Math 332 Winter 2025: lecture diary

Darij Grinberg

draft, March 14, 2025
(This is **NOT** a text or a set of notes. It is just an archive of what I write on my virtual blackboard in class. See `https://www.cip.ifi.lmu.de/~grinberg/t/25wa/index.html` for the actual notes.)

1. Rings and fields

1.1. Defining rings

1.1.1. The definition

Definition 1.1.1. A **ring** means a set R equipped with

- two binary operations (i.e., maps from $R \times R$ to R) that are called **addition** and **multiplication** and are denoted by $+$ and $\cdot$, and
- two elements of R that are called **zero** and **unity** and are denoted by 0 and 1,

such that the following properties (the **ring axioms**) hold:

1. $(R, +, 0)$ is an abelian group. In other words:
   a) The operation $+$ is associative (i.e., $a + (b + c) = (a + b) + c$ for all a, b, c).
   b) The element 0 is a neutral element for $+$ (that is, $a + 0 = 0 + a = a$ for all a).
   c) Each element $a \in R$ has an inverse for the operation $+$ (i.e., an element $b \in R$ such that $a + b = b + a = 0$).
   d) The operation $+$ is commutative (i.e., $a + b = b + a$ for all a, b).
2. $(R, \cdot, 1)$ is a monoid. In other words:
   a) The operation $\cdot$ is associative (i.e., $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all a, b, c).
   b) The element 1 is a neutral element for $\cdot$ (that is, $a \cdot 1 = 1 \cdot a = a$ for all a).

   We **do not** require commutativity of $\cdot$.
3. The **distributive laws** hold in R: That is, for all $a, b, c \in R$, we have
$$a \cdot (b + c) = a \cdot b + a \cdot c;$$
$$(b + c) \cdot a = b \cdot a + c \cdot a.$$
4. We have $a \cdot 0 = 0 \cdot a = 0$ for all $a \in R$.

The zero of R and the unity of R are called 0 and 1 because they behave like the numbers 0 and 1; but they don't have to be these numbers. In case of ambiguity, we fall back to writing 0_R and 1_R for them.

The unity of R is also known as the **identity** of R or the **one** of R.

The product $a \cdot b$ is often written ab.

The inverse of an element $a \in R$ in the group $(R, +, 0)$ is called the **additive inverse** of a, and is denoted by $-a$.

The sum $a + (-b)$ is abbreviated $a - b$ and called the **difference** of a and b.

Definition 1.1.2. A ring R is said to be **commutative** if its multiplication is commutative (i.e., we have $ab = ba$ for all a, b).

1.1.2. Some examples

- The sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ (endowed with the usual addition, the usual multiplication, the usual 0 and the usual 1) are commutative rings.

  (Note that we never required the existence of multiplicative inverses.)

- The set $\mathbb{N} := \{0, 1, 2, \ldots\}$ of nonnegative integers is not a ring, since it has no additive inverses (except for 0). Nevertheless it almost fits the bill, as it satisfies all the other ring axioms. Such objects are called **semirings**.

- We can define a commutative ring $\mathbb{Z}'$ as follows:

  We define a binary operation $\widetilde{\times}$ on the set $\mathbb{Z}$ by setting

  $$a \mathbin{\widetilde{\times}} b := -ab \qquad \text{for all } (a, b) \in \mathbb{Z} \times \mathbb{Z}.$$

  Now, let $\mathbb{Z}'$ be the set $\mathbb{Z}$, endowed with the usual addition $+$ and the unusual multiplication $\widetilde{\times}$ and the usual $0_{\mathbb{Z}'} = 0$ and the unusual $1_{\mathbb{Z}'} = -1$. Then it is easy to check that this $\mathbb{Z}'$ is again a commutative ring. But in fact, this ring $\mathbb{Z}'$ is just a copy of the original ring $\mathbb{Z}$ of integers, but with every integer k renamed as $-k$. To make this more precise, we need the notion of a **ring isomorphism**, which allows us to say that our ring $\mathbb{Z}'$ is **isomorphic** to $\mathbb{Z}$ via the ring isomorphism

  $$\mathbb{Z} \to \mathbb{Z}', \qquad k \mapsto -k.$$

- The quotient rings $\mathbb{Z}/n$ for $n \in \mathbb{Z}$ are further examples of rings. For instance,

  $$\mathbb{Z}/3 = \left\{\overline{0}, \overline{1}, \overline{2}\right\}, \qquad \text{with } \overline{1} + \overline{1} = \overline{2} \text{ and } \overline{2} + \overline{2} = \overline{4} = \overline{1}.$$

- The polynomial rings

  $$\begin{aligned} \mathbb{Q}[x] &= \{\text{all polynomials in the indeterminate } x \text{ over } \mathbb{Q}\}; \\ \mathbb{Z}[x] &= \{\text{all polynomials in the indeterminate } x \text{ over } \mathbb{Z}\}; \\ \mathbb{R}[x] &= \{\text{all polynomials in the indeterminate } x \text{ over } \mathbb{R}\}; \\ \mathbb{Q}[x, y] &= \{\text{all polynomials in the indeterminates } x, y \text{ over } \mathbb{Q}\}. \end{aligned}$$

  And many more along these lines. We will see them in a later chapter.

- The set of all functions from $\mathbb{Q}$ to $\mathbb{Q}$ is a commutative ring, where addition and multiplication are defined pointwise:

$$\begin{aligned} (f+g)(x) &= f(x)+g(x) \qquad \text{for all } f,g:\mathbb{Q}\to\mathbb{Q} \text{ and } x\in\mathbb{Q}; \\ (f\cdot g)(x) &= f(x)\cdot g(x) \qquad \text{for all } f,g:\mathbb{Q}\to\mathbb{Q} \text{ and } x\in\mathbb{Q}, \end{aligned}$$

  where the zero is the "constant-0" function and where the unity is the "constant-1" function.

  The same construction works for functions from $\mathbb{Q}$ to $\mathbb{R}$, or from $\mathbb{R}$ to $\mathbb{Q}$, or from $\mathbb{N}$ to $\mathbb{Q}$.

  More generally, if R is a ring, and if S is any set, then the set of all functions from S to R is a ring (with $+$, $\cdot$, 0 and 1 defined as above). This new ring is commutative if R is.

When we specify a ring, we don't need to prove its 0 and its 1; we only need to ensure that they exist.

Some more examples of rings:

- The ring $\mathbb{H}$ of quarternions:

$$\mathbb{H} = \{a+bi+cj+dk \mid a,b,c,d\in\mathbb{R}\}$$

  with addition being the boring one:

$$\begin{aligned} &(a+bi+cj+dk)+\left(a'+b'i+c'j+d'k\right) \\ &=\left(a+a'\right)+\left(b+b'\right)i+\left(c+c'\right)j+\left(d+d'\right)k \end{aligned}$$

  and multiplication being given by distributivity and

$$\begin{aligned} &ij=-ji=k, \qquad jk=-kj=i, \qquad ki=-ik=j, \\ &i^2=j^2=k^2=-1. \end{aligned}$$

  and the requirement that real numbers commute with everything (so $ai=ia$ and so on when $a\in\mathbb{R}$). This is not a commutative ring.

- There are many rings "between" $\mathbb{Q}$ and $\mathbb{R}$. For instance, let

$$\begin{aligned} \mathbb{S} &= \left\{\text{all real numbers of the form } a+b\sqrt{5} \text{ with } a,b\in\mathbb{Q}\right\} \\ &= \left\{3=3+0\sqrt{5},\ 2+5\sqrt{5},\ 7-8\sqrt{5},\ \frac{23}{6}+19\sqrt{5},\ 2\sqrt{5},\ \ldots\right\}. \end{aligned}$$

This is a ring (with the usual addition, multiplication, 0 and 1). To prove this, we note that the ring axioms (except the existence of additive inverses) are satisfied for $\mathbb{S}$ because they are satisfied for $\mathbb{R}$. The existence of additive inverses because

$$-\left(a+b\sqrt{5}\right) = (-a) + (-b)\sqrt{5} \in \mathbb{S}.$$

It remains to prove "closure" – i.e., to prove that the operations $+$ and $\cdot$ on $\mathbb{S}$ are actually maps from $\mathbb{S} \times \mathbb{S}$ to $\mathbb{S}$. In other words, we must prove that every $x, y \in \mathbb{S}$ satisfy $x + y \in \mathbb{S}$ and $xy \in \mathbb{S}$. We can do this by hand:

$$\begin{aligned}
\left(a+b\sqrt{5}\right) + \left(c+d\sqrt{5}\right) &= (a+c) + (b+d)\sqrt{5}; \\
\left(a+b\sqrt{5}\right)\left(c+d\sqrt{5}\right) &= ac + ad\sqrt{5} + bc\sqrt{5} + bd\underbrace{\sqrt{5}\sqrt{5}}_{=5} \\
&= ac + ad\sqrt{5} + bc\sqrt{5} + 5bd \\
&= (ac+5bd) + (ad+bc)\sqrt{5}.
\end{aligned}$$

So $\mathbb{S}$ is a commutative ring.

- We could define a different ring structure on the same set $\mathbb{S}$: specifically, a ring that, as a set, is identical with $\mathbb{S}$, but has a different choice of multiplication and unity. Namely, we define a binary operation $*$ on $\mathbb{S}$ by

$$\left(a+b\sqrt{5}\right) * \left(c+d\sqrt{5}\right) = ac + bd\sqrt{5} \qquad \text{for all } a,b,c,d \in \mathbb{Q}.$$

  To make sure that this is well-defined, we would need to check that each $x \in \mathbb{S}$ can be written as $a + b\sqrt{5}$ for **unique** $a, b \in \mathbb{Q}$, but this is quite easy using the irrationality of $\sqrt{5}$. It is also easy to check that the set $\mathbb{S}$, equipped with the usual addition $+$, the unusual multiplication $*$, the usual zero 0 and the unusual unity $1 + \sqrt{5}$, is a commutative ring. It is not the same ring as $\mathbb{S}$, not even isomorphic to $\mathbb{S}$.

- Let $\mathbb{S}_3$ be the set of all real numbers of the form $a + b\sqrt[3]{5}$ with $a, b \in \mathbb{Q}$. Is this a ring (endowed with the usual addition, the usual multiplication, the usual 0 and the usual 1)?

  No, because multiplication is not a binary operation on $\mathbb{S}_3$:

$$\left(a+b\sqrt[3]{5}\right)\left(c+d\sqrt[3]{5}\right) = ac + ad\sqrt[3]{5} + bc\sqrt[3]{5} + bd\sqrt[3]{25}.$$

  There is no way to rewrite the RHS here in the form $u + v\sqrt[3]{5}$ with $u, v \in \mathbb{Q}$. (Proving this is not that easy, but doable.)

- For any $n \in \mathbb{N}$, the set $\mathbb{R}^{n \times n}$ of all $n \times n$-matrices with real entries (endowed with matrix addition, matrix multiplication, the zero matrix and the identity matrix) is a ring. It is not commutative unless $n \leq 1$, since usually $AB \neq BA$ for matrices.

  More generally: If R is any ring, and if $n \in \mathbb{N}$, then the set $R^{n \times n}$ of all $n \times n$-matrices with entries in R (endowed with matrix addition, matrix multiplication, the zero matrix and the identity matrix) is a ring. This is called the $n \times n$**-matrix ring** over R; it is denoted by $R^{n \times n}$ or $\mathrm{M}_n(R)$. Note that $R^{n \times n}$ is not commutative even for $n = 1$ if R itself is not commutative.

From now on, we will omit the words "endowed with the usual addition, ...": Any ring with a reasonable addition, multiplication etc. is understood to be endowed with these operations unless we declare otherwise.

- The **zero ring** is the ring consisting of a single element 0. This element serves both as zero and as unity. (So $0 = 1$ in this ring.) Both operations $+$ and $\cdot$ are given by $0 + 0 = 0 \cdot 0 = 0$. The zero ring is commutative.

  More generally, a **trivial ring** means a ring with only one element. Any trivial ring is just the zero ring with its element 0 renamed.

- Let n be an integer.

  Consider the relation "congruent modulo n" on the set $\mathbb{Z}$. It is defined by

  $$a \equiv b \bmod n \iff n \mid a - b.$$

  This relation (for fixed n) is an equivalence relation. Its equivalence classes are called the **residue classes of integers modulo** n. Explicitly for each integer a, the residue class that contains a is

  $$\begin{aligned} &\{\text{all integers that are congruent to } a \text{ modulo } n\} \\ &= \{\text{all integers that differ from } a \text{ by a multiple of } n\} \\ &= \{\ldots,\ a-3n,\ a-2n,\ a-n,\ a,\ a+n,\ a+2n,\ a+3n,\ \ldots\}. \end{aligned}$$

  We denote this class by $\overline{a}$. Two integers a and b satisfy $\overline{a} = \overline{b}$ if and only if $a \equiv b \bmod n$. Thus, working with residue classes of integers modulo n is like working with integers but pretending that n is 0. The set of all these residue classes is called $\mathbb{Z}/n$ (or $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}_n$).

  These residue classes can be added and multiplied by the following rules:

  $$\overline{a} + \overline{b} = \overline{a+b};$$
  $$\overline{a} \cdot \overline{b} = \overline{a \cdot b}.$$

  This turns the set $\mathbb{Z}/n$ into a commutative ring with zero $\overline{0}$ and unity $\overline{1}$.

Note that

$$|\mathbb{Z}/n| = \begin{cases} |n|, & \text{if } n \neq 0; \\ \infty, & \text{if } n = 0. \end{cases}$$

Note that the residue classes in $\mathbb{Z}/0$ are all distinct: No two integers are congruent modulo 0 unless they are equal; thus each residue class is just a singleton: $\overline{a} = \{a\}$.

Examples: In $\mathbb{Z}/12$, we have

$$\begin{aligned} \overline{6} \cdot \overline{7} = \overline{6 \cdot 7} = \overline{42} = \overline{6} \qquad & \text{since } 42 \equiv 6 \bmod 12; \\ \overline{6} \cdot \overline{8} = \overline{6 \cdot 8} = \overline{48} = \overline{0} \qquad & \text{since } 48 \equiv 0 \bmod 12. \end{aligned}$$

In $\mathbb{Z}/15$, we have

$$\begin{aligned} \overline{6} \cdot \overline{7} = \overline{6 \cdot 7} = \overline{42} = \overline{12} \qquad & \text{since } 42 \equiv 12 \bmod 15; \\ \overline{6} \cdot \overline{8} = \overline{6 \cdot 8} = \overline{48} = \overline{3} \qquad & \text{since } 48 \equiv 3 \bmod 15. \end{aligned}$$

- Consider a 4-element set with four elements $0, 1, a, b$. We endow this set with two operations $+$ and $\cdot$ defined by the following tables of values:

$x+y$	$y=0$	$y=1$	$y=a$	$y=b$
$x=0$	0	1	a	b
$x=1$	1	0	b	a
$x=a$	a	b	0	1
$x=b$	b	a	1	0

$x \cdot y$	$y=0$	$y=1$	$y=a$	$y=b$
$x=0$	0	0	0	0
$x=1$	0	1	a	b
$x=a$	0	a	b	1
$x=b$	0	b	1	a

.

For example, $aa = b$ and $ab = 1$. It can be checked in finite time (and a lot of patience) that this really satisfies all the ring axioms, and thus makes a ring. It is a commutative ring. We will soon learn a conceptual way to define this ring (and more general rings like this): it is the field of order 4.

- The ring of **dual numbers**, which are pairs (a, b) of real numbers with addition being entrywise:

$$(a, b) + (c, d) = (a + c,\ b + d),$$

and multiplication being defined by

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

The ring axioms are easy to check. What is interesting is the meaning of this ring: Write (a, b) as $a + b\varepsilon$. Then, addition is

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

Multiplication is

$$(a + b\varepsilon)(c + d\varepsilon) = ac + (ad + bc)\varepsilon.$$

This is what you get multiplying out the LHS and throwing away the $bd\varepsilon^2$ term. So in a sense, ε acts like a first-order infinitesimal: $\varepsilon^2 = 0$. Formally, of course, ε is just the pair $(0, 1)$.

For example,

$$(a + b\varepsilon)^n = a^n + na^{n-1}b\varepsilon.$$

More generally, for any polynomial f, we have

$$f(a + b\varepsilon) = f(a) + f'(a)\, b\varepsilon.$$

1.2. Calculating in rings

1.2.1. What works

Intuitively, the elements of a commutative ring are "numbers in a wider sense" – i.e., objects that behave like numbers. So we expect all the standard rules for numbers to apply more generally in any commutative ring. In a noncommutative ring, things are trickier since the rule $ab = ba$ can fail and therefore all the other rules downstream from it can also fail. Let us be more explicit about what rules we expect to hold.

If $a_1, a_2, \ldots, a_n$ are any n elements of a ring, then the sum $a_1 + a_2 + \cdots + a_n$ is well-defined (i.e., its value does not depend on the order and the parenthesization). More generally, any finite sum of the form $\sum_{s \in S} a_s$ (where S is a finite set) is well-defined whenever the addends a_s belong to a ring. This fact is known as **generalized commutativity**. For rigorous proofs, see some references in the notes.

If our ring is commutative, then the same holds for finite products of the form $\prod_{s \in S} a_s$. If the ring is noncommutative, then $\prod_{s \in S} a_s$ usually does not make sense unless the factors a_s just happen to commute. Nevertheless, a product with a well-specified order, such as $a_1 a_2 \cdots a_n$, makes sense even in a noncommutative ring. This fact is known as **generalized associativity**. Again, see references for proofs. An empty product is defined to be the unity of the underlying ring, whereas an empty sum is defined to be the zero.

There is also a generalized distributivity law saying

$$a(b_1 + b_2 + \cdots + b_n) = ab_1 + ab_2 + \cdots + ab_n$$

and so on.

In any ring, subtraction satisfies the rules you would expect: For any $a, b, c \in R$, we have

$$\begin{aligned}
(-a)\, b &= a\,(-b) = -(ab)\,; \\
(-a)\,(-b) &= ab; \\
(-1)\, a &= -a; \\
a\,(b-c) &= ab - ac; \\
(a-b)\, c &= ac - bc.
\end{aligned}$$

Next, we define some more definitions.

If n is an integer, and a is an element of a ring R, then we define an element na of R by

$$na := \begin{cases} \underbrace{a + a + \cdots + a}_{n \text{ times}}, & \text{if } n \geq 0; \\ -\left(\underbrace{a + a + \cdots + a}_{-n \text{ times}} \right), & \text{if } n < 0. \end{cases}$$

Note that this defines multiplying (aka **scaling**) an element of R by an integer. This is not the same as multiplying two elements of R with each other. (However, if R does contain $\mathbb{Z}$ as a subset, then usually the two operations agree, unless the multiplication on R has been rigged to differ from multiplication of numbers like in our $\mathbb{Z}'$ example above.)

If n is a nonnegative integer, and a is an element of a ring R, then we define an element a^n of R by

$$a^n := \underbrace{aa \cdots a}_{n \text{ times}}.$$

In particular, $a^0 = (\text{empty product}) = 1_R$ by definition.

These scaling and exponentiation operations behave like you would expect, with a couple caveats. For scaling, there are no caveats: We always have

$$\begin{aligned}
(n+m)\, a &= na + ma; \\
n\,(a+b) &= na + nb; \\
(nm)\, a &= n\,(ma)\,; \\
(-1)\, a &= -a
\end{aligned}$$

for $a, b \in R$ and $n, m \in \mathbb{Z}$. For exponentiation, we always have

$$\begin{aligned}
a^{n+m} &= a^n \cdot a^m; \\
a^{nm} &= (a^n)^m
\end{aligned}$$

for $a \in R$ and $n, m \in \mathbb{N}$. Moreover, if $a, b \in R$ commute (i.e., satisfy $ab = ba$), then

$$\begin{aligned} (ab)^n &= a^n b^n; \\ a^i b^j &= b^j a^i; \\ (a+b)^n &= \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \qquad \text{(the binomial theorem)} \end{aligned}$$

for any $n, i, j \in \mathbb{N}$. These are not generally true if a, b do not commute (pitfall!). Also,

$$\begin{aligned} 1_R^n &= 1_R \qquad \text{for any } n \in \mathbb{N}; \\ 0_R^n &= 0_R \qquad \text{for any } n > 0; \\ 0_R^0 &= 1_R. \end{aligned}$$

All of this is proved just like for numbers, except that sometimes commutativity needs to be used explicitly rather than tacitly.

1.2.2. What doesn't work

But rings can be weird:

- It is not always true that $a \neq 0$ and $b \neq 0$ imply $ab \neq 0$. We have seen counterexamples in $\mathbb{Z}/12$. There are also counterexamples in $\mathbb{R}^{2\times 2}$.
- It is not always true that $ab = 1$ implies $ba = 1$. Counterexamples to this are hard to find (for example, $ab = 1 \implies ba = 1$ holds for any finite ring, any field, any integral domain, any matrix ring over a field, ...), but they exist.

1.3. Subrings

1.3.1. Definition

Groups have subgroups; vector spaces have subspaces. For rings, expect the same:

Definition 1.3.1. Let R be a ring. A **subring** of R means a subset S of R such that

- we have $a + b \in S$ for all $a, b \in S$ (that is, S is closed under addition);
- we have $ab \in S$ for all $a, b \in S$ (that is, S is closed under multiplication);

- we have $-a \in S$ for all $a \in S$ (that is, S is closed under negation);
- we have $0 \in S$ (where the 0 means the zero of R);
- we have $1 \in S$ (where the 1 means the unity of R).

These five conditions are called the **subring axioms**.

Proposition 1.3.2. Let S be a subring of a ring R. Then, S itself automatically is a ring (with its operations $+$ and $\cdot$ obtained by restricting the corresponding operations of R, and with the 0 and 1 inherited from R).

1.3.2. Examples

- From the classical construction of the number systems,

$$\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}.$$

  Each of these inclusions is a "subring": that is, $\mathbb{Z}$ is a subring of $\mathbb{Q}$, which in turn is a subring of $\mathbb{R}$, and so on.

- You can extend this chain further to the right: $\mathbb{C}$ is a subring of $\mathbb{H}$ (the Hamilton quaternions).

- Can you extend this chain to the left? Does $\mathbb{Z}$ have any subrings besides itself?

  No, because if S is a subring of $\mathbb{Z}$, then $1 \in S$ (by one of the subring axioms), hence $n \in S$ for each positive integer n (since $n = 1+1+\cdots+1$), thus $-n \in S$ for each positive integer n (since S is closed under negation), and so all integers belong to S.

- Are there rings between $\mathbb{Z}$ and $\mathbb{Q}$? What about

$$\frac{1}{2}\mathbb{Z} = \left\{ \frac{n}{2} \mid n \in \mathbb{Z} \right\} ?$$

  This is not a subring, since it is not closed under multiplication ($\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$). We can, however, fix this by extending it: Consider instead

$$\mathbb{B}_2 = \left\{ \frac{a}{2^k} \mid a \in \mathbb{Z} \text{ and } k \in \mathbb{N} \right\}.$$

  This is a subring of $\mathbb{Q}$ that contains $\mathbb{Z}$ as a subring, so we have

$$\mathbb{Z} \subseteq \mathbb{B}_2 \subseteq \mathbb{Q}.$$

  Similarly you can define $\mathbb{B}_3$ and $\mathbb{B}_4 (= \mathbb{B}_2)$ and $\mathbb{B}_5$ and so on.

