b) \subseteq (g)$. □

Proof of theorem. With $S(a, b)$ and g as in the lemma, the claim is that g is a gcd of a, b. Since $a, b \in S(a, b) = (g)$, a and b are multiples of g, so g is

a *common* divisor of a and b. Moreover, if $c \mid a$ and $c \mid b$ then $c \mid g = s_g a + t_g b$ trivially. Thus g is a *greatest* common divisor of a and b. □

As discussed above, a *unique* gcd can be obtained by choosing a distinguished associate of an equivalence class of gcds.

8.2 Prime factorization in Euclidean domains

A "unique factorization domain" is an integral domain in which every nonunit can be expressed as a product of a unique set of factors. The following lemmas provide the generalization of properties of the integers that are required for proofs by induction in abstract Euclidean domains. Let D be a Euclidean domain with degree function d.

Lemma 40

1. $d(1) \leq d(a)$ *for any* $a \in D^*$.
2. $d(1) = d(a) \iff a$ *is a unit.*

Proof

1. $d(1) \leq d(1a = a)$ by definition of a Euclidean degree function.
2. ($\Leftarrow$) If a is a unit then $d(a) \leq d(aa^{-1} = 1)$, so $d(a) = d(1)$ by (1) above.

   ($\Rightarrow$) Let $d(a) = d(1)$. Euclidean division of 1 by a gives $1 = aq + r$ where $d(r) < d(a)$ or $r = 0$. But $d(r) < d(1)(= d(a))$ is impossible by (1) above. Hence $r = 0$ and so a is a unit.

□

Lemma 41 *In* D^*, *if* $a = bc$ *where* c *is a nonunit then* $d(b) < d(a)$.

Proof A Euclidean degree function must satisfy $d(b) \leq d(a)$ by definition, so it is required to prove that the inequality must be *strict*. Euclidean division of b by a gives

$$b = aq + r \quad \text{where } d(r) < d(a) \text{ or } r = 0.$$

But if $r = 0$ then $b = aq = bcq$, so $cq = 1$ by cancellation of $b \neq 0$, which contradicts the assumption that c is a nonunit. Hence $r \neq 0$ so that $d(r) < d(a)$, and

$$r = b - aq = b - bcq = b(1 - cq),$$

so that $d(b) \leq d(r)$. Hence $d(b) \leq d(r) < d(a)$. □

Corollary 42 *If b is a proper divisor of a then $d(b) < d(a)$.*

Proof If $b \,|\, a$ and $a \nmid b$ then $a = bc$ where c is a nonunit. □

Theorem 43 (Existence of a factorization into primes) *Any $a \in D^*$ is either a unit or can be expressed as a finite product of primes: $a = p_1 p_2 \cdots p_n$.*

Lemma 44 *Let a and b be relatively prime in a Euclidean domain D. Then*

1. *$a \,|\, bc \;\Rightarrow\; a \,|\, c$;*
2. *$a \,|\, c$ and $b \,|\, c \;\Rightarrow\; ab \,|\, c$.*

The requirement that a and b be relatively prime is necessary: in $\mathbb{Z}$ for example $4 \,|\, (2 \times 6)$ but $4 \nmid 6$, and $6 \,|\, 12$ but $(4 \times 6) \nmid 12$.

Corollary 45 (to 1 above). *If p is a prime in a Euclidean domain D then*

$$p \,|\, ab \;\Rightarrow\; p \,|\, a \text{ or } p \,|\, b,$$

and hence

$$p \,|\, \prod_{i=1}^{n} a_i \;\Rightarrow\; p \,|\, a_i \text{ for some } i,\ 1 \leq i \leq n.$$

Theorem 46 (Uniqueness of prime factorization) *In a Euclidean domain D, every nonunit a can be expressed as a product of primes in essentially one way: if*

$$a = p_1 p_2 \cdots p_s = q_1 q_2 \cdots q_t$$

where the p_i and q_i are primes, then $s = t$ and there exists a reordering of the q_i such that

$$p_1 \sim q_1,\ p_2 \sim q_2,\ \cdots,\ p_s \sim q_s.$$

Both the main theorems can be proved by induction, using the lemmas introduced above.