- Are there rings between $\mathbb{Q}$ and $\mathbb{R}$? A lot, such as

$$S = \left\{ a + b\sqrt{5} \mid a, b \in \mathbb{Q} \right\}$$

(called $\mathbb{Q}\left[\sqrt{5}\right]$). Another example is the ring

$$\mathbb{Q}\left[\sqrt{2}\right] = \left\{ a + b\sqrt{2} \mid a, b \in \mathbb{Q} \right\}.$$

Another example is the ring

$$\mathbb{Q}\left[\sqrt[3]{2}\right] = \left\{ a + b\sqrt[3]{2} + c\sqrt[3]{4} \mid a, b, c \in \mathbb{Q} \right\}$$

(exercise: check that this really is a subring of $\mathbb{R}$!). Another is

$$\begin{aligned} \mathbb{Q}[\pi] &= \Big\{ a + b\pi + c\pi^2 + d\pi^3 + \cdots \mid a, b, c, d, \ldots \in \mathbb{Q} \\ &\qquad \text{and only finitely many of } a, b, c, d, \ldots \text{ are } \neq 0 \} \\ &= \{ f(\pi) \mid f \text{ is a polynomial with rational coefficients} \}. \end{aligned}$$

(You need all polynomials here – there is no a-priori bound on the degree after which no new values will appear, since π is transcendental.)

- What about rings between $\mathbb{R}$ and $\mathbb{C}$? There are none. The only subrings of $\mathbb{C}$ that contain $\mathbb{R}$ are $\mathbb{R}$ and $\mathbb{C}$ themselves. The easiest way to see this is by realizing that any such subring would be an $\mathbb{R}$-vector subspace of $\mathbb{C}$ that contains $\mathbb{R}$; but $\dim \mathbb{C} = 2$, so the only such subspaces are $\mathbb{R}$ and $\mathbb{C}$.

- There are rings between $\mathbb{Z}$ and $\mathbb{C}$ that are neither sub- nor superrings of $\mathbb{Q}$ and $\mathbb{R}$.

  A particularly important one is $\mathbb{Z}[i]$, the ring of **Gaussian integers**.

  A **Gaussian integer** is a complex number of the form $a + bi$, where $a, b \in \mathbb{Z}$ (and $i = \sqrt{-1}$). For instance, $3 + 5i$ or $7 - 9i$ but not $\frac{2}{3} + 5i$.

  It is easy to see that $\mathbb{Z}[i]$ is a subring of $\mathbb{C}$ and contains $\mathbb{Z}$ as a subring. But it neither contains nor is contained in any of $\mathbb{Q}$ and $\mathbb{R}$.

  Visually, Gaussian integers are the lattice points of a square lattice (i.e., the points with both coordinates integers) in the plane.

  There are also **Gaussian rationals**, called $\mathbb{Q}[i]$, and defined as $a + bi$ where $a, b \in \mathbb{Q}$.

- Recall the ring of functions from $\mathbb{Q}$ to $\mathbb{Q}$. Similarly, there is a ring of functions from $\mathbb{R}$ to $\mathbb{R}$. The latter ring has a subring that consists of all **continuous** functions from $\mathbb{R}$ to $\mathbb{R}$. Another subring consists of all **smooth** functions from $\mathbb{R}$ to $\mathbb{R}$ (= infinitely often differentiable).

- Let $n \in \mathbb{N}$, and let R be any ring. Recall the matrix ring

$$R^{n\times n} = \left\{ \text{all } n\times n\text{-matrices } \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \text{ with } a_{i,j} \in R \right\}.$$

  Some of its subrings are:

  - the subring

$$\begin{aligned} R^{n\leq n} &= \left\{ \text{all } n\times n\text{-matrices } \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ 0 & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{pmatrix} \text{ with } a_{i,j} \in R \right\} \\ &= \left\{ \text{all upper-triangular matrices in } R^{n\times n} \right\}; \end{aligned}$$

  - the subring

$$\begin{aligned} R^{n\geq n} &= \left\{ \text{all } n\times n\text{-matrices } \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \text{ with } a_{i,j} \in R \right\} \\ &= \left\{ \text{all lower-triangular matrices in } R^{n\times n} \right\}; \end{aligned}$$

  - the subring

$$\begin{aligned} R^{n=n} &= \left\{ \text{all } n\times n\text{-matrices } \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{pmatrix} \text{ with } a_{i,j} \in R \right\} \\ &= \left\{ \text{all diagonal matrices in } R^{n\times n} \right\}; \end{aligned}$$

  - the subring

$$\begin{aligned} RI_n &= \{aI_n \mid a \in R\} \\ &= \left\{ \text{all } n\times n\text{-matrices } \begin{pmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a \end{pmatrix} \text{ with } a \in R \right\} \\ &= \{\text{all scalar multiples of the identity matrix}\}. \end{aligned}$$

– many, many more.

On the other hand,

$$R^{n\times n}_{\text{symm}} = \left\{ \text{all symmetric } n\times n\text{-matrices } \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \text{ with } a_{i,j} \in R \right\}$$

is not a subring of $R^{n\times n}$ (unless $n \leq 1$ or R is trivial), since the product of two symmetric matrix is generally not symmetric: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$.

1.4. Zero divisors and integral domains

Definition 1.4.1. An element of a ring R is said to be **nonzero** if it is $\neq 0_R$.

Definition 1.4.2. Let R be a commutative ring. A nonzero element $a \in R$ is said to be a **zero divisor** if there exists a nonzero $b \in R$ such that $ab = 0$.

For example, in the ring $\mathbb{Z}/6$, both elements $\overline{2}$ and $\overline{3}$ are zero divisors, since $\overline{2}\cdot\overline{3} = \overline{6} = \overline{0}$.

Definition 1.4.3. Let R be a commutative ring. Assume that $0 \neq 1$ in R (that is, $0_R \neq 1_R$). We say that R is an **integral domain** if all nonzero $a, b \in R$ satisfy $ab \neq 0$.

In other words, a commutative ring with $0 \neq 1$ is an integral domain if and only if it has no zero divisors.

Examples:

- The rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are integral domains. So would be the ring $\mathbb{H}$ if it was commutative.
- The ring $\mathbb{Z}/n$ is an integral domain if and only if n is 0 or a prime or minus a prime. We will prove this later.
- The ring $\mathbb{S} = \left\{ a + b\sqrt{5} \mid a, b \in \mathbb{Q} \right\}$ is an integral domain (being a subring of the integral domain $\mathbb{R}$), but the ring $\mathbb{S}' = \left\{ a + b\sqrt{5} \mid a, b \in \mathbb{Q} \right\}$ with the multiplication $*$ is not.

- The ring of all functions from $\mathbb{Q}$ to $\mathbb{Q}$ is not an integral domain, since we can find two functions that are not identically 0 but whose product is identically 0.

  Most rings of functions have this behavior! Even a "nice" function ring such as the ring of smooth functions from $\mathbb{R}$ to $\mathbb{R}$. Only once you get to holomorphic or analytic functions do you get an integral domain.

1.5. Units and fields

1.5.1. Units and inverses

By definition, any ring R has an addition, a subtraction and a multiplication. But a division may or may not be possible, depending on what you want to divide by. Even in $\mathbb{Q}$, you cannot divide by 0. In the ring $\mathbb{Z}$, you can divide every integer by 1 and -1, but other divisions may or may not work. In fact, in any ring R, you can divide any element by 1 and by -1. Let us give such elements (which you can divide every element by) a name:

Definition 1.5.1. Let R be a ring.

(a) An element $a \in R$ is said to be a **unit** of R (or **invertible** in R) if there exists a $b \in R$ such that $ab = ba = 1$. In this case, b is unique and is known as the **inverse** (or **reciprocal**, or **multiplicative inverse**) of a, and is denoted by a^{-1}.

(b) We let $R^\times$ denote the set of all units of R.

Some comments:

- Of course, 1 here means 1_R.
- "Unity" and "unit" are not the same: A ring usually has many units, but only one unity. (Of course, the unity is a unit.)
- We required $ab = ba = 1$ rather than $ab = 1$ only. When R is commutative, $ab = 1$ suffices.
- Why is b unique? If there were two such b's, say b_1 and b_2, then we would have $b_1 \underbrace{ab_2}_{=1} = b_1$, so that $b_1 = \underbrace{b_1 a}_{=1} b_2 = b_2$.

Some examples of units:

- The units of the ring $\mathbb{Q}$ are the nonzero elements of $\mathbb{Q}$. This is because each nonzero $r \in \mathbb{Q}$ has a reciprocal $\dfrac{1}{r} \in \mathbb{Q}$.

  Similarly, the same holds for $\mathbb{R}$ and for $\mathbb{C}$ and for $\mathbb{H}$.

- The units of the ring $\mathbb{Z}$ are 1 and -1. Not 2, because $\frac{1}{2} \notin \mathbb{Z}$.
- The units of the ring $\mathbb{R}^{n \times n}$ are the invertible $n \times n$-matrices.
- In the ring of all functions from $\mathbb{Q}$ to $\mathbb{Q}$ (with pointwise $+$ and $\cdot$), the units are those functions that never take the value 0.

Now what about the units of $\mathbb{Z}/n$?

Proposition 1.5.2. Let $n \in \mathbb{Z}$. Then:
(a) The units of the ring $\mathbb{Z}/n$ are precisely the residue classes $\overline{a}$ where $a \in \mathbb{Z}$ is coprime to n.
(b) Let $a \in \mathbb{Z}$. Then, $\overline{a}$ is a unit of $\mathbb{Z}/n$ if and only if a is coprime to n.

Proof. It suffices to show part **(b)**.
(b) We prove the "if" ($\Longleftarrow$) and "only if" ($\Longrightarrow$) directions separately:
$\Longleftarrow$: Assume that $a \in \mathbb{Z}$ is coprime to n. We must prove that $\overline{a}$ is a unit of $\mathbb{Z}/n$.
Since a is coprime to n, we have $\gcd(a,n) = 1$. But Bezout's theorem yields that there exist $x, y \in \mathbb{Z}$ such that $xa + yn = \gcd(a,n)$. Consider these x, y.
Thus, $xa + yn = \gcd(a,n) = 1$. In other words, $xa - 1 = -yn \equiv 0 \bmod n$. Thus $xa \equiv 1 \bmod n$. Therefore, $\overline{x} \cdot \overline{a} = \overline{xa} = \overline{1}$ in $\mathbb{Z}/n$. Since $\mathbb{Z}/n$ is commutative, this entails that $\overline{x}$ is an inverse of $\overline{a}$. So $\overline{a}$ has an inverse, i.e., is a unit.
$\Longrightarrow$: Assume that $\overline{a}$ is a unit of $\mathbb{Z}/n$. We must prove that a is coprime to n.
Since $\overline{a}$ is a unit, it has an inverse $\overline{x}$. Thus, $x \in \mathbb{Z}$ and $xa \equiv 1 \bmod n$. Hence, $\gcd(xa,n) = \gcd(1,n)$ (by the property of gcds saying that $\gcd(\alpha,\beta) = \gcd(\gamma,\beta)$ whenever $\alpha \equiv \gamma \bmod \beta$). Of course, $\gcd(1,n) = 1$. So $\gcd(xa,n) = \gcd(1,n) = 1$. Therefore, $\gcd(a,n) = 1$ as well (since $\gcd(a,n) \mid a \mid xa$ and $\gcd(a,n) \mid n$ and thus $\gcd(a,n) \mid \gcd(xa,n) = 1$). In other words, a is coprime to n. Thus the proof is done. $\square$

Examples:

- The units of the ring $\mathbb{Z}/12$ are $\overline{1}$, $\overline{5}$, $\overline{7}$, $\overline{11}$, since the numbers $a \in \{0, 1, \ldots, 11\}$ that are coprime to 12 are $1, 5, 7, 11$.
- The units of the ring $\mathbb{Z}/5$ are $\overline{1}$, $\overline{2}$, $\overline{3}$, $\overline{4}$.
- The only unit of the ring $\mathbb{Z}/2$ is $\overline{1}$.
- A trivial ring has only one unit, namely its unity (which is also its zero). This is the only case when 0_R is a unit.

Some general facts about units follow:

Theorem 1.5.3. Let R be a ring. Then, the set $R^\times = \{\text{all units of } R\}$ is a multiplicative group. That is: $(R^\times, \cdot, 1)$ is a group.

Proof. We must show the following facts:

1. The unity 1 of R belongs to $R^\times$.
2. If $a, b \in R^\times$, then $ab \in R^\times$.
3. If $a \in R^\times$, then a has an inverse of $R^\times$.

All the group axioms are then inherited from $R^\times$. So let us prove these three facts:

Proof of Fact 1: Well, 1 is its own inverse: $1 \cdot 1 = 1$.

Proof of Fact 2: Let $a, b \in R^\times$. Why is $ab \in R^\times$? Because we can explicitly construct an inverse for it: namely, we claim that $b^{-1}a^{-1}$ is an inverse of ab. To wit,

$$b^{-1}\underbrace{a^{-1} \cdot a}_{=1} b = b^{-1}b = 1 \qquad \text{and}$$
$$a\underbrace{b \cdot b^{-1}}_{=1}a^{-1} = aa^{-1} = 1.$$

Proof of Fact 3: Let $a \in R^\times$. Then, a has an inverse $a^{-1} \in R$. We must show that $a^{-1} \in R^\times$. But this is clear, since a itself is an inverse of a^{-1} (this follows from the same equalities $aa^{-1} = a^{-1}a = 1$ that say that a^{-1} is an inverse of a).

So the proof is complete. □

As consequences of the above proof, we obtain the following facts:

Theorem 1.5.4 (Shoe-sock theorem)**.** Let R be a ring. Let a, b be two units of R. Then, ab is a unit of R, and its inverse is

$$(ab)^{-1} = b^{-1}a^{-1}.$$

Theorem 1.5.5. Let R be a ring. Let a be a unit of R. Then, a^{-1} is a unit of R, and its inverse is $\left(a^{-1}\right)^{-1} = a$.

Of course, we will use these without mentioning.

1.5.2. Fields

As we have seen, many rings (such as $\mathbb{Z}$) have few units, but many other rings (such as $\mathbb{Q}$ or $\mathbb{R}$) have many. The latter kind of ring has a name:

Definition 1.5.6. Let R be a commutative ring. Assume that $0 \neq 1$ in R. Then, R is said to be a **field** if every nonzero element of R is a unit.

Examples:

- The rings $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields. The ring $\mathbb{Z}$ is not (since 2 is not a unit, for example).
- The ring $\mathbb{S} = \mathbb{Q}\left[\sqrt{5}\right] = \left\{a + b\sqrt{5} \mid a, b \in \mathbb{Q}\right\}$ is a field. Indeed, if $a + b\sqrt{5}$ is a nonzero element of $\mathbb{S}$, then $a + b\sqrt{5}$ is a unit, since its inverse is

$$\begin{aligned}\left(a + b\sqrt{5}\right)^{-1} &= \frac{1}{a + b\sqrt{5}} = \frac{a - b\sqrt{5}}{\left(a - b\sqrt{5}\right)\left(a + b\sqrt{5}\right)} \\ &= \frac{a - b\sqrt{5}}{a^2 - 5b^2} = \frac{a}{a^2 - 5b^2} + \frac{-b}{a^2 - 5b^2}\sqrt{5} \in \mathbb{S}.\end{aligned}$$

  Strictly speaking, this relies on the fact that $a^2 - 5b^2 \neq 0$, which is true because otherwise 5 would be a square of a rational number (which it is not: $\sqrt{5}$ is irrational).
- The Hamilton quaternions $\mathbb{H}$ would be a field if they were commutative. A noncommutative ring R with $0 \neq 1$ whose all nonzero elements are units is called a **division ring** or a **skew-field**. So $\mathbb{H}$ is a skew-field.
- Let n be a positive integer. Then, $\mathbb{Z}/n$ is a field if and only if n is prime. (See below for a proof.)

1.6. Fields and integral domains: some connections

Proposition 1.6.1. **(a)** Every field is an integral domain.
(b) Every **finite** integral domain is a field.

Proof. **(a)** Easy: If $ab = 0$ and $a, b \neq 0$, then you can multiply by $a^{-1}b^{-1}$ to obtain $1 = 0$, which is absurd.

(b) Pigeonhole principle. Jurij argued that each nonzero $a \in R$ must have an inverse, because it has two equal powers $a^i = a^j$ (with $i < j$), which then entails $a^i\left(1 - a^{j-i}\right) = 0$, and because R is an integral domain, you can cancel a^i to obtain $1 - a^{j-i} = 0$, which entails that a^{j-i-1} is an inverse of a.

I have a different argument in the lecture notes, but it also uses the pigeonhole principle. □

Of course, part **(b)** does not hold without the word "finite", since $\mathbb{Z}$ is an integral domain but not a field. Other examples of this nature are polynomial rings (see later).

Corollary 1.6.2. Let n be a positive integer. Then,

$$(\mathbb{Z}/n \text{ is an integral domain}) \iff (\mathbb{Z}/n \text{ is a field}) \iff (n \text{ is prime}).$$

Proof. Since $\mathbb{Z}/n$ is finite, the first $\iff$ sign follows immediately from the previous proposition. As for the second $\iff$ sign, it is not much harder:

$\Longleftarrow$: If n is prime, then all the numbers $1, 2, \ldots, n-1$ are coprime to n, so that their residue classes $\overline{1}, \overline{2}, \ldots, \overline{n-1}$ are units of $\mathbb{Z}/n$, and this means that $\mathbb{Z}/n$ is a field.

$\Longrightarrow$: Read this argument backwards. □

The group of units $(\mathbb{Z}/n)^\times$ of the ring $\mathbb{Z}/n$ is a rather interesting object. It is an abelian group, but is it cyclic? Not always. For instance, $(\mathbb{Z}/12)^\times = \{\overline{1}, \overline{5}, \overline{7}, \overline{11}\}$ is isomorphic to $Z_2 \times Z_2$ (the Klein four-group), which is not cyclic. When is it cyclic?

1.6.1. Division

As we know, rings have addition, subtraction and multiplication, but not always division. Nevertheless, when b is a unit of a ring R, and a is any element of R, it makes sense to define $\dfrac{a}{b}$ to be ab^{-1}.

When R is noncommutative, this is rather misleading: firstly because $b^{-1}a$ has an equally claim at being $\dfrac{a}{b}$; secondly because familiar rules like $\dfrac{a}{b} \cdot \dfrac{c}{d} = \dfrac{ac}{bd}$ no longer hold.

When R is commutative, however, all is fine, so we do make this definition:

Definition 1.6.3. Let R be a commutative ring. Let $a \in R$ and $b \in R^\times = \{\text{units of } R\}$. Then, $\dfrac{a}{b}$ is defined to be the element $ab^{-1} = b^{-1}a \in R$. This is also written as a/b, and is called the **quotient** of a by b. The operation $(a, b) \mapsto \dfrac{a}{b}$ is called **division**.

In particular, when R is a field, we can divide by any nonzero element.

Proposition 1.6.4. Division satisfies the rules you would expect: If R is a commutative ring, and if $a, c \in R$ and $b, d \in R^\times$, then

$$\begin{aligned} \frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd}; \\ \frac{a}{b} \cdot \frac{c}{d} &= \frac{ac}{bd}; \\ \frac{a}{b} \Big/ \frac{c}{d} &= \frac{ad}{bc} \qquad \left(\text{if } c \in R^\times\right). \end{aligned}$$

And, of course, division undoes multiplication: i.e., we have the equivalence $\left(\frac{a}{b} = c\right) \iff (a = bc)$ whenever b is a unit.

Proof. Easy consequences of associativity and distributivity and commutativity(!). □

1.7. Ring morphisms

1.7.1. Definition and examples

In modern mathematics, whenever you define some type of objects, it isn't long until you also define a notion of morphisms between these objects:

- Between vector spaces, you have linear maps.
- Between topological spaces, you have continuous maps.
- Between groups, you have group morphisms (= homomorphisms).

All these notions have a commonality: They are a type of maps that respect/preserve certain structures. So let us define a similar concept for rings:

Definition 1.7.1. Let R and S be two rings.

(a) A **ring homomorphism** (or, for short: a **ring morphism**) from R to S means a map $f : R \to S$ that

- **respects addition:** that is, $f(a+b) = f(a) + f(b)$ for all $a, b \in R$;
- **respects multiplication:** that is, $f(ab) = f(a) \cdot f(b)$ for all $a, b \in R$;
- **respects the zero**: that is, $f(0_R) = 0_S$;
- **respects the unity:** that is, $f(1_R) = 1_S$ (contra Volcic, also contra Dummit/Foote).

(b) A **ring isomorphism** from R to S means an invertible ring morphism $f : R \to S$ whose inverse $f^{-1} : S \to R$ is also a ring morphism.

(c) The rings R and S are said to be **isomorphic** (this is written $R \cong S$, or sometimes $R \approx S$) if there exists a ring isomorphism $f : R \to S$.

Examples:

- Let $n \in \mathbb{Z}$. The map
$$\pi : \mathbb{Z} \to \mathbb{Z}/n,$$
$$a \mapsto \overline{a}$$
(that sends each integer a to its residue class $\overline{a} = a + n\mathbb{Z}$ modulo n) is a ring morphism, because any $a, b \in \mathbb{Z}$ satisfy
$$\overline{a+b} = \overline{a} + \overline{b}, \qquad \overline{ab} = \overline{a} \cdot \overline{b}, \qquad \overline{0} = 0_{\mathbb{Z}/n}, \qquad \overline{1} = 1_{\mathbb{Z}/n}.$$

- The map
$$\mathbb{Z} \to \mathbb{Z},$$
$$a \mapsto 2a$$
is not a ring morphism. It respects addition and zero. It does not respect unity or multiplication (e.g., it sends 2 to 4 but $2 \cdot 2$ to 8 rather than $4 \cdot 4$).

- The map
$$\mathbb{Z} \to \mathbb{Z},$$
$$a \mapsto 0$$
is not a ring morphism. It does not respect unity, although it respects everything else.

- The map
$$\mathbb{Z} \to \mathbb{Z},$$
$$a \mapsto a^2$$
is not a ring morphism. It does not respect addition, although it respects everything else.

- The identity map $\mathrm{id} : \mathbb{Z} \to \mathbb{Z}$ is a ring morphism. It is the only ring morphism from $\mathbb{Z}$ to $\mathbb{Z}$. Indeed, if f is any ring morphism from $\mathbb{Z}$ to $\mathbb{Z}$, then $f(1) = 1$ (since f respects the unity), thus $f(n) = n$ for each $n > 0$ (because $f(n) = f(1 + 1 + \cdots + 1) = f(1) + f(1) + \cdots + f(1) = 1 + 1 + \cdots + 1 = n$) but also $f(0) = 0$ (since f respects the zero) and thus $f(-1) = -1$ (since $f(-1) + f(1) = f((-1) + 1) = f(0) = 0$) and thus $f(n) = n$ also for negative n (exercise), so that $f(n) = n$ for all integers n; but this means $f = \mathrm{id}$.

- Consider the map

$$\begin{aligned} f : \mathbb{C} &\to \mathbb{R}^{2\times 2}, \\ a+bi &\mapsto \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \qquad \text{for all } a,b \in \mathbb{R}. \end{aligned}$$

This map f is a ring morphism. Indeed, it clearly respects addition and zero and unity (since $f(1+0i) = I_2$). To show that it respects multiplication. So let $z, w \in \mathbb{C}$. We want to show that $f(zw) = f(z) \cdot f(w)$.

Write z and w as $z = a+bi$ and $w = c+di$. Then,

$$zw = (a+bi)(c+di) = (ac-bd) + (ad+bc)\,i.$$

So

$$f(zw) = \begin{pmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{pmatrix}.$$

In contrast,

$$f(z) \cdot f(w) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \cdot \begin{pmatrix} c & d \\ -d & c \end{pmatrix} = \begin{pmatrix} ac-bd & ad+bc \\ -ad-bc & ac-bd \end{pmatrix}.$$

The RHSs of these equalities are clearly equal; hence, so are the LHSs. Thus, $f(zw) = f(z) \cdot f(w)$, qed.

Since f is injective, we can use the image $f(z)$ of a complex number z as a "stand-in" for z.

This is not an isolated phenomenon. "Likewise", there is an injective ring morphism

$$\begin{aligned} g : \mathbb{H} &\to \mathbb{R}^{4\times 4}, \\ a+bi+cj+dk &\mapsto \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix}. \end{aligned}$$

Several more rings we will study later can be "represented" by matrices, in the sense that we can find injective morphisms from these rings to matrix rings, and thus we can work with matrices instead of with abstract objects.

- The map

$$\begin{aligned} \mathbb{R}^{2\times 2} &\to \mathbb{R}, \\ A &\mapsto \det A \end{aligned}$$

is not a ring morphism, since it fails to respect addition ($\det(A+B) \neq \det A + \det B$), even though it respects everything else. Actually, this cannot be helped: There exist no ring morphisms from $\mathbb{R}^{2\times 2}$ to $\mathbb{R}$.

- Let S be a subring of a ring R. Let $i : S \to R$ be the **canonical inclusion**; this is simply the map that sends each element $a \in S$ to itself. Then, i is a ring morphism. Indeed, it respects addition, because $i(a+b) = i(a) + i(b)$ is just saying that $a + b = a + b$. Similarly, it respects all the other features.

- Let R be a ring. Let S be any set. Recall that the maps from S to R form a ring (with pointwise $+$ and $\cdot$). We call this ring R^S. Fix any $s \in S$. Then, the map

$$\begin{aligned} R^S &\to R, \\ g &\mapsto g(s) \end{aligned}$$

  is a ring morphism. Indeed, it respects addition because $(f+g)(s) = f(s) + g(s)$ for any $f, g \in R^S$ (and this is because we defined addition in R^S this way!). Similarly for all the other requirements.

1.7.2. Basic properties of ring morphisms

The composition of two ring morphisms is again a ring morphism. In other words:

Proposition 1.7.2. Let R, S, T be three rings. Let $f : S \to T$ and $g : R \to S$ be two ring morphisms. Then, $f \circ g : R \to T$ is a ring morphism.

Proof. Same as for groups. □

The following proposition slightly simplifies proving that a map is a ring morphisms:

Proposition 1.7.3. Let R and S be two rings. Let $f : R \to S$ be a map that respects addition. Then, f respects the zero.

Proof. We have $0_R = 0_R + 0_R$, so $f(0_R) = f(0_R + 0_R) = f(0_R) + f(0_R)$. Now subtract $f(0_R)$ and obtain $0_S = f(0_R)$, qed. □

Thus, the "respects the zero" axiom can be removed from the definition of a ring morphism.

By the way, the definition of a ring morphism can be restated as follows: A **ring morphism** is a map $f : R \to S$ between two rings R and S that is a group morphism from $(R, +, 0)$ to $(S, +, 0)$ and a monoid morphism from $(R, \cdot, 1)$ to $(S, \cdot, 1)$.

By definition, ring morphisms preserve the basic structures of a ring ($+$, $\cdot$, 0 and 1). As a consequence, they also preserve all the structures that are derived from these basic structures:

Proposition 1.7.4. Let R and S be two rings. Let $f : R \to S$ be a ring morphism. Then:

(a) The map f respects finite sums: i.e., we have

$$f(a_1 + a_2 + \cdots + a_n) = f(a_1) + f(a_2) + \cdots + f(a_n)$$

for any $a_1, a_2, \ldots, a_n \in R$.

(b) The map f respects finite products: i.e., we have

$$f(a_1 a_2 \cdots a_n) = f(a_1) \cdot f(a_2) \cdot \cdots \cdot f(a_n)$$

for any $a_1, a_2, \ldots, a_n \in R$.

(c) The map f respects differences: i.e., we have

$$f(a - b) = f(a) - f(b) \qquad \text{for all } a, b \in R.$$

(d) The map f respects inverses: i.e., if a is a unit of R, then $f(a)$ is a unit of S, with inverse $(f(a))^{-1} = f\left(a^{-1}\right)$.

(e) The map f respects integer multiples: i.e., if $a \in R$ and $n \in \mathbb{Z}$, then $f(na) = nf(a)$.

(f) The map f respects powers: i.e., if $a \in R$ and $n \in \mathbb{N}$, then $f(a^n) = (f(a))^n$.

Proof. LTTR. □

1.7.3. The image of a ring morphism

Recall that the **image** of a map $f : R \to S$ is defined to be the set

$$\operatorname{Im} f := f(R) = \{f(r) \mid r \in R\}.$$

This makes sense for any map between any sets. For a ring morphism, it is particularly nice, because it is a subring of S. Namely:

Proposition 1.7.5. Let R and S be two rings. Let $f : R \to S$ be a ring morphism. Then, $\operatorname{Im} f = f(R)$ is a subring of S.

Proof. For instance, $\operatorname{Im} f$ is closed under addition, since $f(a) + f(b) = f(a + b)$. The other subring axioms are similar. □

Thus, for example, the set of all 2×2-matrices of the form $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ is a subring of $\mathbb{R}^{2\times 2}$, because it is the image of the above-defined ring morphism

$$\begin{aligned} f : \mathbb{C} &\to \mathbb{R}^{2\times 2}, \\ a + bi &\mapsto \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \qquad \text{for all } a, b \in \mathbb{R}. \end{aligned}$$

1.7.4. Basic properties of ring isomorphisms

By definition, in order to prove that some map f is a ring isomorphism, you have to check (1) that f is a ring morphism, (2) that f has an inverse, and (3) that this inverse f^{-1} is also a ring morphism. In truth, (3) is unnecessary, because of the following:

Proposition 1.7.6. Let R and S be two rings. Let $f : R \to S$ be an invertible ring morphism. Then, f is a ring isomorphism.

Proof. We need to show that f^{-1} is a ring morphism as well. So we need to show that $f^{-1}(c+d) = f^{-1}(c) + f^{-1}(d)$ for all $c, d \in S$, and likewise for multiplication, zero and unity. But this is not hard: Set $a = f^{-1}(c)$ and $b = f^{-1}(d)$ and recall that f is a ring morphism, so $f(a+b) = f(a) + f(b) = c + d$ (by the definitions of a and b), and thus $a + b = f^{-1}(c+d)$, so that $f^{-1}(c+d) = a + b = f^{-1}(c) + f^{-1}(d)$. □

Proposition 1.7.7. Let R, S and T be three rings. Let $f : S \to T$ and $g : R \to S$ be two ring isomorphisms. Then, $f \circ g : R \to T$ is a ring isomorphism as well.

Proposition 1.7.8. The inverse $f^{-1} : S \to R$ of a ring isomorphism $f : R \to S$ is a ring isomorphism.

Corollary 1.7.9. The relation $\cong$ for rings is an equivalence relation.

The most useful property of ring isomorphisms is the following "meta-theorem":

> **Isomorphism principle for rings:** Let R and S be two isomorphic rings. Then, any "ring-theoretic" property of R (that is, any property that does not refer to specific elements, but can be stated entirely in terms of ring operations) that holds for R must also hold for S.

What is a ring-theoretic property? Here are some examples and non-examples:

- "The ring R has 15 elements": yes.
- "The ring R is commutative": yes.
- "The ring R is a field": yes.
- "The ring R is a subring of $\mathbb{R}$": no.
- "There is an injective ring morphism from R to $\mathbb{R}$": yes.
- "There exists a field F and a ring morphism from R to F": yes.
- "The ring R is an integral domain": yes.

- "For any $a, b, c \in R$, we have $3abc\,(a+b+c) = 0$ (where 0 is the zero of R)" : yes.
- "The center of R has 10 elements": yes.
- "There exist two nonzero elements $a, b \in R$ such that $a^2 + b^2 = 0$": yes.
- "The set R is contains the complex number $i = \sqrt{-1}$": no.

Clearly, an isomorphism can destroy properties that are not ring-theoretical, just since it can send elements to different elements.

1.8. Ideals and kernels

1.8.1. Kernels

In linear algebra, you learn that linear maps have images (= ranges = column spaces) and kernels (= nullspaces); both are vector subspaces.

In ring theory, ring morphisms also have images and kernels. Images are subrings, but kernels are not (unless, like Jurij, you allow rings to be nonunital).

Let us recall the definition of a kernel:

Definition 1.8.1. Let R and S be two rings. Let $f : R \to S$ be a ring morphism. Then, the **kernel** of f is defined to be the set

$$\operatorname{Ker} f := \{r \in R \mid f(r) = 0_S\}.$$

This is a subset (but usually not a subring) of R.

Examples:

- Let $n \in \mathbb{Z}$. The kernel of the ring morphism
$$\begin{aligned} \pi : \mathbb{Z} &\to \mathbb{Z}/n, \\ a &\mapsto \overline{a} \end{aligned}$$
is the set $n\mathbb{Z} = \{\text{all multiples of } n\}$.
- Let R be a ring. Let S be any set. Recall the ring R^S of all functions from S to R (with pointwise $+$ and $\cdot$). Fix an element $s \in S$. Then, the kernel of the ring morphism
$$\begin{aligned} R^S &\to R, \\ f &\mapsto f(s) \end{aligned}$$
is the set of all functions $f \in R^S$ that vanish at s.
- The kernel of an injective ring morphism $f : R \to S$ is always $\{0_R\}$.

1.8.2. Ideals

Kernels of ring morphisms are not always subrings, but here is what they are:

Definition 1.8.2. Let R be a ring. An **ideal** of R means a subset I of R such that

- we have $a + b \in I$ for any $a, b \in I$ (that is, I is closed under addition);
- we have $ab \in I$ and $ba \in I$ for any $a \in R$ and $b \in I$ (that is, I is closed under multiplication by **arbitrary** elements of R – not just within itself);
- we have $0 \in I$ (where 0 means 0_R).

These three requirements are called the **ideal axioms**. The second is called **absorption**. Note that they imply that I is a nonunital subring of R. In particular, I must be an additive subgroup of R (that is, a subgroup of $(R, +, 0)$).

Theorem 1.8.3. Let R and S be two rings. Let $f : R \to S$ be a ring morphism. Then, $\operatorname{Ker} f$ is an ideal of R.

Proof. Let us check the absorption axiom (as the other two are similar).

Let $a \in R$ and $b \in \operatorname{Ker} f$. We must show that $ab \in \operatorname{Ker} f$ and $ba \in \operatorname{Ker} f$.

By assumption, $b \in \operatorname{Ker} f$, so that $f(b) = 0$. Now, since f is a ring morphism,

$$f(ab) = f(a) \cdot \underbrace{f(b)}_{=0} = 0,$$

so that $ab \in \operatorname{Ker} f$. Similarly, $ba \in \operatorname{Ker} f$, and we are done. □

We will soon see that this theorem has a converse: Any ideal of R can be written as the kernel of a ring morphism from R to some other ring. Thus, ideals and kernels are "the same thing, viewed from different angles".

1.8.3. Principal ideals

The simplest way to construct ideals in a commutative ring is by fixing an element and taking all its multiples:

Proposition 1.8.4. Let R be a **commutative** ring. Let $u \in R$. We define uR to be the set $\{ur \mid r \in R\}$. The elements of this set uR are called the **multiples** of u (in R).

Then, uR is an ideal of R. This ideal is known as a **principal ideal** of R. In particular, $0R = \{0\}$ and $1R = R$ are therefore principal ideals of R.

Proof. We must prove that uR is an ideal of R. So let's check the ideal axioms:

- Closure under addition: $ua + ub = u(a + b)$.
- Absorption: $(ua) \cdot b = u(ab)$ and $b \cdot (ua) = u(ba)$, the latter by commutativity. (Actually, it would suffice that u commutes with every element of R, meaning that u is central.)
- Zero: $0 = u \cdot 0$.

□

For example, $2\mathbb{Z} = \{\text{all even integers}\}$ is a principal ideal of $\mathbb{Z}$.

Principal ideals can also be defined for noncommutative rings, but this is more complicated. The simple definition $uR = \{ur \mid r \in R\}$ works nicely when u lies in the center of R.

1.8.4. Other examples of ideals

In the classical number rings $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, all ideals are principal (this will be proved below). Ideals get more interesting when the ring R is more complicated:

- Consider the set of all polynomials

$$\begin{aligned} f &\in \mathbb{Q}[x, y] \\ &= \{\text{polynomials with rational coefficients in } x, y\} \end{aligned}$$

  that have constant term 0. This set is an ideal of $\mathbb{Q}[x, y]$ (why?), but not a principal ideal (why?).
- Consider the set of all polynomials $f \in \mathbb{Z}[x]$ whose constant term is even. This set is an ideal of $\mathbb{Z}[x]$, but not a principal ideal.

1.9. Quotient rings

What follows is one of the most abstract topics in this course: the definition of quotient rings, and their basic properties.

Recall the idea behind modular arithmetic: By passing from the integers to their residue classes mod n (for a given $n \in \mathbb{Z}$), we are essentially equating n with 0, so that two integers become "equal" if they differ by a multiple of n. Thus, the residue classes mod n are "what remains" of the integers after equating n with 0. They form the ring $\mathbb{Z}/n$.

The same passage can be made in greater generality: We can start with any commutative ring R and any element u of R, and we can equate u with 0 in R. "What remains" of R is called the **quotient ring** R/u.

Even more generally: For any ring R and any ideal I of R, we can equate all the elements of I with 0. The result is a new ring, called the **quotient ring** R/I. If I is a principal ideal uR, then R/I is just R/u.

This was the rough idea behind quotient rings. We will now define them rigorously.

1.9.1. Quotient groups

We don't have to reinvent the wheel: You have already seen residue classes in a first course on groups; in that context, they are known as "**cosets**". We will only need to turn them into a ring.

Let me recall the definition of cosets:

- If H is a subgroup of a group G, then the **left cosets** of H in G are the subsets
$$gH := \{gh \mid h \in H\} \qquad \text{for all } g \in G.$$
There is one left coset gH for each $g \in G$; but different g's often lead to the same gH. Thus, there are usually fewer left cosets than elements of G. The set of all left cosets of H is called G/H.

- If G is an abelian group, then we can also rewrite gH as Hg and refer to our left cosets as right cosets, or just simply as **cosets**. We will apply this theory to the case when G is the additive group $(R, +, 0)$ of a ring R, which of course is abelian, so we will just speak of cosets. But we should call them $g + H$ rather than gH because the group operation is now $+$. So they now have the form
$$g + H := \{g + h \mid h \in H\} \qquad \text{for all } g \in R.$$

- Let G be an **additive** group (i.e., it is abelian, and its binary operation is called $+$). Then, the cosets of H in G are denoted by $g + H$ rather than gH. We can define an addition on these cosets by setting
$$(g_1 + H) + (g_2 + H) = g_1 + g_2 + H \qquad \text{for all } g_1, g_2 \in G.$$
This turns G/H (that is, the set of all these cosets $g + H$) into an additive group with neutral element $0_G + H$. This group G/H is called the **quotient group** of G by H.

- Best-known example: $\mathbb{Z}/n\mathbb{Z}$, also called $\mathbb{Z}/n$. This is the quotient group of the additive group $\mathbb{Z}$ by its subgroup $n\mathbb{Z} = \{\text{all multiples of } n\}$. The cosets here are known as residue classes modulo n. This quotient $\mathbb{Z}/n$ is called the **cyclic group of order** n, at least when n is positive.

  This construction accounts for the addition on $\mathbb{Z}/n$, but not for the multiplication.

1.9.2. Quotient rings

Now, piggybacking on the construction of quotient groups we just recalled, we shall define a similar concept of quotient rings (generalizing $\mathbb{Z}/n$, now including its multiplication). Instead of subgroups, we will now use ideals:

Definition 1.9.1. Let I be an ideal of a ring R. Then, I is a subgroup of the additive group $(R, +, 0)$. Thus, the quotient group R/I is a well-defined additive group. Its elements are the cosets $r + I$ for $r \in R$. These cosets are called the **residue classes** modulo I. A coset $r + I$ is also denoted by $\overline{r}$ or $[r]$ or $[r]_I$ or $r \bmod I$. (We will only use the notations $r + I$ and $\overline{r}$.)

Note that the addition on R/I is defined by

$$(a + I) + (b + I) = a + b + I \qquad \text{for all } a, b \in R.$$

Now, we define a multiplication on R/I by setting

$$(a + I)(b + I) = ab + I \qquad \text{for all } a, b \in R.$$

(We will prove below that this is well-defined.)

The set R/I, equipped with the addition and the multiplication we just introduced, and with the zero $0 + I$ and the unity $1 + I$, is a ring (as we will soon see). This ring is called the **quotient ring** of R by the ideal I, and is denoted by R/I. It is pronounced "R **modulo** I".

Note that the rules

$$(a + I) + (b + I) = a + b + I \qquad \text{for all } a, b \in R$$

and

$$(a + I)(b + I) = ab + I \qquad \text{for all } a, b \in R$$

can be rewritten in the more familiar form

$$\overline{a} + \overline{b} = \overline{a + b} \qquad \text{for all } a, b \in R;$$
$$\overline{a} \cdot \overline{b} = \overline{ab} \qquad \text{for all } a, b \in R.$$

Before we prove the above definition (well, the claims made therein), let us see a few examples:

- Let $n \in \mathbb{Z}$. Then, the set $n\mathbb{Z} = \{\text{all multiples of } n\}$ is an ideal of $\mathbb{Z}$ (a principal ideal, in fact). The quotient ring $\mathbb{Z}/n\mathbb{Z}$ is exactly the ring $\mathbb{Z}/n$ of residue classes modulo n. So our definition of R/I is a generalization of $\mathbb{Z}/n$, replacing the integers by the ring R and replacing the multiples of n by the elements of I.

- Two stupid general examples:

  Recall that every ring R has at least the ideals $\{0_R\}$ and R. What are the respective quotient rings?

  - The quotient ring $R/\{0_R\}$ is isomorphic to R. Indeed, each residue class modulo $\{0_R\}$ has the form $r + \{0_R\} = \{r\}$, which is a 1-element set. Hence, there is an obvious bijection from R to $R/\{0_R\}$ that sends each element r to $\{r\}$. This is a ring isomorphism.
  - The quotient ring R/R is a trivial ring, i.e., isomorphic to the zero ring. Indeed, there is only one residue class, say $0 + R$, which contains all the elements of R, so that the quotient ring R/R has just one element.

- Let R be the ring $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ of Gaussian integers. Consider its principal ideal

$$\begin{aligned} 3R &= \{3r \mid r \in R\} \\ &= \{3a + 3bi \mid a, b \in \mathbb{Z}\} \\ &= \{c + di \mid c, d \in \mathbb{Z} \text{ are multiples of } 3\}. \end{aligned}$$

  What is the quotient ring $R/3R$? The elements of this ring have the form

$$\overline{a + bi} \qquad \text{with } a, b \in \{0, 1, 2\}$$

  (do not mistake the line over the $a + bi$ for the "complex conjugate" notation). In fact, any Gaussian integer can be reduced to a Gaussian integer of the form $a + bi$ with $a, b \in \{0, 1, 2\}$ by subtracting appropriate Gaussian-integer multiples of 3 from its real part and from its imaginary part:

$$\overline{5 + 8i} = \overline{2 + 2i}, \qquad \text{since } (5 + 8i) - (2 + 2i) = 3 + 6i = 3(1 + 2i) \in 3R.$$

  In other words,

$$R/3R = \left\{\overline{0},\ \overline{1},\ \overline{2},\ \overline{i},\ \overline{1 + i},\ \overline{2 + i},\ \overline{2i},\ \overline{1 + 2i},\ \overline{2 + 2i}\right\}.$$

  It is easy to see that these 9 elements are actually distinct (for instance, the difference $(2 + i) - 2i = 2 - 2i$ is not a Gaussian-integer multiple of 3, so we have $\overline{2 + i} \neq \overline{2i}$).

  Furthermore, it is easy to see that all of these 9 elements, except for $\overline{0}$, are units of $R/3R$. Thus, $R/3R$ is a field with 9 elements.

  Let us do some computations in this field:

$$\begin{aligned} \overline{2 + i} + \overline{2 + 2i} &= \overline{(2 + i) + (2 + 2i)} = \overline{4 + 3i} = \overline{1}; \\ \overline{2 + i} \cdot \overline{2 + 2i} &= \overline{(2 + i)(2 + 2i)} = \overline{2 + 6i} = \overline{2}; \\ \overline{2 + i} \cdot \overline{1 + i} &= \overline{(2 + i)(1 + i)} = \overline{1 + 3i} = \overline{1}. \end{aligned}$$

If we replace 3 by any other positive integer n, then the quotient ring R/nR will be a finite ring with n^2 elements. But it will not always be a field. For instance, for $n = 5$, we have

$$\overline{1+2i} \cdot \overline{1-2i} = \overline{(1+2i)(1-2i)} = \overline{1+4} = \overline{5} = \overline{0} \qquad \text{in } R/5R,$$

which shows that $\overline{1+2i}$ and $\overline{1-2i}$ cannot be units (since they are nonzero). So $R/5R$ is not a field.

We will learn more about when R/nR is a field later on.

- Again take $R = \mathbb{Z}[i]$, but now consider the quotient ring $R/((1+i)R)$. How many elements does it have? The answer is 2, but the reason is not that obvious, since we need to understand which Gaussian integers belong to $(1+i)R$.

  Here is one way to prove the answer:

  1. Observe: $2 \in (1+i)R$ (because $2 = (1+i)(1-i)$). Thus, every Gaussian integer can be reduced to a Gaussian integer of the form $a+bi$ with $a, b \in \{0, 1\}$ by adding an element of $(1+i)R$.
  2. Thus, $R/((1+i)R) = \left\{\overline{0}, \overline{1}, \overline{i}, \overline{1+i}\right\}$.
  3. Furthermore, $\overline{1} = \overline{i}$ (since $1 - i = -i(1+i) \in (1+i)R$) and $\overline{1+i} = \overline{0}$ (since $1 + i \in (1+i)R$).
  4. Thus, $R/((1+i)R) = \left\{\overline{0}, \overline{1}\right\}$.
  5. Finally, $\overline{0} \neq \overline{1}$, since $0 - 1$ is not a multiple of $1 + i$ (because $\dfrac{0-1}{1+i} = \dfrac{-1}{1+i} = -\dfrac{1}{2} + \dfrac{1}{2}i \notin R$). So $R/((1+i)R)$ consists of the two distinct elements $\overline{0}$ and $\overline{R}$.

  Actually, $R/((1+i)R) \cong \mathbb{Z}/2$ as a ring.

  Can we analyze $R/((7+5i)R)$ likewise? How many elements does this have? Tricky question. Start with

  $$(7+5i)(7-5i) = 7^2 + 5^2 = 74,$$

  so at least we know that every element of $R/((7+5i)R)$ can be written as $a+bi$ with $a, b \in \{0, 1, \ldots, 73\}$. But what then? This will be an exercise later on, once we've learned a few more things about rings.

More examples appear in the text. For now, let me quickly go over the proof of well-definedness of quotient rings:

Theorem 1.9.2. Let R be a ring, and let I be an ideal of R. Then, the quotient ring R/I is a well-defined ring.

Proof. We need to show two things:

1. that the operations $+$ and $\cdot$ on R/I are well-defined;
2. that they satisfy the ring axioms.

Part 2 is straightforward (all the ring axioms are inherited from R: for example, $\overline{a} \cdot \left(\overline{b} \cdot \overline{c}\right) = \left(\overline{a} \cdot \overline{b}\right) \cdot \overline{c}$ follows from $a \cdot (b \cdot c) = (a \cdot b) \cdot c$).

So we only need to do Part 1. For the operation $+$, we already know from group theory that it is well-defined. Thus, we only need to prove it for the operation $\cdot$.