Corollary 47 *In a Euclidean domain, any nonunit a has a prime decomposition of the form*

$$a = up_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$$

where u is a unit, the p_i are the distinct non-associated prime factors of a, and each exponent $e_i > 0$ is uniquely determined by the $\sim$-equivalence class of p_i.

9 Exercises

The assessed questions in this set of exercises are the last three (which does not necessarily mean that they are the hardest)!

1. Prove Lemma 1.

2. Prove that "equivalence mod m" as defined in §1.2.1 is indeed an equivalence relation.

3. Show the uniqueness of an identity element in any groupoid.

4. In a group G, prove that if $(ab)^n = a^n b^n$ for all $a, b \in G$ and positive integers n then G is abelian (commutative), and conversely.

5. In a *finite* monoid $[M; \cdot, 1]$, prove that $uv = 1 \;\Rightarrow\; vu = 1$. (Thus in a finite monoid either left or right invertibility implies two-sided invertibility.) [*Hint*: Since M is finite, successive powers of u cannot all be distinct.] Give a counterexample for the case of an infinite monoid.

6. (*Idempotents in finite semigroups.*) In a semigroup $[S; \cdot]$ an element s is an *idempotent* if $s^2 = s$. Show that for every element s in a *finite* semigroup some power s^n $(n \in \mathbb{Z}^+)$ is an idempotent.

7. Prove the fact asserted in the example at the end of §4 that "requirements 2 and 3 for a map to be a morphism from a multiplicative to an additive group are consequences of 1 and the definition of a group".

8. Complete the sketch proof given in the notes that

$$S = \{a + b\alpha + c\alpha^2 \mid a, b, c \in \mathbb{Q}\},$$

where $\alpha = \sqrt[3]{2}$, is a subring of $\mathbb{R}$, i.e. perform explicitly the required "direct computations".

9. Complete the proof that the *unital subring* of any ring R is given by

$$[1_R] = \{n \cdot 1_R \mid n \in \mathbb{Z}\},$$

by proving that it satisfies the subring conditions.

10. Prove that the set $U(R)$ of all units of a ring R forms a multiplicative group.

11. (a) Show that every nonzero element of $\mathbb{Z}_m$ is either a unit or a zerodivisor.

(b) What is $U(\mathbb{Z}_m)$?

12. In an integral domain prove that $a^2 = 1 \Rightarrow a = \pm 1$. Does this result hold in a commutative ring with zerodivisors?

13. If a commutative ring R has prime characteristic p, is R necessarily an integral domain? What if R has characteristic zero?

14. Prove Proposition 33 on divisibility by associate elements in integral domains.

15. Prove Proposition 34 that the associate relation is an equivalence relation.

16. Prove Proposition 35 concerning gcds and associates from the definitions of gcd and associate.

17. Let D be a Euclidean domain, p a prime in D. Prove that $\sqrt{p} \notin Q(D)$. Conclude that $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, etc., are *irrational*. [*Hint*: Assume, to the contrary, that $\sqrt{p} = a/b$. Then consider $b^2 p = a^2$ in the light of unique factorization.]

18. (** **Assessed** **)
In $R[x]$, what is $[R \cup \{x^2\}]$? Prove your claim.

19. (** **Assessed** **)
Factor into irreducibles the following polynomials in $\mathbb{Z}_3[x]$:
(a) $x^2 + 1$; (b) $x^3 + x + 1$; (c) $x^3 + 2x + 2$;
(d) $x^4 + x^3 + x^2 + x + 1$; (e) $x^4 + x^3 + x + 1$.

20. (** **Assessed** **)
In a Euclidean domain, prove that $a \sim b \Rightarrow d(a) = d(b)$.