Recall that the operation $\cdot$ was defined by

$$(a + I)(b + I) = ab + I \qquad \text{for all } a, b \in R.$$

Thus, "well-defined" means that the right hand side $ab + I$ depends not on the specific elements a and b but only on their cosets $a + I$ and $b + I$. In other words, it means that if we write a given coset x as $x = a_1 + I = a_2 + I$ for two elements $a_1, a_2 \in R$, and if we write a given coset y as $y = b_1 + I = b_2 + I$ for two elements $b_1, b_2 \in R$, then $a_1 b_1 + I = a_2 b_2 + I$ (so that we get the same value for xy no matter which of our presentations of x and y we are using).

So let us prove this. We must show that $a_1 b_1 + I = a_2 b_2 + I$.

(For comparison: When $R = \mathbb{Z}$ and $I = n\mathbb{Z}$, then we are proving that $a_1 \equiv a_2 \bmod n$ and $b_1 \equiv b_2 \bmod n$ imply $a_1 b_1 \equiv a_2 b_2 \bmod n$. This is a classical fact in elementary number theory, and can be proved e.g. by

$$a_1 b_1 \equiv a_1 b_2 \qquad \left(\text{since } a_1 b_1 - a_1 b_2 = a_1 \underbrace{(b_1 - b_2)}_{\in n\mathbb{Z}} \in n\mathbb{Z} \right)$$

$$\equiv a_2 b_2 \bmod n \qquad \left(\text{since } a_1 b_2 - a_2 b_2 = \underbrace{(a_1 - a_2)}_{\in n\mathbb{Z}} b_2 \in n\mathbb{Z} \right).$$

)

In the general case (R and I arbitrary), we argue similarly: We have

$$a_1 b_1 + I = a_1 b_2 + I$$

since $a_1 b_1 - a_1 b_2 = a_1 \underbrace{(b_1 - b_2)}_{\substack{\in I \\ (\text{since } b_1 + I = b_2 + I)}} \in I$ (by the absorption axiom). We have

$$a_1 b_2 + I = a_2 b_2 + I$$

since $a_1 b_2 - a_2 b_2 = \underbrace{(a_1 - a_2)}_{\in I} b_2 \in I$ (by the absorption axiom). So altogether

$$a_1 b_1 + I = a_1 b_2 + I = a_2 b_2 + I,$$

as desired. (See the notes for a slightly different proof.)

So we have shown that $\cdot$ is well-defined. As we said, this completes the proof. □

1.9.3. More examples of quotient rings

Here are some more examples of quotient rings.

- As we recall, if R is a ring and $n \in \mathbb{N}$ is an integer, then

$$\begin{aligned} R^{n \leq n} &= \{\text{all upper-triangular } n \times n\text{-matrices with entries in } R\} \\ &= \left\{ \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ & a_{2,2} & \cdots & a_{2,n} \\ & & \ddots & \vdots \\ & & & a_{n,n} \end{pmatrix} \;\mid\; a_{i,j} \in R \text{ for all } i \leq j \right\} \end{aligned}$$

  (where empty spaces stand for entries that are 0) is a ring.

  Consider the special case $R = \mathbb{Q}$ and $n = 3$ for simplicity. Thus,

$$R^{n \leq n} = \mathbb{Q}^{3 \leq 3} = \left\{ \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \;\mid\; a, b, c, d, e, f \in \mathbb{Q} \right\}.$$

  I claim that the subset

$$\begin{aligned} \mathbb{Q}^{3<3} &= \{\text{all strictly upper-triangular } 3 \times 3\text{-matrices with entries in } \mathbb{Q}\} \\ &= \left\{ \begin{pmatrix} 0 & b & c \\ 0 & 0 & e \\ 0 & 0 & 0 \end{pmatrix} \;\mid\; b, c, e \in \mathbb{Q} \right\} \end{aligned}$$

  is an ideal of $\mathbb{Q}^{3 \leq 3}$. To see this, we verify absorption:

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \begin{pmatrix} 0 & x & y \\ 0 & 0 & z \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & ax & ay + bz \\ 0 & 0 & dz \\ 0 & 0 & 0 \end{pmatrix};$$

$$\begin{pmatrix} 0 & x & y \\ 0 & 0 & z \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} 0 & dx & xe + fy \\ 0 & 0 & fz \\ 0 & 0 & 0 \end{pmatrix}.$$

Alternatively (and more generally), we can prove this by observing that when we multiply two triangular matrices, their diagonal entries just get multiplied:

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \begin{pmatrix} a' & b' & c' \\ 0 & d' & e' \\ 0 & 0 & f' \end{pmatrix} = \begin{pmatrix} aa' & bd'+ab' & be'+cf'+ac' \\ 0 & dd' & ef'+de' \\ 0 & 0 & ff' \end{pmatrix}.$$

So if one of the matrices has zeros on its diagonal, then so will the product.

Now what is the quotient ring $\mathbb{Q}^{3\leq 3}/\mathbb{Q}^{3<3}$? The elements of this quotient ring are cosets

$$\overline{A} = A + \mathbb{Q}^{3<3} \qquad \text{for } A \in \mathbb{Q}^{3\leq 3}.$$

For instance, if $A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}$, then

$$\begin{aligned}
\overline{A} &= \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix} + \mathbb{Q}^{3<3} \\
&= \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix} + \left\{ \begin{pmatrix} 0 & b & c \\ 0 & 0 & e \\ 0 & 0 & 0 \end{pmatrix} \;\middle|\; b,c,e \in \mathbb{Q} \right\} \\
&= \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix} + \begin{pmatrix} 0 & b & c \\ 0 & 0 & e \\ 0 & 0 & 0 \end{pmatrix} \;\middle|\; b,c,e \in \mathbb{Q} \right\} \\
&= \left\{ \begin{pmatrix} 1 & 2+b & 3+c \\ 0 & 4 & 5+e \\ 0 & 0 & 6 \end{pmatrix} \;\middle|\; b,c,e \in \mathbb{Q} \right\} \\
&= \left\{ \begin{pmatrix} 1 & x & y \\ 0 & 4 & z \\ 0 & 0 & 6 \end{pmatrix} \;\middle|\; x,y,z \in \mathbb{Q} \right\} \\
&= \begin{pmatrix} 1 & \mathbb{Q} & \mathbb{Q} \\ 0 & 4 & \mathbb{Q} \\ 0 & 0 & 6 \end{pmatrix},
\end{aligned}$$

where the $\mathbb{Q}$s mean "you can put arbitrary elements of $\mathbb{Q}$ here". So you can think of $\overline{A} = \begin{pmatrix} 1 & \mathbb{Q} & \mathbb{Q} \\ 0 & 4 & \mathbb{Q} \\ 0 & 0 & 6 \end{pmatrix}$ as a "matrix" in which the three entries above the diagonal are undetermined. Formally, it is a set of matrices.

The rules for adding and multiplying such "partly determined matrices"

are what you would expect:

$$\begin{pmatrix} a & \mathbb{Q} & \mathbb{Q} \\ 0 & b & \mathbb{Q} \\ 0 & 0 & c \end{pmatrix} + \begin{pmatrix} d & \mathbb{Q} & \mathbb{Q} \\ 0 & e & \mathbb{Q} \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} a+d & \mathbb{Q} & \mathbb{Q} \\ 0 & b+e & \mathbb{Q} \\ 0 & 0 & c+f \end{pmatrix};$$
$$\begin{pmatrix} a & \mathbb{Q} & \mathbb{Q} \\ 0 & b & \mathbb{Q} \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} d & \mathbb{Q} & \mathbb{Q} \\ 0 & e & \mathbb{Q} \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} ad & \mathbb{Q} & \mathbb{Q} \\ 0 & be & \mathbb{Q} \\ 0 & 0 & cf \end{pmatrix}.$$

These equalities look exactly like the rules for adding and multiplying diagonal matrices:

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} + \begin{pmatrix} d & 0 & 0 \\ 0 & e & 0 \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} a+d & 0 & 0 \\ 0 & b+e & 0 \\ 0 & 0 & c+f \end{pmatrix};$$
$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} d & 0 & 0 \\ 0 & e & 0 \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} ad & 0 & 0 \\ 0 & be & 0 \\ 0 & 0 & cf \end{pmatrix}.$$

So the quotient ring $\mathbb{Q}^{3\leq 3}/\mathbb{Q}^{3<3}$ is isomorphic to the ring

$$\mathbb{Q}^{3=3} = \{\text{diagonal } 3\times 3\text{-matrices over } \mathbb{Q}\}.$$

Formally, the map

$$\mathbb{Q}^{3\leq 3}/\mathbb{Q}^{3<3} \to \mathbb{Q}^{3=3},$$
$$\begin{pmatrix} a & \mathbb{Q} & \mathbb{Q} \\ 0 & b & \mathbb{Q} \\ 0 & 0 & c \end{pmatrix} \mapsto \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$

is a ring isomorphism.

- Here is a slightly more interesting example. Again consider the ring $\mathbb{Q}^{3\leq 3}$ of upper-triangular 3×3-matrices over $\mathbb{Q}$, but now take the smaller ideal

$$\mathbb{Q}^{3<<3} = \left\{ \begin{pmatrix} 0 & 0 & y \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mid y \in \mathbb{Q} \right\}$$

of $\mathbb{Q}^{3\leq 3}$. What is the quotient ring $\mathbb{Q}^{3\leq 3}/\mathbb{Q}^{3<<3}$? A residue class $\overline{A} =$

$A + \mathbb{Q}^{3<<3}$ in this ring looks as follows:

$$\begin{aligned}
\overline{A} &= \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} + \mathbb{Q}^{3<<3} \\
&= \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} + \left\{ \begin{pmatrix} 0 & 0 & y \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mid y \in \mathbb{Q} \right\} \\
&= \left\{ \begin{pmatrix} a & b & c+y \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \mid y \in \mathbb{Q} \right\} \\
&= \begin{pmatrix} a & b & \mathbb{Q} \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \qquad \text{(with notation as before)}.
\end{aligned}$$

The rules for adding and multiplying these residue classes are

$$\begin{pmatrix} a & b & \mathbb{Q} \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} + \begin{pmatrix} a' & b' & \mathbb{Q} \\ 0 & d' & e' \\ 0 & 0 & f' \end{pmatrix} = \begin{pmatrix} a+a' & b+b' & \mathbb{Q} \\ 0 & d+d' & e+e' \\ 0 & 0 & f+f' \end{pmatrix};$$

$$\begin{pmatrix} a & b & \mathbb{Q} \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \begin{pmatrix} a' & b' & \mathbb{Q} \\ 0 & d' & e' \\ 0 & 0 & f' \end{pmatrix} = \begin{pmatrix} aa' & ab'+bd' & \mathbb{Q} \\ 0 & bb' & de'+ef' \\ 0 & 0 & cc' \end{pmatrix}.$$

This is no longer a subring of $\mathbb{Q}^{3\times 3}$ in disguise. Indeed, if we replace the $\mathbb{Q}$s by 0s, then we get

$$\begin{aligned}
\begin{pmatrix} a & b & 0 \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \begin{pmatrix} a' & b' & 0 \\ 0 & d' & e' \\ 0 & 0 & f' \end{pmatrix} &= \begin{pmatrix} aa' & ab'+bd' & be' \\ 0 & bb' & de'+ef' \\ 0 & 0 & cc' \end{pmatrix} \\
&\neq \begin{pmatrix} aa' & ab'+bd' & 0 \\ 0 & bb' & de'+ef' \\ 0 & 0 & cc' \end{pmatrix},
\end{aligned}$$

so that our "partly determined matrices" do not multiply like some kind of actual fully-determined 3×3-matrices. And so the quotient ring $\mathbb{Q}^{3\leq 3}/\mathbb{Q}^{3<<3}$ is a genuinely new ring, not just a subring in disguise.

- There are many more complicated examples. In our quotient rings above, the $\mathbb{Q}$s were mutually independent. But there can be partially determined matrices whose undetermined entries nevertheless must satisfy some kind of relation; we cannot just denote them by $\mathbb{Q}$s any more.

1.9.4. The canonical projection

Back a few lectures ago, we said that the kernel of any ring morphism is an ideal. Now we will prove the converse: Any ideal is a kernel. Even better:

Theorem 1.9.3. Let R be a ring. Let I be an ideal of R. Consider the map

$$\begin{aligned} \pi : R &\to R/I, \\ r &\mapsto \overline{r} = r + I. \end{aligned}$$

This map π is a surjective ring morphism with kernel I.

This map π is called the **canonical projection** from R to R/I.

Proof. We need to prove that:

1. the map π is a ring morphism (i.e., respects addition, multiplication, zero and unity);
2. the map π is surjective;
3. we have $I = \operatorname{Ker} \pi$.

Part 1 is straightforward: To show that π respects multiplication, we must check that $\pi(ab) = \pi(a) \cdot \pi(b)$. In light of the definition of π, this rewrites as $\overline{ab} = \overline{a} \cdot \overline{b}$ (aka $ab + I = (a + I) \cdot (b + I)$). But this is true since we defined the $\cdot$ on R/I by this exact formula. Similarly, π respects all the other things.

Part 2 is trivial: Each element of R/I is a residue class, so by definition it has the form $\overline{r}$ for some $r \in R$.

Remains Part 3. For this, we observe that

$$\begin{aligned} \operatorname{Ker} \pi &= \{r \in R \mid \pi(r) = 0_{R/I}\} \\ &= \left\{r \in R \mid \overline{r} = \overline{0}\right\} \\ &= \{r \in R \mid r + I = 0 + I\} \\ &= \{r \in R \mid r - 0 \in I\} \\ &= \{r \in R \mid r \in I\} = I. \end{aligned}$$

□

For example, if we take $R = \mathbb{Z}$ and $I = 2\mathbb{Z}$, then the canonical projection $\pi : \mathbb{Z} \to \mathbb{Z}/2$ is the map that sends each even integer to $\overline{0}$ and each odd integer to $\overline{1}$. In other words, it assigns to each integer its parity (as an element of $\mathbb{Z}/2$).

1.9.5. The universal property of quotient rings

When trying to understand a quotient ring R/I, it can be helpful to construct ring morphisms into and out of it.

Constructing a morphism $\alpha : S \to R/I$ into a quotient ring R/I is generally easy (we just did so in the above proof).

Constructing a morphism $\beta : R/I \to S$ out of a quotient ring R/I is harder: Not only do you have to specify $\beta(\overline{r})$ for each residue class $\overline{r}$, but you also need to make sure that this value $\beta(\overline{r})$ depends only on the class $\overline{r}$ and not on the chosen representative r. This is called "well-definedness", and often takes some work to verify, since one and the same residue class can be written as $\overline{r}$ for different r's.

This can be done by hand, but it is work. The **universal property of quotient rings** is a theorem that does some of this work for you. It gives a way to define a ring morphism $\beta : R/I \to S$ by providing a ring morphism $f : R \to S$ and showing that $f(I) = 0$ (that is, f sends all elements of I to 0). Once you have done this part, the theorem automatically gives you a ring morphism $f' : R/I \to S$ that sends each residue class $\overline{r} \in R/I$ to $f(r)$. Here is the precise statement:

Theorem 1.9.4 (Universal property of quotient rings)**.** Let R be a ring. Let I be an ideal of R.

Let S be a ring. Let $f : R \to S$ be a ring morphism. Assume that $f(I) = 0$ (that is, $f(i) = 0$ for all $i \in I$). Then, the map

$$\begin{aligned} f' : R/I &\to S, \\ \overline{r} &\mapsto f(r) \end{aligned}$$

is well-defined (i.e., the value $f(r)$ depends only on the residue class $\overline{r}$ and not on r itself) and is a ring morphism.

Before we prove this, an example:

- Consider the canonical projections

$$\begin{aligned} \pi_6 : \mathbb{Z} &\to \mathbb{Z}/6, \\ r &\mapsto r + 6\mathbb{Z} \end{aligned}$$

and

$$\begin{aligned} \pi_3 : \mathbb{Z} &\to \mathbb{Z}/3, \\ r &\mapsto r + 3\mathbb{Z}. \end{aligned}$$

(We don't write $\overline{r}$ for the residue classes, because that would give the same notation $\overline{r}$ to the two different classes $r + 6\mathbb{Z}$ and $r + 3\mathbb{Z}$).

Then, $\pi_3(6\mathbb{Z}) = 0$ (since all multiples of 6 are multiples of 3). Therefore, the universal property stated above (applied to $R = \mathbb{Z}$ and $I = 6\mathbb{Z}$ and $\pi = \pi_3$) shows that the map

$$\begin{aligned} \pi_3' : \mathbb{Z}/6 &\to \mathbb{Z}/3, \\ r + 6\mathbb{Z} &\mapsto r + 3\mathbb{Z} \end{aligned}$$

is a ring morphism. Explicitly, this morphism π_3' sends

$$\begin{aligned} &\text{the mod-6 residue classes } \overline{0},\ \overline{1},\ \overline{2},\ \overline{3},\ \overline{4},\ \overline{5} \\ &\text{to the mod-3 residue classes } \overline{0},\ \overline{1},\ \overline{2},\ \overline{3},\ \overline{4},\ \overline{5} \\ &\text{(that is, to the mod-3 residue classes } \overline{0},\ \overline{1},\ \overline{2},\ \overline{0},\ \overline{1},\ \overline{2}). \end{aligned}$$

More generally, if n and m are two integers such that $m \mid n$, then there is a ring morphism

$$\begin{aligned} \mathbb{Z}/n &\to \mathbb{Z}/m, \\ \overline{r} = r + n\mathbb{Z} &\mapsto \overline{r} = r + m\mathbb{Z}. \end{aligned}$$

This follows from the universal property, applied to $R = \mathbb{Z}$, $I = n\mathbb{Z}$, $S = \mathbb{Z}/m$ and $f = \pi_m : \mathbb{Z} \to \mathbb{Z}/m$.

Incidentally, this accounts for all ring morphisms that go between quotient ring of $\mathbb{Z}$. That is: For two integers n and m, there is a ring morphism from $\mathbb{Z}/n$ to $\mathbb{Z}/m$ if and only if $m \mid n$. In that case, there is only one, namely the one we just found. Proving this is a nice exercise.

Proof of the universal property of quotient rings. First, we must show that f' is well-defined (i.e., the formula $f'(\overline{r}) = f(r)$ does not assign two different output values to the same input).

That is: We must show that if two elements $a, b \in R$ satisfy $\overline{a} = \overline{b}$ in R/I, then $f(a) = f(b)$.

So let $a, b \in R$ be two elements such that $\overline{a} = \overline{b}$ in R/I. This assumption $\overline{a} = \overline{b}$ in R/I is just saying that $a - b \in I$. Hence, $f(a-b) \in f(I) = 0$, so that $f(a-b) = 0$. But f is a ring morphism, thus respects differences. So $f(a) - f(b) = f(a-b) = 0$. In other words, $f(a) = f(b)$.

So we have shown that f' is well-defined. We still need to show that f' is a ring morphism. In other words, we must show that f' respects addition, multiplication, zero and unity. Let me show that f' respects multiplication; the rest is analogous.

So we must show that $f'(xy) = f'(x) \cdot f'(y)$ for all $x, y \in R/I$.

Fix $x, y \in R/I$. Write the residue classes x, y as $x = \overline{a}$ and $y = \overline{b}$ for some $a, b \in R$. Then, $xy = \overline{a} \cdot \overline{b} = \overline{ab}$, so that the definition of f' yields

$$f'(xy) = f'\left(\overline{ab}\right) = f(ab) = f(a) \cdot f(b) \qquad \text{(since } f \text{ is a ring morphism)}.$$

Comparing this with

$$f'(x) \cdot f'(y) = f'(\overline{a}) \cdot f'\left(\overline{b}\right) = f(a) \cdot f(b) \qquad \left(\text{by the definition of } f'\right),$$

we obtain $f'(xy) = f'(x) \cdot f'(y)$, just as we wanted to prove. So we have shown that f' respects multiplication. With similar arguments for the other axioms, we conclude that f' is a ring morphism. □

So we have proved the universal property of quotient rings. For various reasons, it is helpful to have a restatement of this property that does not talk about elements but instead "implicitly" describes f' by an equality:

Theorem 1.9.5 (Universal property of quotient rings, abstract/element-free form)**.** Let R be a ring. Let I be an ideal of R. Let $\pi : R \to R/I$ be the canonical projection.

Let S be a ring. Let $f : R \to S$ be a ring morphism. Assume that $f(I) = 0$ (that is, $f(i) = 0$ for all $i \in I$). Then, there is a unique ring morphism $f' : R/I \to S$ such that

$$f = f' \circ \pi.$$

Proof. The previous version of the universal property shows that there is a unique ring morphism $f' : R/I \to S$ that satisfies

$$f'(\overline{r}) = f(r) \qquad \text{for all } r \in R.$$

Now I claim that this latter condition is equivalent to

$$f = f' \circ \pi.$$

Indeed, we have the following chain of equivalences:

$$\begin{aligned}
&\left(f = f' \circ \pi\right) \\
&\iff \left(f(r) = \left(f' \circ \pi\right)(r) \text{ for all } r \in R\right) \\
&\iff \left(f(r) = f'(\pi(r)) \text{ for all } r \in R\right) \\
&\iff \left(f(r) = f'(\overline{r}) \text{ for all } r \in R\right) \qquad (\text{since } \pi(r) = \overline{r}) \\
&\iff \left(f'(\overline{r}) = f(r) \text{ for all } r \in R\right).
\end{aligned}$$

□

The equality $f = f' \circ \pi$ can be restated as "the diagram on the whiteboard (or in the lecture notes) commutes / is commutative". In general, a **diagram** is a bunch of sets (drawn as nodes) and a bunch of maps between these sets (drawn as arrows). In our case, the sets are R, R/I and S, and the maps are f, f' and π. We say that a diagram **commutes** (or **is commutative**) if, for any two nodes in the diagram, all ways of going from the first to the second node result in the same composed map.

1.9.6. Injectivity means zero kernel

Taking a break from these abstractions, let us prove a simple lemma about injectivity of ring morphisms:

Lemma 1.9.6. Let R and S be two rings. Let $f : R \to S$ be a ring morphism. Then, f is injective if and only if $\operatorname{Ker} f = \{0_R\}$.

Proof. $\Longleftarrow$: Assume that $\operatorname{Ker} f = \{0_R\}$. Let $a, b \in R$ be such that $f(a) = f(b)$. We want to show that $a = b$. Since f respects differences, we have $f(a-b) = f(a) - f(b) = 0$ (since $f(a) = f(b)$). In other words, $a - b \in \operatorname{Ker} f = \{0_R\}$, so that $a - b = 0_R$ and thus $a = b$.

This shows that f is injective.

$\Longrightarrow$: Assume that f is injective. If $a \in \operatorname{Ker} f$, then $f(a) = 0_S = f(0_R)$, so that $a = 0_R$ by injectivity of f. So we conclude that $\operatorname{Ker} f \subseteq \{0_R\}$. Thus, $\operatorname{Ker} f = \{0_R\}$ (since $0_R \in \operatorname{Ker} f$). □

Similar lemmas (with the same proof) hold for group morphisms and vector space morphisms (= linear maps).

1.9.7. The First Isomorphism Theorem for sets

The next topic is again more abstract. We will state and prove the **First Isomorphism Theorem** for rings. But first, we state its analogue for sets, which is really basic and merely serves as a simile.

Consider a map $f : R \to S$ from some set R to some set S. Then, I claim that there is a bijection (= bijective map) hiding inside f.

I mean that we can write f as a composition $f = \iota \circ f' \circ \pi$ of

- a surjection π from R to a certain set of equivalence classes;
- a bijection f' between this set and the image $f(R)$ of f;
- the inclusion ι from $f(R)$ into S.

The set of equivalence classes is

$$R/f := \{\text{equivalence classes of elements of } R \text{ under } \sim\},$$

where $\sim$ is the equivalence relation given by

$$(a \sim b) \Longleftrightarrow (f(a) = f(b)).$$

The surjection $\pi : R \to R/f$ simply sends each $r \in R$ to its equivalence class $\overline{r}$. The bijection $f' : R/f \to f(R)$ is given by $f'(\overline{r}) = f(r)$ for all $r \in R$ (just as in the universal property). This is a bijection, since the taking of equivalence classes gets rid of the non-injectivity of f, whereas the restriction of the codomain to $f(R)$ instead of S ensures that our map becomes surjective. The formal proof is not much harder. Let us state the result as a theorem:

Theorem 1.9.7 (First Isomorphism Theorem for sets)**.** Let R and S be two sets, and let $f : R \to S$ be any map.

Let $\sim$ be the binary relation on the set R defined by

$$(a \sim b) \Longleftrightarrow (f(a) = f(b)).$$

(a) This relation $\sim$ is an equivalence relation.

Let us refer to this relation $\sim$ as **f-equivalence**, and to its equivalence classes as **f-classes**. Let R/f denote the set of all f-classes. For any $r \in R$, we let $\overline{r}$ denote the f-class that contains r.

(b) The image $f(R)$ is a subset of S.

(c) The map

$$\begin{aligned} f' : R/f &\to f(R), \\ \overline{r} &\mapsto f(r) \end{aligned}$$

is well-defined and bijective.

(d) Let $\pi : R \to R/f$ denote the **canonical projection** (i.e., the map that sends each r to $\overline{r}$). Let $\iota : f(R) \to S$ denote the **canonical inclusion** (i.e., the map that sends each s to s). Then, the map f' in part **(c)** satisfies

$$f = \iota \circ f' \circ \pi.$$

In other words, the diagram

$$\begin{array}{ccc} R & \xrightarrow{f} & S \\ {\scriptstyle\pi}\downarrow & & \uparrow{\scriptstyle\iota} \\ R/f & \xrightarrow[f']{} & f(R) \end{array}$$

is commutative.

1.9.8. The First Isomorphism Theorem for rings

Now let us extend the above theorem to rings.

Theorem 1.9.8 (First Isomorphism Theorem for rings, elementwise form)**.** Let R and S be two rings, and let $f : R \to S$ be a ring morphism. Then:

(a) The kernel $\operatorname{Ker} f$ is an ideal of R. Thus, $R/\operatorname{Ker} f$ is a quotient ring of R. As a set, $R/\operatorname{Ker} f$ is precisely the set R/f defined in the previous theorem. The f-classes are precisely the cosets of $\operatorname{Ker} f$.

(b) The image $f(R) := \{f(r) \mid r \in R\}$ of f is a subring of S.

(c) The map

$$f' : R/\operatorname{Ker} f \to f(R), \\ \overline{r} \mapsto f(r)$$

is well-defined and is a ring isomorphism.

(d) This map f' is precisely the map f' defined in the previous theorem.

(e) Let $\pi : R \to R/\operatorname{Ker} f$ denote the **canonical projection** (i.e., the map that sends each r to $\overline{r}$). Let $\iota : f(R) \to S$ denote the **canonical inclusion** (i.e., the map that sends each s to s). Then, the map f' in part **(c)** satisfies

$$f = \iota \circ f' \circ \pi.$$

In other words, the diagram

$$\begin{array}{ccc} R & \xrightarrow{f} & S \\ {\scriptstyle\pi}\downarrow & & \uparrow{\scriptstyle\iota} \\ R/\operatorname{Ker} f & \xrightarrow[f']{} & f(R) \end{array}$$

is commutative.

(f) We have $R/\operatorname{Ker} f \cong f(R)$ as rings.

Proof. **(a)** We have to show that the f-classes are precisely the cosets of $\operatorname{Ker} f$.

For each $a \in R$, we have

$$\begin{aligned} &(\text{the } f\text{-class that contains } a) \\ &= \{b \in R \mid f(a) = f(b)\} \\ &= \{b \in R \mid f(a) - f(b) = 0\} \\ &= \{b \in R \mid f(a-b) = 0\} \qquad (\text{since } f \text{ respects differences}) \\ &= \{b \in R \mid a - b \in \operatorname{Ker} f\} \\ &= \{b \in R \mid a + \operatorname{Ker} f = b + \operatorname{Ker} f\} \\ &= \{b \in R \mid b \text{ lies in the same coset of } \operatorname{Ker} f \text{ as } a\} \\ &= (\text{the coset of } \operatorname{Ker} f \text{ that contains } a). \end{aligned}$$

So the f-classes are precisely the cosets of $\operatorname{Ker} f$. Thus, $R/f = R/\operatorname{Ker} f$ as sets. Of course, $R/\operatorname{Ker} f$ is a ring, since $\operatorname{Ker} f$ is an ideal of R (by what we proved a while ago). Thus, part **(a)** is proved.

(b) Done before.

(c) Since we know that $R/\operatorname{Ker} f = R/f$, we see that our map

$$f' : R/\operatorname{Ker} f \to f(R), \\ \overline{r} \mapsto f(r)$$

is precisely the map

$$\begin{aligned} f' : R/f &\to f(R), \\ \overline{r} &\mapsto f(r) \end{aligned}$$

that was proved to be well-defined and bijective in the previous theorem. It remains to prove that f' is a ring isomorphism. Since f' is bijective, it suffices to show that f' is a ring morphism. This is easy (e.g., it respects multiplication since $f'\left(\overline{a} \cdot \overline{b}\right) = f'\left(\overline{ab}\right) = f(ab) = f(a) f(b) = f'(\overline{a}) f'\left(\overline{b}\right)$), but actually is also a particular case of the universal property of quotient rings (applied to $I = \operatorname{Ker} f$, which is allowed since $f(\operatorname{Ker} f) = 0$). So part **(c)** is proved.

(d) This was already done in the proof of part **(c)**.

(e) This is just part **(e)** of the previous theorem.

(f) Follows from **(c)**. □

As our proof has shown, the First Isomorphism Theorem (for rings) is merely a partial (i.e., less general) improvement on the universal property of quotient rings: The latter yields a ring morphism from R/I, while the former produces a ring isomorphism from $R/\operatorname{Ker} f$.

Here are some examples for using the First Isomorphism Theorem:

- Consider the map

$$f : \mathbb{Q}^{4\leq 4} \to \mathbb{Q}^{2\leq 2},$$
$$\begin{pmatrix} a & b & c & d \\ 0 & u & v & w \\ 0 & 0 & x & y \\ 0 & 0 & 0 & z \end{pmatrix} \mapsto \begin{pmatrix} u & v \\ 0 & x \end{pmatrix},$$

  which removes the "outer shell" from an upper-triangular matrix. This

map f is a ring morphism; for instance, it respects multiplication because

$$\begin{aligned}
&f\left(\begin{pmatrix} a & b & c & d \\ 0 & u & v & w \\ 0 & 0 & x & y \\ 0 & 0 & 0 & z \end{pmatrix}\begin{pmatrix} a' & b' & c' & d' \\ 0 & u' & v' & w' \\ 0 & 0 & x' & y' \\ 0 & 0 & 0 & z' \end{pmatrix}\right) \\
&= f\begin{pmatrix} aa' & bu'+ab' & cx'+bv'+ac' & ad'+cy'+dz'+bw' \\ 0 & uu' & vx'+uv' & vy'+wz'+uw' \\ 0 & 0 & xx' & xy'+yz' \\ 0 & 0 & 0 & zz' \end{pmatrix} \\
&= \begin{pmatrix} uu' & vx'+uv' \\ 0 & xx' \end{pmatrix} = \begin{pmatrix} u & v \\ 0 & x \end{pmatrix}\begin{pmatrix} u' & v' \\ 0 & x' \end{pmatrix} \\
&= f\begin{pmatrix} a & b & c & d \\ 0 & u & v & w \\ 0 & 0 & x & y \\ 0 & 0 & 0 & z \end{pmatrix} \cdot f\begin{pmatrix} a' & b' & c' & d' \\ 0 & u' & v' & w' \\ 0 & 0 & x' & y' \\ 0 & 0 & 0 & z' \end{pmatrix}.
\end{aligned}$$

The kernel of this morphism f is

$$\begin{aligned}
\operatorname{Ker} f &= \left\{ \begin{pmatrix} a & b & c & d \\ 0 & u & v & w \\ 0 & 0 & x & y \\ 0 & 0 & 0 & z \end{pmatrix} \in \mathbb{Q}^{4\leq 4} \;\mid\; \begin{pmatrix} u & v \\ 0 & x \end{pmatrix} = 0 \right\} \\
&= \left\{ \begin{pmatrix} a & b & c & d \\ 0 & 0 & 0 & w \\ 0 & 0 & 0 & y \\ 0 & 0 & 0 & z \end{pmatrix} \in \mathbb{Q}^{4\leq 4} \right\}.
\end{aligned}$$

So you can conclude right away that this $\operatorname{Ker} f$ is an ideal of $\mathbb{Q}^{4\leq 4}$. Moreover, the image $f\left(\mathbb{Q}^{4\leq 4}\right)$ is the whole $\mathbb{Q}^{2\leq 2}$. The First Isomorphism Theorem yields a ring isomorphism

$$\begin{aligned}
f' : \mathbb{Q}^{4\leq 4}/\operatorname{Ker} f &\to f\left(\mathbb{Q}^{4\leq 4}\right), \\
\overline{r} &\mapsto f(r).
\end{aligned}$$

In other words, it yields a ring isomorphism

$$\begin{aligned}
&f' : \mathbb{Q}^{4\leq 4}/\operatorname{Ker} f \to \mathbb{Q}^{2\leq 2}, \\
&\overline{\begin{pmatrix} a & b & c & d \\ 0 & u & v & w \\ 0 & 0 & x & y \\ 0 & 0 & 0 & z \end{pmatrix}} \mapsto \begin{pmatrix} u & v \\ 0 & x \end{pmatrix}.
\end{aligned}$$

In particular, $\mathbb{Q}^{4\leq 4}/\operatorname{Ker} f \cong \mathbb{Q}^{2\leq 2}$.

- Polynomials provide a great source of examples for the First Isomorphism Theorem. The typical example will look as follows:

$$(\text{a polynomial ring}) \,/\, (\text{an ideal}) \cong (\text{a ring of numbers}).$$

For example,

$$\underbrace{\mathbb{R}[x]}_{\substack{\text{the ring of polynomials} \\ \text{in } x \text{ with real coefficients}}} \quad / \quad \underbrace{\left(x^2+1\right)}_{\substack{\text{I really mean the} \\ \text{principal ideal } \left(x^2+1\right)\mathbb{R}[x] \\ \text{here}}} \quad \cong \mathbb{C}.$$

Informally, this is saying that if you are working with polynomials in an indeterminate x with real coefficients, but you equate the polynomial x^2+1 to zero, then you obtain the complex numbers. Even more informally, the complex numbers are obtained from the real numbers by "adjoining a root of x^2+1", that is, conjuring an element i satisfying $i^2+1=0$ out of thin air and computing with it. We will make this precise later.

1.10. Direct products of rings

1.10.1. Direct products of two rings

Here is a way to generate a new ring out of two existing rings (proof straightforward):

Proposition 1.10.1. Let R and S be two rings. Then, the Cartesian product

$$R \times S = \{(r,s) \mid r \in R \text{ and } s \in S\}$$

becomes a ring if we endow it with the entrywise addition

$$(r,s) + \left(r',s'\right) = \left(r+r',\ s+s'\right)$$

and the entrywise multiplication

$$(r,s)\left(r',s'\right) = \left(rr',\ ss'\right)$$

and the zero $(0_R, 0_S)$ and the unity $(1_R, 1_S)$.

Definition 1.10.2. This ring is denoted by $R \times S$, and is called the **direct product** of R and S.

1.10.2. Direct products of any number of rings

More generally, we can define the direct product $R_1 \times R_2 \times \cdots \times R_n$ of any n rings $R_1, R_2, \ldots, R_n$, and even better, the direct product $\prod_{i \in I} R_i$ of any family (finite or infinite) of rings R_i:

Proposition 1.10.3. Let I be a set. Let $(R_i)_{i \in I}$ be a family of rings (i.e., let R_i be a ring for each $i \in I$). Then, the Cartesian product

$$\prod_{i \in I} R_i = \left\{\text{all families } (r_i)_{i \in I} \text{ with } r_i \in R_i \text{ for each } i \in I\right\}$$

becomes a ring if we endow it with the entrywise addition

$$(r_i)_{i \in I} + (s_i)_{i \in I} = (r_i + s_i)_{i \in I}$$

and the entrywise multiplication

$$(r_i)_{i \in I} \, (s_i)_{i \in I} = (r_i s_i)_{i \in I}$$

and the zero $\left(0_{R_i}\right)_{i \in I}$ and the unity $\left(1_{R_i}\right)_{i \in I}$.

Definition 1.10.4. This ring is called the **direct product** of the rings R_i, and is denoted by $\prod_{i \in I} R_i$.

In particular:

- If $I = \{1, 2, \ldots, n\}$, then $\prod_{i \in I} R_i$ is also denoted by $R_1 \times R_2 \times \cdots \times R_n$, and its elements $(r_i)_{i \in \{1,2,\ldots,n\}}$ are written as $(r_1, r_2, \ldots, r_n)$.
- If all the rings R_i are the same ring R, then $\prod_{i \in I} R_i$ is also denoted by R^I. This is actually just the ring of all functions from I to R (with pointwise $+$ and $\cdot$), which we have seen before among our first examples of rings.
- If $n \in \mathbb{N}$, and if R is a ring, then the ring $R^{\{1,2,\ldots,n\}} = \underbrace{R \times R \times \cdots \times R}_{n \text{ times}}$ is just called R^n.
- If $I = \{1, 2, 3, \ldots\}$, then the families $(r_i)_{i \in I}$ can be written as infinite sequences $(r_1, r_2, r_3, \ldots)$.
- If $I = \mathbb{Z}$, then the families $(r_i)_{i \in I}$ can be written as "infinite-both-ways sequences" $(\ldots, r_{-2}, r_{-1}, r_0, r_1, r_2, \ldots)$.

1.10.3. Examples

- The ring $\mathbb{Z}^3 = \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ consists of all triples (r, s, t) of integers. Addition and multiplication are entrywise: e.g.,

$$(r, s, t) \cdot (r', s', t') = (rr', ss', tt').$$

  This ring is not an integral domain, since $(0, 1, 2) \cdot (1, 0, 0) = (0, 0, 0)$.

- If R, S and T are three rings, then the direct products

$$R \times S \times T, \qquad R \times (S \times T), \qquad (R \times S) \times T$$

  are not the same! They are not even the same set, since their elements have the respective forms

$$(r, s, t), \qquad (r, (s, t)), \qquad ((r, s), t);$$

  these forms clearly "carry the same information" but are organized differently.

  However, these three direct products are isomorphic. The isomorphisms go as follows:

$$\begin{aligned} R \times S \times T &\to R \times (S \times T), \\ (r, s, t) &\mapsto (r, (s, t)) \end{aligned}$$

  and

$$\begin{aligned} R \times S \times T &\to (R \times S) \times T, \\ (r, s, t) &\mapsto ((r, s), t). \end{aligned}$$

  So we say that the direct product operation (on rings) is "associative up to isomorphism". It is also "commutative up to isomorphism".

- Complex numbers are defined as pairs of real numbers. Thus, $\mathbb{C}$ is $\mathbb{R} \times \mathbb{R}$ as a set (since $a + bi = (a, b)$), even as an additive group, but not as a ring! This is because $(a, b) \cdot (c, d) \neq (ac, bd)$ in $\mathbb{C}$. Instead, $(a, b) \cdot (c, d) = (ac - bd,\ ad + bc)$.

  Actually, $\mathbb{C}$ is not isomorphic to $\mathbb{R} \times \mathbb{R}$ as rings, since $\mathbb{C}$ is a field but $\mathbb{R} \times \mathbb{R}$ is not ($\mathbb{R} \times \mathbb{R}$ is not even an integral domain, because $(1, 0) \cdot (0, 1) = (0, 0)$).

- Let R be any ring. Let $n \in \mathbb{N}$. Let $R^{n=n}$ be the set of all **diagonal** $n \times n$-matrices over R. That is,

$$R^{n=n} = \left\{ \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix} \ \mid \ a_1, a_2, \ldots, a_n \in R \right\}.$$

Then, $R^{n=n}$ is a subring of $R^{n\times n}$. Moreover, $R^{n=n} \cong R^n$ as rings (where R^n is a direct product, as defined above). Specifically, the map

$$R^n \to R^{n=n},$$
$$(a_1, a_2, \ldots, a_n) \mapsto \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix}$$

is a ring isomorphism.

Note: A direct product of commutative rings is commutative.

1.10.4. Direct products and idempotents

Direct products are closely related to idempotents. Recall that **idempotents** in a ring R are elements $r \in R$ satisfying $r^2 = r$. Two idempotents that always exist are 0 and 1, but some rings have more. For instance, $(0,1,0,1,1,1,0)$ is an idempotent in a direct product of seven rings. So a direct product of k nontrivial rings has at least 2^k idempotents. Another example of an idempotent is any idempotent matrix; these matrices are called **projections**.

I claim that idempotents can be used to reveal rings as direct products. More precisely, **central** idempotents can do this.

One direction is clear: Given two rings R and S, the pairs $a := (1_R, 0_S)$ and $b := (0_R, 1_S)$ are central idempotents in $R \times S$. These idempotents allow you to reconstruct the factors R and S back from the direct product $R \times S$: namely, the principal ideals

$$\begin{aligned} a(R \times S) &= \{ax \mid x \in R \times S\} \\ &= \{(1_R, 0_S)(r,s) \mid (r,s) \in R \times S\} \\ &= \{(r, 0_S) \mid (r,s) \in R \times S\} \\ &= \{(r, 0_S) \mid r \in R\} \cong R \end{aligned}$$

and $b(R \times S) \cong S$ are themselves rings that are isomorphic to R and S. So the central idempotents a and b allow us to decompose the direct product $R \times S$ into its factors R and S. Actually, just a suffices, since $a + b = (1_R, 1_S) = 1_{R\times S}$ and thus $b = 1_{R\times S} - a$.

Conversely:

Proposition 1.10.5. Let e be a central idempotent in a ring R. Then, the principal ideals $eR = Re$ and $(1-e)R = R(1-e)$ themselves are rings (with addition, multiplication and zero inherited from R, and with unities e and $1-e$, respectively), and there is a ring isomorphism

$$\begin{aligned} (eR) \times ((1-e)R) &\to R, \\ (x, y) &\mapsto x + y. \end{aligned}$$

In particular, $R \cong (eR) \times ((1-e)\,R)$.

Note that this would not hold for non-central idempotents. For instance, the matrix ring $\mathbb{R}^{2\times 2}$ has lots of idempotents (e.g., $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$), but is not a nontrivial direct product.

Let's see an example: The ring $\mathbb{Z}/6$ has idempotents $\overline{0}$, $\overline{1}$, $\overline{3}$, $\overline{4}$. These are the trivial ones plus two more. All of them are central, since $\mathbb{Z}/6$ is commutative. So we should obtain an isomorphism from $\mathbb{Z}/6$ to a direct product of nontrivial rings! Using the idempotent $e = \overline{3}$, we obtain

$$\overline{3}\,(\mathbb{Z}/6) = \{\overline{0}, \overline{3}\} \cong \mathbb{Z}/2;$$
$$(1-\overline{3})\,(\mathbb{Z}/6) = \overline{4}\,(\mathbb{Z}/6) = \{\overline{0}, \overline{4}, \overline{2}\} \cong \mathbb{Z}/3.$$

Thus, the above proposition yields

$$\mathbb{Z}/6 \cong (\mathbb{Z}/2) \times (\mathbb{Z}/3) \qquad \text{as rings.}$$

We will soon see how to generalize this.

1.11. Ideal arithmetic

Next, we shall define two ways to build new ideals from old:

Definition 1.11.1. Let I and J be two ideals of a ring R.

(a) Then, $I + J$ denotes the subset

$$\{i + j \mid i \in I \text{ and } j \in J\} \text{ of } R.$$

(b) Next, we define a further subset IJ of R, also denoted $I \cdot J$. Unlike $I + J$, this will **not** be defined as $\{ij \mid i \in I \text{ and } j \in J\}$, since that would not be an ideal.

Instead, $IJ = I \cdot J$ will be defined as the set

$$\{\text{all finite sums of } (I,J)\text{-products}\},$$

where an (I,J)**-product** means a product ij with $i \in I$ and $j \in J$.

Note that our definition of IJ was more complicated than our definition of $I + J$, since it involved the extra step of forming finite sums. This ensures that IJ is closed under addition. For $I + J$, this step was not necessary, since the set $\{i + j \mid i \in I \text{ and } j \in J\}$ is already closed under addition:

$$(i_1 + j_1) + (i_2 + j_2) = \underbrace{(i_1 + i_2)}_{\in I} + \underbrace{(j_1 + j_2)}_{\in J}.$$

But this would not hold for $\{ij \mid i \in I \text{ and } j \in J\}$.

A third way to combine ideals is taking their intersection: $I \cap J$.

Here is a bunch of properties of these operations:

Proposition 1.11.2 (ideal arithmetic)**.** Let R be a ring.

(a) Let I and J be two ideals of R. Then, $I + J$ and $I \cap J$ and IJ are ideals of R as well.

(b) Let I and J be two ideals of R. Then, $IJ \subseteq I \cap J \subseteq I \subseteq I + J$ and $IJ \subseteq I \cap J \subseteq J \subseteq I + J$.

(c) The set of all ideals of R is a monoid with respect to the binary operation $+$, with neutral element $\{0_R\}$. That is,

$$(I + J) + K = I + (J + K) \qquad \text{for any three ideals } I, J, K \text{ of } R;$$
$$I + \{0_R\} = \{0_R\} + I = I \qquad \text{for any ideal } I \text{ of } R.$$

(d) The set of all ideals of R is a monoid with respect to the binary operation $\cap$, with neutral element R. That is,

$$(I \cap J) \cap K = I \cap (J \cap K) \qquad \text{for any three ideals } I, J, K \text{ of } R;$$
$$I \cap R = R \cap I = I \qquad \text{for any ideal } I \text{ of } R.$$

(e) The set of all ideals of R is a monoid with respect to the binary operation $\cdot$, with neutral element R. That is,

$$(IJ) K = I (JK) \qquad \text{for any three ideals } I, J, K \text{ of } R;$$
$$IR = RI = I \qquad \text{for any ideal } I \text{ of } R.$$

(f) Addition and intersection of ideals are commutative:

$$I + J = J + I \qquad \text{and} \qquad I \cap J = J \cap I.$$

(g) If R is commutative, then $IJ = JI$ for any ideals I and J of R.

This proposition shows that the operations $+$, $\cap$ and $\cdot$ on the set of all ideals of R satisfy a lot of rules that resemble those of arithmetic. This is called **ideal arithmetic**. But don't get too relaxed: You cannot subtract ideals. In particular, you cannot reconstruct I from J and $I + J$.

To see these things on an example, let us check what these operations give for principal ideals of $\mathbb{Z}$:

Proposition 1.11.3. Let $n, m \in \mathbb{Z}$. Let $I = n\mathbb{Z}$ and $J = m\mathbb{Z}$. Then:

(a) We have $IJ = nm\mathbb{Z}$.

(b) We have $I \cap J = \operatorname{lcm}(n, m)\,\mathbb{Z}$.

(c) We have $I + J = \gcd(n, m)\,\mathbb{Z}$.

(d) We have $I \subseteq J$ if and only if $m \mid n$.

(e) We have $I = J$ if and only if $|n| = |m|$.

Proof. **(a)** Let $c \in nm\mathbb{Z}$. Then, $c = nmd$ for some integer d. Then,

$$c = nmd = \underbrace{n}_{\in n\mathbb{Z} = I} \underbrace{(md)}_{= m\mathbb{Z} \in J}$$

is an (I, J)-product, hence a finite sum of (I, J)-products. So $c \in IJ$.

This shows that $nm\mathbb{Z} \subseteq IJ$.

Conversely, any (I, J)-product is a multiple of nm (since it is ij for some $i = nk$ and $j = m\ell$, and thus equals $(nk)(m\ell) = nm(k\ell)$). Hence, any finite sum of (I, J)-products is a multiple of nm as well. In other words, it belongs to $nm\mathbb{Z}$. This shows that $IJ \subseteq nm\mathbb{Z}$. Combined with $nm\mathbb{Z} \subseteq IJ$, this yields $IJ = nm\mathbb{Z}$.

(b) We have

$$\begin{aligned} I \cap J &= \{\text{all elements of } I \text{ that also belong to } J\} \\ &= \{\text{all elements of } n\mathbb{Z} \text{ that also belong to } m\mathbb{Z}\} \\ &= \{\text{all multiples of } n \text{ that also are multiples of } m\} \\ &= \{\text{all common multiples of } n \text{ and } m\} \\ &= \{\text{all multiples of } \operatorname{lcm}(n, m)\} = \operatorname{lcm}(n, m)\,\mathbb{Z}. \end{aligned}$$

(c) We must show that $I + J = \gcd(n, m)\,\mathbb{Z}$.

Bezout's theorem tells us that there exist two integers a and b such that $na + mb = \gcd(n, m)$. Using these integers, we obtain

$$\gcd(n, m) = \underbrace{na}_{\in I} + \underbrace{mb}_{\in J} \in I + J,$$

and consequently $\gcd(n, m)\,\mathbb{Z} \subseteq I + J$ (because $I + J$ is an ideal and thus closed under taking multiples).

Remains to prove that $I + J \subseteq \gcd(n, m)\,\mathbb{Z}$. In other words, we must prove that $i + j \in \gcd(n, m)\,\mathbb{Z}$ for any $i \in I$ and $j \in J$. Let us just do this: Let $i \in I$ and $j \in J$. Thus, $i = nx$ and $j = my$ for some integers x, y. Then,

$$\begin{aligned} i + j &= \underbrace{n}_{= \gcd(n,m) \cdot z} x + \underbrace{m}_{= \gcd(n,m) \cdot w} y \\ &= \gcd(n, m) \cdot zx + \gcd(n, m) \cdot wy \\ &= \gcd(n, m) \cdot (zx + wy) \in \gcd(n, m)\,\mathbb{Z}, \end{aligned}$$

qed.

(d), **(e)** LTTR. □

We will soon see how this proposition can be generalized to a wider class of rings (but not all rings).

1.12. The Chinese Remainder Theorem

1.12.1. Introduction

Let us recall our ring isomorphism

$$\mathbb{Z}/6 \cong (\mathbb{Z}/2) \times (\mathbb{Z}/3)\,.$$

This can be generalized:

Theorem 1.12.1 (The Chinese Remainder Theorem for two integers)**.** Let n and m be two coprime integers. Then,

$$\mathbb{Z}/(nm) \cong (\mathbb{Z}/n) \times (\mathbb{Z}/m) \qquad \text{as rings.}$$

More concretely, there is a ring isomorphism

$$\begin{aligned} &\mathbb{Z}/(nm) \to (\mathbb{Z}/n) \times (\mathbb{Z}/m)\,, \\ &\qquad \overline{r} \mapsto (\overline{r},\ \overline{r})\,, \qquad \text{or, to be precise:} \\ &r + nm\mathbb{Z} \mapsto (r + n\mathbb{Z},\ r + m\mathbb{Z})\,. \end{aligned}$$

As usual, the notation $\overline{r}$ is confusing but practical, since it avoids the mention of the modulus. Context can make it unambiguous.

Rather than prove this theorem directly, I will generalize it and prove it in the general setting.

1.12.2. The Chinese Remainder Theorem for two ideals

We replace our two coprime integers n and m by two ideals I and J of an arbitrary ring R. Instead of the coprimality we now require the ideals I and J to be **comaximal** – meaning that they satisfy $I + J = R$. In fact, by the above result $n\mathbb{Z} + m\mathbb{Z} = \gcd(n, m)\,\mathbb{Z}$, the coprimality of n and m is equivalent to the comaximality of the principal ideals $n\mathbb{Z}$ and $m\mathbb{Z}$.

Definition 1.12.2. Let I and J be two ideals of a ring R. We say that I and J are **comaximal** if $I + J = R$.

Now we can generalize the Chinese Remainder Theorem above to ideals:

Theorem 1.12.3 (The Chinese Remainder Theorem for two ideals)**.** Let I and J be two comaximal ideals of a commutative ring R. Then:

(a) We have $I \cap J = IJ$.

(b) We have $R/(IJ) \cong (R/I) \times (R/J)$.

(c) More concretely, there is a ring isomorphism

$$\begin{aligned} &R/(IJ) \to (R/I) \times (R/J)\,, \\ &\qquad \overline{r} \mapsto (\overline{r}, \overline{r}) \qquad (\text{that is, } r + IJ \mapsto (r + I,\ r + J))\,. \end{aligned}$$

Part **(a)** generalizes the fact that $\operatorname{lcm}(n,m) = \pm nm$ for any two coprime integers n and m; parts **(b)** and **(c)** generalize the above CRT for integers.

Let us now prove the generalized theorem. We will abbreviate products such as $(R/I) \times (R/J)$ by removing the parentheses: $R/I \times R/J$. Division goes before multiplication.

Proof. We know that I and J are comaximal, i.e., we have $I + J = R$. Hence, $1 \in R = I + J$. In other words, $1 = i + j$ for some $i \in I$ and $j \in J$. Consider these i and j.

(a) It is easy to see that $IJ \subseteq I \cap J$ (since any (I,J)-product lies both in I and in J, hence lies in $I \cap J$, and of course the same must hold for the finite sums of (I,J)-products). It remains to show that $I \cap J \subseteq IJ$.

Let $a \in I \cap J$. Then,

$$\begin{aligned} a &= \underbrace{1}_{=i+j} \cdot a = (i+j) \cdot a = \underbrace{i}_{\in I} \underbrace{a}_{\in I\cap J\subseteq J} + \underbrace{j}_{\in J} \underbrace{a}_{\in I\cap J\subseteq I} \\ &\in IJ + JI = IJ \qquad (\text{since } R \text{ is commutative}) . \end{aligned}$$

So we have shown that $a \in IJ$ for each $a \in I \cap J$. In other words, $I \cap J \subseteq IJ$. So part **(a)** is proved.

(c) Define the map

$$\begin{aligned} f : R &\to (R/I) \times (R/J) , \\ r &\mapsto (\overline{r}, \overline{r}) \end{aligned}$$

(where $(\overline{r}, \overline{r})$ means $(r + I,\ r + J)$). Easily, f is a ring morphism. Its kernel is

$$\begin{aligned} \operatorname{Ker} f &= \left\{ r \in R \mid (\overline{r}, \overline{r}) = 0_{(R/I)\times(R/J)} \right\} \\ &= \{ r \in R \mid \overline{r} = 0 \text{ in } R/I, \text{ and } \overline{r} = 0 \text{ in } R/J \} \\ &= \{ r \in R \mid r \in I, \text{ and } r \in J \} \\ &= \{ r \in R \mid r \in I \cap J \} \\ &= I \cap J = IJ \qquad (\text{by part } \textbf{(a)}) . \end{aligned}$$

Now I claim that f is surjective, i.e., that the image of f is the whole ring $(R/I) \times (R/J)$.

To prove this, recall again that $1 = i + j$, so that $1 - i = j \in J$. Hence, $\overline{1} = \overline{i}$ in R/J. In other words, $\overline{i} = \overline{1}$ in R/J. Similarly, $\overline{j} = \overline{1}$ in R/I. Therefore,

$$\begin{aligned} f(i) &= \left(\overline{i}, \overline{i}\right) = \left(0_{R/I}, 1_{R/J}\right) \qquad \text{and} \\ f(j) &= \left(\overline{j}, \overline{j}\right) = \left(1_{R/I}, 0_{R/J}\right) . \end{aligned}$$

Now, for every $x \in R$ and $y \in R$, we have

$$\begin{aligned} f(yi + xj) &= f(y) f(i) + f(x) f(j) \qquad (\text{since } f \text{ is a ring morphism}) \\ &= (\overline{y}, \overline{y}) \left(0_{R/I}, 1_{R/J}\right) + (\overline{x}, \overline{x}) \left(1_{R/I}, 0_{R/J}\right) \\ &= \left(0_{R/I}, \overline{y}\right) + \left(\overline{x}, 0_{R/J}\right) \\ &= (\overline{x}, \overline{y}) . \end{aligned}$$

This shows that every $(\overline{x}, \overline{y}) \in R/I \times R/J$ lies in the image of f. In other words, f is surjective. That is,

$$f(R) = R/I \times R/J.$$

The first isomorphism theorem for rings says that the map

$$\begin{aligned} f' : R/\operatorname{Ker} f &\to f(R), \\ \overline{r} &\mapsto f(r) \end{aligned}$$

is well-defined and is a ring isomorphism. In view of $\operatorname{Ker} f = IJ$ and $f(R) = R/I \times R/J$ and $f(r) = (\overline{r}, \overline{r})$, we can rewrite this as follows: The map

$$\begin{aligned} f' : R/IJ &\to R/I \times R/J, \\ \overline{r} &\mapsto (\overline{r}, \overline{r}) \end{aligned}$$

is well-defined and is a ring isomorphism. This proves part **(c)**, and thus also part **(b)**. □

1.12.3. Application to integers

Proof of the CRT for two integers. Let $R = \mathbb{Z}$ and $I = n\mathbb{Z}$ and $J = m\mathbb{Z}$. Then, something we proved a while ago says that $IJ = nm\mathbb{Z}$ and $I + J = \underbrace{\gcd(n,m)}_{\substack{=1 \\ \text{(since } n \text{ and } m \text{ are coprime)}}} \mathbb{Z} = \mathbb{Z}$. So the ideals I and J are comaximal. Hence, the general CRT for two ideals yields that

$$\begin{aligned} R/IJ &\cong R/I \times R/J, \qquad \text{that is,} \\ \mathbb{Z}/nm &\cong \mathbb{Z}/n \times \mathbb{Z}/m. \end{aligned}$$

The specific isomorphism also follows from the general CRT. □

If n and m are two integers, then the **extended Euclidean algorithm** provides a quick way of computing integers x and y such that $xn + ym = \gcd(n,m)$. Thus, the ideal equality $n\mathbb{Z} + m\mathbb{Z} = \gcd(n,m)\,\mathbb{Z}$ can be realized by a pretty efficient algorithm. This makes the CRT itself rather efficient, giving a quick way of evaluating not just the isomorphism $\mathbb{Z}/nm \to \mathbb{Z}/n \times \mathbb{Z}/m$ (this one is easy to compute) but also its inverse.

1.12.4. Interlude: Multiplying comaximal ideals

Our next goal is to extend the CRT from two ideals to k ideals for any k. To do so, we need some auxiliary results about how comaximality behaves under ideal multiplication.

Recall the classical fact from number theory saying that if $\gcd(i,k) = 1$ and $\gcd(j,k) = 1$, then $\gcd(ij,k) = 1$ (for integers i,j,k). More generally, for any three integers i,j,k, we have

$$\gcd(ij,k) \mid \gcd(i,k)\gcd(j,k).$$

We can generalize these facts to any ideals in any ring:

Proposition 1.12.4. Let I, J and K be three ideals of a ring R. Then:
(a) We have $(I+K)(J+K) \subseteq IJ+K$.
(b) If $I+K=R$ and $J+K=R$, then $IJ+K=R$.

Proof. All three sets $I+K$ and $J+K$ and $IJ+K$ are ideals of R.

(a) It suffices to prove that any $(I+K,\ J+K)$-product lies in $IJ+K$ (since the ideal $(I+K)(J+K)$ consists of finite sums of such products, but $IJ+K$ is closed under addition).

In other words, we must prove that $xy \in IJ+K$ for any $x \in I+K$ and $y \in J+K$. Let's prove this.

Fix $x \in I+K$ and $y \in J+K$. Since $x \in I+K$, we can write x as $x = i+a$ with $i \in I$ and $a \in K$. Likewise, we can write y as $y = j+b$ with $j \in J$ and $b \in K$. Using these i,a,j,b, we now have

$$\begin{aligned} xy &= (i+a)(j+b) = \underbrace{ij}_{\in IJ} + \underbrace{ib}_{\in K} + \underbrace{aj}_{\in K} + \underbrace{ab}_{\in K} \\ &\in IJ + \underbrace{K+K+K}_{\subseteq K} \subseteq IJ+K, \end{aligned}$$

(since K is an ideal)

as desired. So part **(a)** of the proposition is proved.

(b) Assume that $I+K=R$ and $J+K=R$. But part **(a)** says that $(I+K)(J+K) \subseteq IJ+K$. Hence, $IJ+K \supseteq \underbrace{(I+K)}_{=R}\underbrace{(J+K)}_{=R} = \underbrace{R}_{\ni 1} R = R$, thus $IJ+K=R$. □

We can extend part **(b)** of this proposition to products of multiple ideals:

Proposition 1.12.5. Let $I_1, I_2, \ldots, I_k$ be k ideals of a ring R. Let K be a further ideal of R. Assume that

$$I_i + K = R \qquad \text{for each } i \in \{1,2,\ldots,k\}.$$

Then, $I_1 I_2 \cdots I_k + K = R$.

Proof. Induction on k. The base case ($k=0$) relies on the empty product of 0 ideals being R (this is a definition). For the induction step, use part **(b)** of the previous proposition. □

1.12.5. The Chinese Remainder Theorem for k ideals

To generalize the CRT to k rather than 2 ideals, we need one more piece of notation:

Definition 1.12.6. Let $I_1, I_2, \ldots, I_k$ be k ideals of a ring R. We say that these k ideals $I_1, I_2, \ldots, I_k$ are **mutually comaximal** if $I_i + I_j = R$ holds for all $i < j$ (that is, I_i is comaximal with I_j for all $i < j$).

For $k > 2$, this is a **much stronger** requirement than $I_1 + I_2 + \cdots + I_k = R$. This is a generalization of the mutual coprimality of k integers, which is much stronger than just saying that their cumulative gcd is 1. For instance, the three integers $6, 10, 15$ have $\gcd(6, 10, 15) = 1$, but no two of them are coprime; they are certainly not mutually coprime.

When $n_1, n_2, \ldots, n_k$ are k mutually coprime integers, the corresponding principal ideals $n_1\mathbb{Z}$, $n_2\mathbb{Z}$, $\ldots$, $n_k\mathbb{Z}$ are mutually coprime.

Now we can state the generalized CRT for k ideals:

Theorem 1.12.7 (The Chinese Remainder Theorem for k ideals)**.** Let $I_1, I_2, \ldots, I_k$ be k mutually comaximal ideals of a commutative ring R. Then:

(a) We have

$$I_1 \cap I_2 \cap \cdots \cap I_k = I_1 I_2 \cdots I_k.$$

(b) We have

$$R/(I_1 I_2 \cdots I_k) \cong R/I_1 \times R/I_2 \times \cdots \times R/I_k.$$

(c) More concretely, there is a ring isomorphism

$$\begin{aligned} R/(I_1 I_2 \cdots I_k) &\to R/I_1 \times R/I_2 \times \cdots \times R/I_k, \\ \overline{r} &\mapsto (\overline{r}, \overline{r}, \ldots, \overline{r}). \end{aligned}$$

Proof. Induct on k. The induction step uses

- the previous proposition to argue that $I_1 I_2 \cdots I_{k-1}$ is comaximal with I_k;
- the CRT for two ideals;
- the simple fact that if three rings A, B, C satisfy $A \cong B$, then $A \times C \cong B \times C$ (and more concretely: if $f : A \to B$ is a ring isomorphism, then

$$\begin{aligned} A \times C &\to B \times C, \\ (a, c) &\mapsto (f(a), c) \end{aligned}$$

  is a ring isomorphism as well).

See the text for more details. □

1.12.6. Applying to integers again

Applying the general CRT for k ideals to k principal ideals of $R = \mathbb{Z}$, we find:

Theorem 1.12.8 (The Chinese Remainder Theorem for k integers)**.** Let $n_1, n_2, \ldots, n_k$ be k mutually coprime integers (that is, k integers satisfying $\gcd\left(n_i, n_j\right) = 1$ for all $i < j$). Then,

$$\mathbb{Z}/\left(n_1 n_2 \cdots n_k\right) \cong \mathbb{Z}/n_1 \times \mathbb{Z}/n_2 \times \cdots \times \mathbb{Z}/n_k.$$

More concretely, there is a ring isomorphism

$$\begin{aligned} \mathbb{Z}/\left(n_1 n_2 \cdots n_k\right) &\to \mathbb{Z}/n_1 \times \mathbb{Z}/n_2 \times \cdots \times \mathbb{Z}/n_k, \\ \overline{r} &\mapsto \left(\overline{r}, \overline{r}, \ldots, \overline{r}\right). \end{aligned}$$

Proof. Apply the previous theorem to $R = \mathbb{Z}$ and $I_i = n_i\mathbb{Z}$. Comaximality follows from coprimality. □

Corollary 1.12.9. Let n be a positive integer with prime factorization $n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$ (where $a_1, a_2, \ldots, a_k \in \mathbb{N}$), where the $p_1, p_2, \ldots, p_k$ are distinct primes. Then,

$$\mathbb{Z}/n \cong \mathbb{Z}/p_1^{a_1} \times \mathbb{Z}/p_2^{a_2} \times \cdots \times \mathbb{Z}/p_k^{a_k}.$$

More concretely, there is a ring isomorphism

$$\begin{aligned} \mathbb{Z}/n &\to \mathbb{Z}/p_1^{a_1} \times \mathbb{Z}/p_2^{a_2} \times \cdots \times \mathbb{Z}/p_k^{a_k}, \\ \overline{r} &\mapsto \left(\overline{r}, \overline{r}, \ldots, \overline{r}\right). \end{aligned}$$

Proof. Apply the theorem above to $n_i = p_i^{a_i}$, after observing that powers of distinct primes are coprime. □

This corollary can be used to break rings of the form $\mathbb{Z}/n$ down into simpler rings of the form $\mathbb{Z}/p^a$. This has many applications:

- Counting squares (or, more generally, solutions to polynomial equations) in $\mathbb{Z}/n$.

  Recall HW#0 Exercise 7: How many remainders do perfect squares leave when divided by 7 ? by 14 ? by a general $n > 0$?

  In other words, how many squares are there in $\mathbb{Z}/n$? Here, a **square** in a ring R means an element of the form a^2 with $a \in R$. For instance, $\mathbb{Z}/5$ has 3 squares: $\overline{0}$, $\overline{1}$, $\overline{4}$.

  It is easy to see that if A and B are two rings, then

  $$(\#\text{ of squares in } A \times B) = (\#\text{ of squares in } A) \cdot (\#\text{ of squares in } B)$$

(since the squares in $A \times B$ are the pairs (a, b) of a square in A with a square in B). Moreover, isomorphic rings have the same # of squares. Thus, if n has the prime factorization $n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, then

$$\mathbb{Z}/n \cong \mathbb{Z}/p_1^{a_1} \times \mathbb{Z}/p_2^{a_2} \times \cdots \times \mathbb{Z}/p_k^{a_k}$$

yields

$$\begin{aligned}
&(\#\text{ of squares in } \mathbb{Z}/n) \\
&= \left(\#\text{ of squares in } \mathbb{Z}/p_1^{a_1} \times \mathbb{Z}/p_2^{a_2} \times \cdots \times \mathbb{Z}/p_k^{a_k}\right) \\
&= \prod_{i=1}^{k} \left(\#\text{ of squares in } \mathbb{Z}/p_i^{a_i}\right).
\end{aligned}$$

Hence, it remains to compute the # of squares in $\mathbb{Z}/p^a$ for any prime p and any positive integer a.

A good first step is to count the squares in $\mathbb{Z}/p$ for any prime p. Their number turns out to be

$$\begin{cases} 2, & \text{if } p = 2; \\ \dfrac{p+1}{2}, & \text{if } p \neq 2. \end{cases}$$

This is not hard to show by noticing that each square x in $\mathbb{Z}/p$ other than $\overline{0}$ is taken twice (i.e., it is the square of two distinct elements of $\mathbb{Z}/p$), unless $p = 2$. Details on the course website (Spring 2019 HW#6 Exercise 5).

The next step is counting squares in $\mathbb{Z}/p^2$. Their number is

$$\begin{cases} 2, & \text{if } p = 2; \\ \dfrac{p^2-p}{2} + 1, & \text{if } p \neq 2, \end{cases}$$

as can again be shown without too much trouble. (Again, see the website.)

Now you can trust me that the analogous problem for $\mathbb{Z}/p^a$ is solvable, but the answer is not particularly nice. It is

$$\begin{aligned}
&(\#\text{ of squares in } \mathbb{Z}/p^a) \\
&= \begin{cases} \dfrac{p^{a+1}+p+2}{2(p+1)}, & \text{if } p \neq 2 \text{ and if } a \text{ is even;} \\ \dfrac{p^{a+1}+2p+1}{2(p+1)}, & \text{if } p \neq 2 \text{ and if } a \text{ is odd;} \\ \dfrac{2^{a-1}+4}{3}, & \text{if } p = 2 \text{ and if } a \text{ is even;} \\ \dfrac{2^{a-1}+5}{3}, & \text{if } p = 2 \text{ and if } a \text{ is odd.} \end{cases}
\end{aligned}$$

Plugging this into our product formula, we find a formula for the # of squares in $\mathbb{Z}/n$.

- What is an integer a that leaves the remainder 3 when divided by 5, the remainder 2 when divided by 6, and the remainder 9 when divided by 23 ?

  This is just asking for an integer a that satisfies $\overline{a} = \overline{3}$ in $\mathbb{Z}/5$, satisfies $\overline{a} = \overline{2}$ in $\mathbb{Z}/6$, and satisfies $\overline{a} = \overline{9}$ in $\mathbb{Z}/23$. In other words, this is asking for an integer a whose image under the ring morphism

  $$\begin{aligned} \mathbb{Z} &\to \mathbb{Z}/5 \times \mathbb{Z}/6 \times \mathbb{Z}/23, \\ r &\mapsto (\overline{r}, \overline{r}, \overline{r}) \end{aligned}$$

  is the triple $\left(\overline{3}, \overline{2}, \overline{9}\right)$.

  Since the integers $5, 6, 23$ are mutually coprime, the Chinese Remainder Theorem shows that there is a ring isomorphism

  $$\begin{aligned} \mathbb{Z}/\left(5 \cdot 6 \cdot 23\right) &\to \mathbb{Z}/5 \times \mathbb{Z}/6 \times \mathbb{Z}/23, \\ \overline{r} &\mapsto \left(\overline{r}, \overline{r}, \overline{r}\right). \end{aligned}$$

  By tracking our way through the proof of this theorem, we can obtain an algorithm for constructing the inverse of this isomorphism. Thus, we can find the preimage of the triple $\left(\overline{3}, \overline{2}, \overline{9}\right)$ under our isomorphism: It is $\overline{308}$. So the integer 308 is the simplest answer to our question.

- A modern application of the Chinese Remainder Theorem is a computational technique called **Chinese remaindering**. It can be used to parallelize computations with integers. For instance, assume that you want to compute

  $$a = 77^2 \cdot 80^2 - 78^2 \cdot 79^2.$$

  By some kind of theoretical knowledge, you know that $|a| < 50\,000$. What is a quick way to find a without actually doing all the computations?

  It suffices to compute the remainder $a\%100\,001$, because this remainder, when taken together with the condition $|a| < 50\,000$, will allow you to reconstruct a. In other words, it suffices to compute $\overline{a}$ in $\mathbb{Z}/100\,001$. For the same reason, it suffices to compute $\overline{a}$ in $\mathbb{Z}/n$ for any $n > 100\,000$.

  So now you can work in $\mathbb{Z}/n$ instead of working in $\mathbb{Z}$. This is already a simplification, but we can do even better: If n has many distinct prime factors, say $n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, then $\mathbb{Z}/n \cong \mathbb{Z}/p_1^{a_1} \times \mathbb{Z}/p_2^{a_2} \times \cdots \times \mathbb{Z}/p_k^{a_k}$, and so it will suffice to compute $\overline{a}$ in each of the quotient rings $\mathbb{Z}/p_i^{a_i}$ and then use the CRT to collate the results together to get $\overline{a}$ in $\mathbb{Z}/n$ (this is possible since the extended Euclidean algorithm is quite fast).

  A good choice for n is $n = 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 = 510\,510 > 100\,000$. So, if you compute $\overline{a}$ in $\mathbb{Z}/p$ for all $p \in \{2, 3, 5, 7, 11, 13, 17\}$, then you can

reconstruct $\overline{a}$ in $\mathbb{Z}/n$ and thus a. This is easy. For instance, in $\mathbb{Z}/5$, we have

$$\begin{aligned}\overline{a} &= \overline{77^2 \cdot 80^2 - 78^2 \cdot 79^2} \\ &= \overline{77}^2 \cdot \overline{80}^2 - \overline{78}^2 \cdot \overline{79}^2 \\ &= \underbrace{\overline{2}^2 \cdot \overline{0}^2}_{=0} - \underbrace{\overline{3}^2}_{=\overline{4}} \cdot \underbrace{\overline{4}^2}_{=\overline{1}} = -\overline{4} \cdot \overline{1} = \overline{1}.\end{aligned}$$

There are many more situations where this is useful (see, e.g., the Vogan and Knuth references in the notes).

1.12.7. A few words about noncommutative rings

Recall the CRT as we stated it:

Theorem 1.12.10 (The Chinese Remainder Theorem for k ideals)**.** Let $I_1, I_2, \ldots, I_k$ be k mutually comaximal ideals of a commutative ring R. Then:

(a) We have

$$I_1 \cap I_2 \cap \cdots \cap I_k = I_1 I_2 \cdots I_k.$$

(b) We have

$$R/\left(I_1 I_2 \cdots I_k\right) \cong R/I_1 \times R/I_2 \times \cdots \times R/I_k.$$

(c) More concretely, there is a ring isomorphism

$$\begin{aligned}R/\left(I_1 I_2 \cdots I_k\right) &\to R/I_1 \times R/I_2 \times \cdots \times R/I_k, \\ \overline{r} &\mapsto \left(\overline{r}, \overline{r}, \ldots, \overline{r}\right).\end{aligned}$$

We outlined its proof and applied it to $R = \mathbb{Z}$.

But in fact, we can try to generalize it further: Does R really need to be commutative?

Literally, the answer is "yes": Our theorem that $I + J = R$ entails $I \cap J = IJ$ requires $JI = IJ$. If we do not assume that R is commutative, we have to replace it by

$$I \cap J = IJ + JI.$$

Moreover, our proof of

$$R/IJ \cong R/I \times R/J$$

can be generalized to noncommutative R if we replace IJ by $IJ + JI$: we get

$$R/\left(IJ + JI\right) \cong R/I \times R/J,$$

or simply

$$R/\left(I \cap J\right) \cong R/I \times R/J.$$

We end up with the following:

Theorem 1.12.11 (The Chinese Remainder Theorem for k ideals, noncommutative version)**.** Let $I_1, I_2, \ldots, I_k$ be k mutually comaximal ideals of a ring R. Then:

(a) We have

$$I_1 \cap I_2 \cap \cdots \cap I_k = \sum_{\sigma \in S_k} I_{\sigma(1)} I_{\sigma(2)} \cdots I_{\sigma(k)},$$

where S_k is the group of all permutations of the set $\{1, 2, \ldots, k\}$. For instance, for $k = 3$, this is saying that

$$I_1 \cap I_2 \cap I_3 = I_1 I_2 I_3 + I_1 I_3 I_2 + I_2 I_1 I_3 + I_2 I_3 I_1 + I_3 I_1 I_2 + I_3 I_2 I_1.$$

(b) We have

$$R/\left(I_1 \cap I_2 \cap \cdots \cap I_k\right) \cong R/I_1 \times R/I_2 \times \cdots \times R/I_k.$$

(c) More concretely, there is a ring isomorphism

$$\begin{aligned} R/\left(I_1 \cap I_2 \cap \cdots \cap I_k\right) &\to R/I_1 \times R/I_2 \times \cdots \times R/I_k, \\ \overline{r} &\mapsto \left(\overline{r}, \overline{r}, \ldots, \overline{r}\right). \end{aligned}$$

Proof. More or less the same as for commutative R. □

It turns out that the above theorem can still be improved! This was noticed only recently and proved by Birgit van Dalen in 2005:

Theorem 1.12.12. Let $I_1, I_2, \ldots, I_k$ be k mutually comaximal ideals of a ring R. Then,

$$\begin{aligned} I_1 \cap I_2 \cap \cdots \cap I_k &= \sum_{\sigma \in S_k} I_{\sigma(1)} I_{\sigma(2)} \cdots I_{\sigma(k)} \\ &= I_1 I_2 \cdots I_k + I_k I_{k-1} \cdots I_1 \\ &= I_1 I_2 \cdots I_k + I_2 I_3 \cdots I_k I_1 + I_3 I_4 \cdots I_k I_1 I_2 + \cdots \\ &\qquad \left(\text{sum of all cyclic rotations of } I_1 I_2 \cdots I_k\right). \end{aligned}$$

See the notes (and the HW) for more about comaximal ideals.

1.13. Euclidean rings and Euclidean domains

1.13.1. All ideals of $\mathbb{Z}$ are principal

We have been talking about ideals of $\mathbb{Z}$ for quite a while now, but we always dealt with principal ideals. Are there any others? No, because:

Proposition 1.13.1. Any ideal of $\mathbb{Z}$ is principal.

Proof. Let I be an ideal of $\mathbb{Z}$. We must show that I is principal.

If $I = \{0\}$, then this is clear (since $I = 0\mathbb{Z}$). So we assume that $I \neq \{0\}$. Then, I contains a nonzero integer, therefore a positive integer (since I is closed under negation). Let b be the **smallest** positive integer in I.

Now I want to prove that $I = b\mathbb{Z}$ (which will yield that I is principal).

Since $b \in I$, we have $b\mathbb{Z} \subseteq I$ (since I is an ideal). Remains to prove that $I \subseteq b\mathbb{Z}$.

To do so, we let $a \in I$ be arbitrary. We must show that $a \in b\mathbb{Z}$, that is, $b \mid a$, that is, $a\%b = 0$. But $a\%b = a - qb$ for some $q \in \mathbb{Z}$, and therefore $a\%b \in I$ (since $a, b \in I$). Moreover, $a\%b$ is nonnegative and smaller than b. So $a\%b$ cannot be positive (since if it was, then it would be a smaller positive integer in I than b, but b was already the smallest positive integer in I). Thus, $a\%b$ must be 0. Hence, $b \mid a$, so that $a \in b\mathbb{Z}$. This shows that $I \subseteq b\mathbb{Z}$, and we are done. □

Note that division with remainder is what made this proof work! Thus, commutative rings in which you can "divide with remainder" have the property that all their ideals are principal, at least if there is a reasonable notion of "smallest element" in them. Next time, we will make this vague concept formal and study it.

The proposition above is not constructive, since ideals of $\mathbb{Z}$ can be specified in arbitrarily non-explicit forms (e.g., the set of all integers that are 0 or multiples of an odd perfect number is an ideal of $\mathbb{Z}$, thus principal by the proposition above, but this tells you nothing about what b generates this ideal). However, when an ideal I is given in certain simple ways, there are often algorithms for finding a generator. The most common case is when I is a sum of two principal ideals: $I = a\mathbb{Z} + b\mathbb{Z}$. In this case, we are looking for a $c \in \mathbb{Z}$ such that $a\mathbb{Z} + b\mathbb{Z} = c\mathbb{Z}$. By what we proved before, we know that this c must be $\pm \gcd(a, b)$, so the question is how to compute $\gcd(a, b)$. The answer is given by the Euclidean algorithm, which (just like the above proof) rests on division with remainder. There is furthermore an algorithm called the extended Euclidean algorithm, which actually constructs two integers x and y such that $xa + yb = \gcd(a, b)$.

These algorithms are useful in number theory, so it would be good to generalize them to other rings. The rings for which this can be done are known as the **Euclidean rings**.

1.13.2. Euclidean rings

Definition 1.13.2. Let R be a commutative ring.

(a) A **norm** on R means a function $N : R \to \mathbb{N}$ with $N(0) = 0$.

(b) A norm N on R is called **Euclidean** if for any $a \in R$ and any nonzero $b \in R$, there exist $q, r \in R$ with

$$a = qb + r \qquad \text{and} \qquad (r = 0 \text{ or } N(r) < N(b)).$$

(c) We say that R is a **Euclidean ring** if R has a Euclidean norm.

(d) We say that R is a **Euclidean domain** if R is a Euclidean ring and an integral domain.

You can think of the norm as a measure of "how big" an element of R is; particular cases are the absolute value of an integer and the degree of a polynomial. Note that we are **not** requiring the norm to be multiplicative or to satisfy the triangle inequality. We are also not requiring the q and the r to be unique.

Some examples:

- Any field F is a Euclidean domain. Indeed, any map $N : F \to \mathbb{N}$ with $N(0) = 0$ is a Euclidean norm on F, since you can always divide **without** remainder (i.e., with remainder $r = 0$).
- The ring $\mathbb{Z}$ is a Euclidean domain. Indeed, the map

$$\begin{aligned} N : \mathbb{Z} &\to \mathbb{N}, \\ n &\mapsto |n| \end{aligned}$$

  is a Euclidean norm on $\mathbb{Z}$. The Euclideanness follows from division with remainder (once you check that it works for negative b). Note that the q and the r are not unique: For example, for $a = 7$ and $b = 5$, there are **two** pairs $(q, r) \in \mathbb{Z} \times \mathbb{Z}$ satisfying

$$a = qb + r \qquad \text{and} \qquad (r = 0 \text{ or } N(r) < N(b)).$$

  These two pairs are $(1, 2)$ and $(2, -3)$.
- If F is a field, then the ring $F[x]$ of polynomials in a single indeterminate x with coefficients in F is a Euclidean domain, with Euclidean norm

$$\begin{aligned} N : F[x] &\to \mathbb{N}, \\ p &\mapsto \deg p \qquad \text{when } p \neq 0, \\ 0 &\mapsto 0. \end{aligned}$$

  We will see this in more detail later on. (This is just polynomial division with remainder. Note that here, the q and the r are unique.)

  However, polynomials in more than 1 variable do not form a Euclidean domain. Neither do polynomials in 1 variable over $\mathbb{Z}$.

- The ring $\mathbb{Z}[i]$ of Gaussian integers is a Euclidean domain.

  Indeed, we claim that the map

  $$\begin{aligned} N : \mathbb{Z}[i] &\to \mathbb{N}, \\ a+bi &\mapsto a^2+b^2 \qquad (\text{for } a,b \in \mathbb{Z}) \end{aligned}$$

  is a Euclidean norm.

  To prove this, we must show that for any $\alpha \in \mathbb{Z}[i]$ and any nonzero $\beta \in \mathbb{Z}[i]$, there exist elements $q, r \in \mathbb{Z}[i]$ such that

  $$\alpha = q\beta + r \qquad \text{and} \qquad (r = 0 \text{ or } N(r) < N(\beta)).$$

  So let us fix α and β. How do we find q and r ?

  Actually, we don't need the $r = 0$ option; I claim that we can find $q, r \in \mathbb{Z}[i]$ such that

  $$\alpha = q\beta + r \qquad \text{and} \qquad N(r) < N(\beta).$$

  To find such q and r, we observe that each $z \in \mathbb{Z}[i]$ satisfies $N(z) = |z|^2$. Hence, we have the following chain of equivalences:

  $$(N(r) < N(\beta)) \iff (|r| < |\beta|) \iff \left(\left|\frac{r}{\beta}\right| < 1\right)$$

  (since $\dfrac{|z|}{|w|} = \left|\dfrac{z}{w}\right|$ for any complex z, w with $w \neq 0$). Moreover, we have the equivalence

  $$(\alpha = q\beta + r) \iff \left(\frac{\alpha}{\beta} = q + \frac{r}{\beta}\right) \iff \left(\frac{\alpha}{\beta} - q = \frac{r}{\beta}\right).$$

  Thus, we need to find $q, r \in \mathbb{Z}[i]$ such that

  $$\frac{\alpha}{\beta} - q = \frac{r}{\beta} \qquad \text{and} \qquad \left|\frac{r}{\beta}\right| < 1.$$

  In other words, we need to find $q \in \mathbb{Z}[i]$ such that

  $$\left|\frac{\alpha}{\beta} - q\right| < 1.$$

  (If we can find such a q, then $r = \alpha - q\beta$ will automatically be the r to match it.) In other words, we want to find a Gaussian integer q such that $\dfrac{\alpha}{\beta}$ lies in the open circle with center q and radius 1.

But this can be done: The open circles with centers at all Gaussian integers and radius 1 cover the whole complex plane. Geometrically, this can be seen using the convexity of the circles and the fact that every complex number lies in one of the "lattice squares" with corners $a+bi$, $(a+1)+bi$, $a+(b+1)\,i$, $(a+1)+(b+1)\,i$ for some $a, b \in \mathbb{Z}$.

There is also an algebraic argument: Write the complex number $\dfrac{\alpha}{\beta}$ as $(u+x)+(v+y)\,i$, where u and v are integers and $x, y \in [0,1]$ (since each real number can be written as $u+x$ for some $u \in \mathbb{Z}$ and $x \in [0,1]$). Then, the required point q will be

$$\begin{cases} u+vi, & \text{if } x \leq \dfrac{1}{2} \text{ and } y \leq \dfrac{1}{2}; \\ u+(v+1)\,i, & \text{if } x \leq \dfrac{1}{2} \text{ and } y > \dfrac{1}{2}; \\ (u+1)+vi, & \text{if } x > \dfrac{1}{2} \text{ and } y \leq \dfrac{1}{2}; \\ (u+1)+(v+1)\,i, & \text{if } x > \dfrac{1}{2} \text{ and } y > \dfrac{1}{2}. \end{cases}$$

To prove that $\left|\dfrac{\alpha}{\beta} - q\right| < 1$, you can just argue using Pythagoras.

Hence, $\mathbb{Z}[i]$ is a Euclidean ring, thus a Euclidean domain.

- The ring
$$\mathbb{Z}\left[\sqrt{-3}\right] = \left\{a+b\sqrt{-3} \mid a, b \in \mathbb{Z}\right\}$$
is **not** Euclidean. This is not obvious. (It is not hard to see that the "obvious" norm $(a,b) \mapsto a^2+3b^2$ is not Euclidean. But it can be shown that no other norm is Euclidean either.)

- The ring
$$\mathbb{Z}\left[\sqrt{2}\right] = \left\{a+b\sqrt{2} \mid a, b \in \mathbb{Z}\right\}$$
(a subring of $\mathbb{R}$) is Euclidean. A Euclidean norm for it is the map
$$\begin{aligned} \mathbb{Z}\left[\sqrt{2}\right] &\to \mathbb{N}, \\ a+b\sqrt{2} &\mapsto \left|a^2-2b^2\right|. \end{aligned}$$
(This is not obvious, but can be proved.)

- The ring
$$\mathbb{Z}\left[\sqrt{14}\right] = \left\{a+b\sqrt{14} \mid a, b \in \mathbb{Z}\right\}$$
is Euclidean, but the obvious norm $a+b\sqrt{14} \mapsto \left|a^2-14b^2\right|$ is not Euclidean. An actually Euclidean norm for this ring is notoriously hard to construct, but it exists.

- The ring $\mathbb{Z}\left[\sqrt{5}\right] = \left\{a + b\sqrt{5} \mid a, b \in \mathbb{Z}\right\}$ is not Euclidean.
- For any $n \in \mathbb{Z}$, the quotient ring $\mathbb{Z}/n$ is Euclidean. More generally, if R is a Euclidean ring, then any quotient ring R/I is also Euclidean.

We can now generalize the last proposition:

Proposition 1.13.3. Let R be a Euclidean ring. Then, any ideal of R is principal.

Proof. Same argument as for $\mathbb{Z}$. The only change is that now you take a nonzero element $b \in I$ with minimum $N(b)$ (instead of taking the smallest positive $b \in I$). □

Again, this proof is not constructive, but there are algorithms for many given types of ideals. In particular, for any Euclidean ring R with a division-with-remainder algorithm (i.e., an algorithm that computes q and r for given a and b), there is an algorithm that, for any $a, b \in R$, computes an element $c \in R$ satisfying $aR + bR = cR$, and also computes two elements $x, y \in R$ such that $c = xa + yb$. This is a generalization of the extended Euclidean algorithm for integers, and proceeds in exactly the same way. See §2.13.3 in the text for the details of this algorithm. The element c is essentially a "gcd" of a and b (we will say more about this later).

1.14. An introduction to divisibility theory

We will now generalize the basics of elementary number theory to commutative rings – as generally as possible. Some things generalize easily; others less so; some don't generalize at all. Often, properties hold only if the ring satisfies certain extra conditions. More can be found in textbooks.

1.14.1. Principal ideal domains

The last proposition we proved says the following:

Proposition 1.14.1. In a Euclidean ring, any ideal is principal.

This motivates the following definition:

Definition 1.14.2. An integral domain R is said to be a **principal ideal domain** (short: **PID**) if each ideal of R is principal.

Thus, the proposition we proved yields:

Proposition 1.14.3. Any Euclidean domain is a PID.

But there are other PIDs which are not Euclidean. The examples are rather exotic. One such non-Euclidean PID is the ring

$$\mathbb{Z}[\alpha] = \{a + b\alpha \mid a, b \in \mathbb{Z}\} \qquad \text{for } \alpha = \frac{1+\sqrt{-19}}{2}.$$

For a proof, see [Dummit/Foote].

1.14.2. Divisibility in commutative rings

Let us define the notions we will study:

Definition 1.14.4. Let R be a commutative ring. Let $a \in R$.

(a) A **multiple** of a means an element of the principal ideal aR.

(b) A **divisor** of a means an element $d \in R$ such that $a \in dR$. We write "$d \mid a$" for "d is a divisor of a".

Now, let $a, b \in R$.

(c) A **common divisor** of a and b means a divisor of a that is also a divisor of b.

(d) A **common multiple** of a and b means a multiple of a that is also a multiple of b.

(e) A **greatest common divisor** (short: **gcd**) of a and b means a common divisor d of a and b such that every common divisor of a and b is a divisor of d.

(f) A **least common multiple** (short: **lcm**) of a and b means a common multiple m of a and b such that every common divisor of a and b is a multiple of m.

These definitions of gcd and lcm are perhaps not the most straightforward generalizations of the gcd and lcm from number theory, and in fact, they are not even literally unique: By this definition, both 2 and -2 are gcds of 4 and 6. Likewise, both 12 and -12 are lcms of 4 and 6. Thus we use the indefinite article "a" here.

In general, it is not guaranteed that a gcd and an lcm exist in the first place. In many rings, they often don't. And as we just noticed, when they do exist, they need not be unique.

1.14.3. Gcds and lcms for integers

Proposition 1.14.5. Let a and b be two integers. Let $g = \gcd(a, b)$ and $\ell = \operatorname{lcm}(a, b)$ in the sense of elementary number theory. Then:

(a) The gcds of a and b in the new sense are g and $-g$.

(b) The lcms of a and b in the new sense are ℓ and $-\ell$.

Proof. Easy if you remember your number theory. □

In general commutative rings, gcds might not even exist. For instance, if R is the ring

$$\mathbb{Z}\left[\sqrt{-3}\right] = \left\{a + b\sqrt{-3} \mid a, b \in \mathbb{Z}\right\},$$

then the elements $a = 4$ and $b = 2\left(1 + \sqrt{-3}\right)$ have neither a gcd nor an lcm.

What about uniqueness? As we saw, in $\mathbb{Z}$, gcds and lcms are unique up to sign. How do we generalize this "up to sign"-uniqueness to arbitrary rings?

1.14.4. Associate elements

Definition 1.14.6. Let R be a commutative ring. Let $a, b \in R$. We say that a is **associate** to b (and we write $a \sim b$) if there exists a unit u of R such that $a = bu$.

For example:

- For $R = \mathbb{Z}$, two integers a and b are associate if and only if $a = \pm b$.
- In a field, two nonzero elements are always associate.
- In a polynomial ring $F[x]$ over a field F, two polynomials f and g are associate if and only if $f = \lambda g$ for some nonzero $\lambda \in F$. In particular, any nonzero polynomial is associate to a unique monic polynomial.
- What about $R = \mathbb{Z}[i]$? The units of $\mathbb{Z}[i]$ are 1, -1, i, $-i$ (this is not hard to prove; see later). So two elements $a, b \in \mathbb{Z}[i]$ are associate if and only if a is b or $-b$ or ib or $-ib$.

Here are some general properties of associateness:

Proposition 1.14.7. Let R be a commutative ring. The relation $\sim$ is an equivalence relation.

Proof. Easy. □

Proposition 1.14.8. Let R be an integral domain. Let $a, b \in R$. Then, $a \sim b$ if and only if both $a \mid b$ and $b \mid a$ hold.

Proof. $\Longrightarrow$: Assume that $a \sim b$. Thus, $a = bu$ for some unit u, so that $b \mid a$. Moreover, $a = bu$ yields $b = au^{-1}$ (since u is a unit), so that $a \mid b$.

$\Longleftarrow$: Assume that $a \mid b$ and $b \mid a$. That is, $b = ax$ and $a = by$ for some $x, y \in R$. So $b = \underbrace{a}_{=by} x = byx = bxy$. If $b \neq 0$, then this entails $1 = xy$ because we can cancel b in the integral domain R; thus we conclude that x is a unit (since R is commutative) and so $a \sim b$. If $b = 0$, then $a = 0$ (since $b \mid a$), and so $a \sim b$ (since $a = b \cdot 1$). □

The condition "R is an integral domain" cannot be omitted in this last proposition, but the counterexamples are rather obscure.

Associate elements "look the same" to divisibility, meaning that in a divisibility $a \mid b$ we can always replace a by any element associate to a, or replace b by any element associate to b. In other words:

Proposition 1.14.9. Let R be a commutative ring. Let $a \sim a'$ and $b \sim b'$. Then, $a \mid b$ if and only if $a' \mid b'$.

1.14.5. Uniqueness of gcds and lcms in an integral domain

Proposition 1.14.10. Let R be an integral domain. Let $a, b \in R$. Then:
(a) Any two gcds of a and b are associate.
(b) Any two lcms of a and b are associate.

Proof. **(a)** Let g and h be two gcds of a and b. We must show that $g \sim h$.

By one of the propositions above, it suffices to show that $g \mid h$ and $h \mid g$.

Why does $g \mid h$? Because g is a common divisor of a and b, but h is a greatest common divisor of a and b and thus divisible by any common divisor of a and b, including g.

So $g \mid h$. Similarly, $h \mid g$. And we are done.

(b) Analogous. □

The proposition we just proved is the most reasonable sense in which gcds and lcms are unique.

1.14.6. Existence of gcds and lcms in a PID

Existence is a trickier question: Sometimes, two elements of an integral domain have a gcd; sometimes, they don't. At least in a PID (and thus in any Euclidean domain), gcds and lcms always exist:

Theorem 1.14.11. Let R be a PID (for example, a Euclidean domain). Let $a, b \in R$. Then, there exist a gcd and an lcm of a and b.

More concretely:

Proposition 1.14.12. Let R be a commutative ring. Let $a, b, c \in R$. Then:
(a) If $aR + bR = cR$, then c is a gcd of a and b. [NB: This is **not** an "if and only if"!]
(b) If $aR \cap bR = cR$, then c is an lcm of a and b. [This is actually an "if and only if".]

Proof. **(a)** Assume that $aR + bR = cR$. Then,

$$a = a \cdot 1 + b \cdot 0 \in aR + bR = cR,$$

so that c is a divisor of a. Similarly, c is a divisor of b. Hence, c is a common divisor of a and b.

Why is it a greatest common divisor of a and b ? We must prove that $d \mid c$ for any common divisor d of a and b. So let d be a common divisor of a and b. Why is $d \mid c$? We have

$$c = c \cdot 1 \in cR = aR + bR,$$

so that $c = ax + by$ for some $x, y \in R$. But $a = du$ and $b = dv$ for some $u, v \in R$ (since d divides a and b). Thus,

$$c = \underbrace{a}_{=du} x + \underbrace{b}_{=dv} y = dux + dvy = d\left(ux + vy\right) \in dR.$$

In other words, $d \mid c$, as desired. So part **(a)** is proved.

(b) The equation $aR \cap bR = cR$ is saying literally that "the common multiples of a and b are the multiples of c". This means precisely that c is an lcm of a and b (why?). □

Warning 1.14.13. Adding principal ideals like aR and bR is ideal addition, not addition of elements. So $aR + bR \neq (a + b) R$. Indeed, $(a + b) R$ consists of the multiples of $a + b$, whereas $aR + bR$ consists of the sums of a multiple of a with a multiple of b. Usually there are many more of the latter than of the former, so $(a + b) R$ is a proper subset of $aR + bR$.

Proof of the theorem. Since R is a PID, the ideal $aR + bR$ is principal. This means that $aR + bR = cR$ for some $c \in R$. By part **(a)** of the above proposition, this c must then be a gcd of a and b. So a and b have a gcd.

Similarly, a and b have an lcm. □

This proof is, of course, non-constructive, since it relies on a magical machine that writes every ideal of R as cR for some $c \in R$, and this cannot be done algorithmically even for $R = \mathbb{Z}$. However, for many rings R, we can make the theorem constructive, by finding algorithms to write sums $aR + bR$ and intersections $aR \cap bR$ in the form cR. One important class of such algorithms is the **extended Euclidean algorithm**, which finds a gcd and an lcm whenever R is a Euclidean ring. In the notes, I do this in some detail (§2.13.3). Let me just mention the main idea: To find $\gcd(a, b)$, we replace (a, b) by (b, r), where r is a remainder upon division of a by b (that is, r is the r in a pair (q, r) satisfying $a = qb + r$ and $r = 0$ or $N(r) < N(b)$, as required in the definition of a Euclidean ring). This computes a gcd of a and b. Then, to find the lcm, we use the formula

$$\gcd(a, b) \cdot \operatorname{lcm}(a, b) \sim ab,$$

which holds for any two elements a, b of an integral domain R that have a gcd and an lcm (see Winter 2021 homework set #2 Exercise 3).

Last time, we saw a bunch of Euclidean domains, such as $\mathbb{Z}[i]$. So we obtain algorithms for gcds and lcms in all of these Euclidean domains.

1.14.7. Irreducible and prime elements

Now let us generalize the prime numbers from the ring $\mathbb{Z}$ to an arbitrary commutative ring R. There are two ways to do so, both useful. They start from two different characterizations of prime numbers in $\mathbb{Z}$:

1. The prime numbers are the integers $p > 1$ that have no positive divisors apart from 1 and p. In other words, they are the integers $p > 1$ such that whenever p is written as $p = ab$ for $a, b \in \mathbb{Z}$, at least one of a and b must be ± 1.

2. They are the integers $p > 1$ with the property that if $p \mid ab$, then $p \mid a$ or $p \mid b$.

You know that these two characterizations are equivalent... for numbers, but not for elements of a general commutative ring! To generalize them to arbitrary rings, we have to treat them separately. The "charactierzation-1" prime numbers are called **irreducible** elements, whereas the "charactierzation-2" prime numbers are called **prime** elements. In other words:

Definition 1.14.14. Let R be a commutative ring. Let $r \in R$ be nonzero and not a unit.

(a) We say that r is **irreducible** (in R) if it has the following property: Whenever $a, b \in R$ satisfy $r = ab$, at least one of a and b is a unit.

(b) We say that r is **prime** (in R) if it has the following property: Whenever $a, b \in R$ satisfy $r \mid ab$, we have $r \mid a$ or $r \mid b$.

Note that we have replaced the condition "$p > 1$" by "r is nonzero and not a unit". This preserves the spirit of the requirement (we don't want to count 0 and 1 as primes), but allows for a bit more "freedom of association", meaning that with any irreducible or prime element, all elements associate to it will also be irreducible or prime. In particular, the integer -2, while not a prime number by definition, does count as a prime element of $\mathbb{Z}$.

When $R = \mathbb{Z}$, then the prime and the irreducible elements are exactly the prime numbers and their negatives:

Proposition 1.14.15. Let $r \in \mathbb{Z}$. Then, we have the following equivalence:

$$(r \text{ is prime in } \mathbb{Z}) \iff (r \text{ is irreducible in } \mathbb{Z}) \iff (|r| \text{ is a prime number}).$$

Proof. Easy. □

Thus, in the ring $\mathbb{Z}$, being prime and being irreducible is the same thing. In an arbitrary integral domain, this is not always the case. For instance:

- In the ring $\mathbb{Z}\left[\sqrt{-5}\right]$, the element 3 is irreducible but not prime.
- Here is an example using polynomials: Consider the univariate polynomial ring

$$\mathbb{Q}[x] = \left\{a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \mid a_i \in \mathbb{Q}\right\}$$

and its subring

$$\begin{aligned} R &= \left\{a_0 + a_2 x^2 + \cdots + a_n x^n \mid a_i \in \mathbb{Q}\right\} \\ &= \left\{f \in \mathbb{Q}[x] \mid \text{the } x^1\text{-coefficient of } f \text{ is } 0\right\} \end{aligned}$$

(yes, this is a subring – check it!). The element x^3 of R is irreducible, but not prime: $x^3 \mid x^2 x^2$ but $x^3 \nmid x^2$.

The converse, however, never happens: A prime element is always irreducible.

More on this next time.

The following connection between "irreducible" and "prime" has already been noticed in Lecture 16:

Proposition 1.14.16. Let R be an integral domain. Then, any prime element of R is irreducible.

In a PID (= integral domain where each ideal is principal), this goes both ways:

Proposition 1.14.17. Let R be a PID (for example, a Euclidean domain). Let $r \in R$. Then, r is prime if and only if r is irreducible.

One proof of this proposition is given in the text (Proposition 2.15.4); we shall give another. This latter proof will rely on the following two general lemmas:

Lemma 1.14.18. Let R be an integral domain. Let $a, b, c \in R$ be such that $a \neq 0$. If $ab \mid ac$, then $b \mid c$.

Proof. Assume that $ab \mid ac$. In other words, $ac = abr$ for some $r \in R$. Consider this r. We have $a(c - br) = ac - abr = 0$ (since $ac = abr$). Since R is an integral domain, we thus conclude that $c - br = 0$ (since $a \neq 0$). In other words, $c = br$. This entails that $b \mid c$. Thus, Lemma 1.14.18. □

Lemma 1.14.19. Let R be an integral domain. Let $a, b, c \in R$ be such that $a \neq 0$. Assume that the elements ab and ac have a gcd g. Then, the elements b and c have a gcd h such that $g = ah$.

Proof. The element a is a common divisor of ab and ac (obviously), and thus must divide g (by the definition of a gcd, since g is a gcd of ab and ac). In other words, there exists some $r \in R$ such that $g = ar$. Consider this r.

Since g is a gcd of ab and ac, we have $g \mid ab$ and $g \mid ac$. From $ar = g \mid ab$, we obtain $r \mid b$ (by Lemma 1.14.18, applied to r and b instead of b and c). Similarly, $r \mid c$. Thus, r is a common divisor of b and c.

Now, let s be any common divisor of b and c. Then, $s \mid b$, so that $b = sb'$ for some $b' \in R$. This b' then satisfies $a \underbrace{b}_{=sb'} = asb'$, so that $as \mid ab$. Similarly, $as \mid ac$. Hence, as is a common divisor of ab and ac. Therefore, as divides g (since g is a gcd of ab and ac). In other words, $as \mid g = ar$. Hence, Lemma 1.14.18 (applied to s and r instead of b and c) yields $s \mid r$.

Forget that we fixed s. We thus have shown that any common divisor s of b and c satisfies $s \mid r$. In other words, any common divisor of b and c divides r. Since we also know that r is a common divisor of b and c, we thus conclude that r is a gcd of b and c (by the definition of a gcd). Hence, the elements b and c have a gcd h such that $g = ah$ (namely, $h = r$), since we know that $g = ar$. This proves Lemma 1.14.19. □

Proof of Proposition 1.14.17. $\Longrightarrow$: This follows from Proposition 1.14.16.

$\Longleftarrow$: Assume that r is irreducible. We must prove that r is prime.

So let $a, b \in R$ be such that $r \mid ab$. We must show that $r \mid a$ or $r \mid b$.

If $a = 0$, then this is obvious (since $r \mid 0$). Thus, we WLOG assume that $a \neq 0$.

Theorem 1.14.11 in Lecture 16 shows that ab and ar have a gcd in R. Let g be this gcd. Then, $g \mid ab$ and $g \mid ar$.

We have $r \mid ab$ (by assumption) and $r \mid ar$ (obviously). Hence, r is a common divisor of ab and ar. Thus, r divides g (by the definition of a gcd, since g is a gcd of ab and ar). That is, $r \mid g$.

Lemma 1.14.19 (applied to $c = r$) yields that the elements b and r have a gcd h such that $g = ah$. Consider this h. Since h is a gcd of b and r, we have $h \mid b$ and $h \mid r$.

In particular, $h \mid r$. In other words, there exists some $k \in R$ such that $r = kh$. Consider this k.

So we have $kh = r$. Since r is irreducible, this entails that at least one of the elements k and h is a unit (by the definition of "irreducible"). Thus, we are in one of the following cases:

Case 1: The element k is a unit.

Case 2: The element h is a unit.[1]

[1] Cases 1 and 2 cannot overlap, because if k and h were both units, then their product $kh = r$ would be a unit as well, but r is irreducible and thus not a unit. But we don't care about this, since cases in a proof can overlap.

Let us first consider Case 1. In this case, the element k is a unit. Hence, k has an inverse k^{-1}. From $r = kh$, we thus obtain $h = rk^{-1}$, so that $r \mid h \mid b$. Thus, $r \mid a$ or $r \mid b$. So we are done in Case 1.

Let us next consider Case 2. In this case, the element h is a unit. Hence, h has an inverse h^{-1}. From $g = ah$, we thus obtain $a = gh^{-1}$. Thus, $g \mid a$. Hence, $r \mid g \mid a$. Thus, $r \mid a$ or $r \mid b$. So we are done in Case 2.

Hence, in both cases, we have shown that $r \mid a$ or $r \mid b$. As we explained, this completes the proof of the "$\Longleftarrow$" direction of Proposition 1.14.17. □

1.14.8. Irreducible factorizations and UFDs

The following theorem generalizes the classical "Fundamental Theorem of Arithmetic" (i.e., the fact that each positive integer has a prime factorization, which is unique up to reordering the factors):

Theorem 1.14.20. Let R be a PID. Then, any nonzero element $r \in R$ can be decomposed (up to associates) into a product of irreducible (i.e., prime) elements of R. Moreover, this product is unique up to order and associateness.

In detail: Let $r \in R$ be a nonzero element. Then, there is a tuple $(p_1, p_2, \ldots, p_n)$ of irreducible (i.e., prime) elements of R such that

$$r \sim p_1 p_2 \cdots p_n.$$

If $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_m)$ are two such tuples, then $(p_1, p_2, \ldots, p_n)$ can be obtained from $(q_1, q_2, \ldots, q_m)$ by reordering the entries and replacing them by associate entries.

Proof. See a textbook, e.g., [**?**, §8.3, Theorem 14] or [**?**, Theorem 8.15]. Just a few words about the proof:

Uniqueness is proved just as for integers.

Existence is tricky: Just as for integers, you start with r and keep factoring it further and further (avoiding unit factors) until no more divisors remain. But you have to argue that this factoring process won't go on forever, and this is no longer as easy as for integers. (It is easy when R has a "multiplicative norm", i.e., a map $N : R \to \mathbb{N}$ such that $N(a) < N(ab)$ whenever $a, b \in R$ are nonzero and b is not a unit. For example, if $R = \mathbb{Z}[i]$, then the Euclidean norm $N : \mathbb{Z}[i] \to \mathbb{N}$ defined by $N(a + bi) = a^2 + b^2$ has this property.) □

Integral domains R in which the claim of Theorem 1.14.20 holds are called **UFDs** (short for **unique factorization domains**). This class is wider than the PIDs. For instance, the polynomial rings $\mathbb{Z}[x]$ (the ring of all univariate polynomials in x with integer coefficients) and $\mathbb{Q}[x, y]$ (the ring of all polynomials in two variables x and y with rational coefficients) are UFDs but not PIDs.

We will not focus on UFDs in this course, but we briefly note that they have some (but not all) of the nice properties of PIDs. In particular, in a UFD, any two elements have a gcd and an lcm. (But we shall not prove this.)

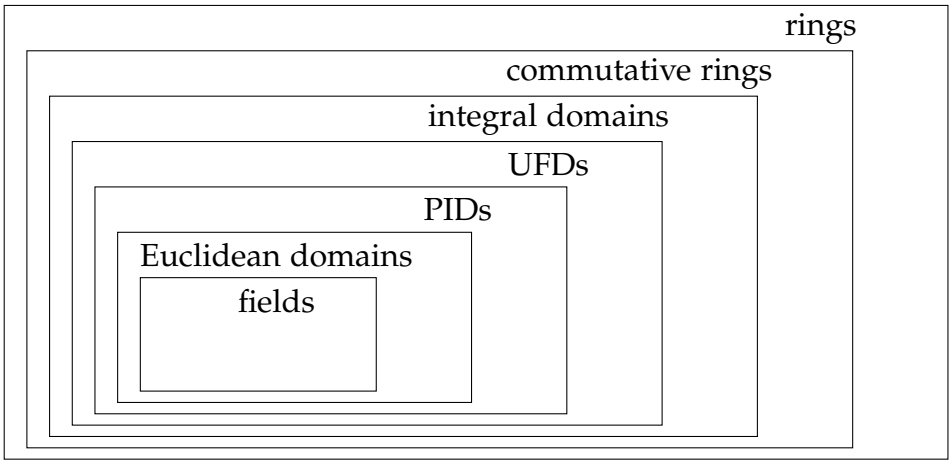
1.14.9. A synopsis

The following corollary combines several results we have seen above in a convenient hierarchy:

Corollary 1.14.21. We have

$$\{\text{fields}\} \subseteq \{\text{Euclidean domains}\} \subseteq \{\text{PIDs}\} \subseteq \{\text{UFDs}\}$$
$$\subseteq \{\text{integral domains}\} \subseteq \{\text{commutative rings}\} \subseteq \{\text{rings}\}.$$

Let us illustrate this hierarchy in a symbolic picture:



All the "$\subseteq$" signs in Corollary 1.14.21 are strict inclusions; let us briefly recall some examples showing this:

- The rings $\mathbb{Z}$ and $\mathbb{Z}[i]$ and $\mathbb{Z}\left[\sqrt{-2}\right]$ and $\mathbb{Z}\left[\sqrt{2}\right]$ and the polynomial ring $\mathbb{Q}[x]$ are Euclidean domains, but not fields.
- The ring $\mathbb{Z}[\alpha]$ for $\alpha = \dfrac{1+\sqrt{-19}}{2}$ is a PID, but not a Euclidean domain.
- The polynomial rings $\mathbb{Q}[x,y]$ and $\mathbb{Z}[x]$ are UFDs, but not PIDs.
- The rings $\mathbb{Z}[2i]$ and $\mathbb{Z}\left[\sqrt{-3}\right]$ are integral domains, but not UFDs.
- The ring $\mathbb{Z}/6 \cong \mathbb{Z}/2 \times \mathbb{Z}/3$ is a commutative ring, but not an integral domain.
- The matrix ring $\mathbb{Q}^{2\times 2}$ and the ring of quaternions $\mathbb{H}$ are not commutative.

1.15. Application: Fermat's $p = x^2 + y^2$ theorem

As an application of some of the above, we will show a result of Fermat:

Theorem 1.15.1 (Fermat's two-squares theorem)**.** Let p be a prime number[2] such that $p \equiv 1 \bmod 4$. Then, p can be written as a sum of two perfect squares.

For example,

$$5 = 1^2 + 2^2;$$
$$13 = 2^2 + 3^2;$$
$$17 = 1^2 + 4^2;$$
$$29 = 2^2 + 5^2.$$

I will prove Theorem 1.15.1 using rings (specifically, the ring $\mathbb{Z}/p$ of residue classes and the ring $\mathbb{Z}[i]$ of Gaussian integers). The first ingredient of the proof is a curious fact about primes, known as **Wilson's theorem**:

Theorem 1.15.2 (Wilson's theorem)**.** Let p be a prime. Then, $(p-1)! \equiv -1 \bmod p$.

Proof. We must show that $\overline{(p-1)!} = \overline{-1}$ in $\mathbb{Z}/p$.

In $\mathbb{Z}/p$, we have

$$\overline{(p-1)!} = \overline{1 \cdot 2 \cdot \cdots \cdot (p-1)} = \overline{1} \cdot \overline{2} \cdot \cdots \cdot \overline{p-1}. \tag{1}$$

Recall that every ring R has a group of units, which is denoted by $R^\times$. (Its elements are the units of R, and its operation is multiplication.) Since the ring $\mathbb{Z}/p$ is a field (because p is prime), its group of units $(\mathbb{Z}/p)^\times$ consists of all nonzero elements of $\mathbb{Z}/p$. Thus,

$$(\mathbb{Z}/p)^\times = \left\{\overline{1}, \overline{2}, \ldots, \overline{p-1}\right\},$$

with all the $p-1$ elements $\overline{1}, \overline{2}, \ldots, \overline{p-1}$ being distinct. Hence, the product of all units of $\mathbb{Z}/p$ is

$$\prod_{a \in (\mathbb{Z}/p)^\times} a = \overline{1} \cdot \overline{2} \cdot \cdots \cdot \overline{p-1}.$$

Comparing this with (1), we find

$$\overline{(p-1)!} = \prod_{a \in (\mathbb{Z}/p)^\times} a. \tag{2}$$

[2] in the sense of classical number theory, i.e., an integer $p > 1$ with no positive divisors other than 1 and p

Now, recall that any unit a of $\mathbb{Z}/p$ (or of any other ring) has an inverse a^{-1}, which is also a unit and satisfies $\left(a^{-1}\right)^{-1} = a$. Thus, the units of $\mathbb{Z}/p$ can be paired up in pairs $\left\{a, a^{-1}\right\}$ consisting of a unit a and its inverse a^{-1}. The only units left unpaired will be the units that are their own inverses. These units are the elements $a \in \mathbb{Z}/p$ that satisfy $a^2 = \overline{1}$, and a moment of thought reveals that they are $\overline{1}$ and $\overline{-1}$ (because $a^2 = \overline{1}$ entails $0 = a^2 - \overline{1} = \left(a - \overline{1}\right)\left(a + \overline{1}\right)$, and since $\mathbb{Z}/p$ is an integral domain, this equality can only hold if either $a - \overline{1}$ or $a + \overline{1}$ is 0). Thus, all units other than $\overline{1}$ and $\overline{-1}$ are paired. Hence, in the product of all units of $\mathbb{Z}/p$, we can pair up each factor other than $\overline{1}$ and $\overline{-1}$ with its inverse:

$$\prod_{a \in (\mathbb{Z}/p)^\times} a = \underbrace{\left(a_1 \cdot a_1^{-1}\right)}_{=1} \cdot \underbrace{\left(a_2 \cdot a_2^{-1}\right)}_{=1} \cdot \cdots \cdot \underbrace{\left(a_k \cdot a_k^{-1}\right)}_{=1} \cdot \overline{1} \cdot \overline{-1} = \overline{-1}.$$

Hence, (2) can be rewritten as $\overline{(p-1)!} = \overline{-1}$, which means precisely that $(p-1)! \equiv -1 \bmod p$. This proves Theorem 1.15.2.

(Caution: The above argument breaks down a bit for $p = 2$, but this case is trivial anyway.) □

Corollary 1.15.3. Let p be an odd prime (i.e., a prime distinct from 2). Let $u = \dfrac{p-1}{2} \in \mathbb{N}$. Then, $u!^2 \equiv -(-1)^u \bmod p$.

Proof. Theorem 1.15.2 yields

$$(p-1)! \equiv -1 \bmod p.$$

However,

$$\begin{aligned}
(p-1)! &= 1\cdot 2\cdot \cdots \cdot (p-1) \\
&= \left(1\cdot 2\cdot \cdots \cdot \frac{p-1}{2}\right) \cdot \left(\underbrace{\left(\frac{p-1}{2}+1\right)}_{\equiv -\frac{p-1}{2} \bmod p} \cdot \cdots \cdot \underbrace{(p-2)}_{\equiv -2 \bmod p} \cdot \underbrace{(p-1)}_{\equiv -1 \bmod p} \right) \\
&\equiv \left(1\cdot 2\cdot \cdots \cdot \frac{p-1}{2}\right) \cdot \left(\left(-\frac{p-1}{2}\right) \cdot \cdots \cdot (-2)\cdot(-1)\right) \\
&= (1\cdot 2\cdot \cdots \cdot u) \cdot \underbrace{((-u)\cdot \cdots \cdot (-2)\cdot(-1))}_{\substack{=(-1)^u\cdot(u\cdot\cdots\cdot 2\cdot 1) \\ =(-1)^u\cdot(1\cdot 2\cdot\cdots\cdot u)}} \qquad \left(\text{since } \frac{p-1}{2} = u\right) \\
&= (1\cdot 2\cdot \cdots \cdot u)\cdot (-1)^u \cdot (1\cdot 2\cdot \cdots \cdot u) \\
&= (-1)^u \cdot \left(\underbrace{1\cdot 2\cdot \cdots \cdot u}_{=u!}\right)^2 = (-1)^u \cdot u!^2 \bmod p,
\end{aligned}$$

so that

$$(-1)^u \cdot u!^2 \equiv (p-1)! \equiv -1 \bmod p.$$

Multiplying both sides of this congruence by $(-1)^u$, we obtain

$$u!^2 \equiv -(-1)^u \bmod p.$$

This proves Corollary 1.15.3. □

Corollary 1.15.4. Let p be a prime such that $p \equiv 1 \bmod 4$. Let $u = \dfrac{p-1}{2} \in \mathbb{N}$. Then, $u!^2 \equiv -1 \bmod p$.

Proof. Apply Corollary 1.15.3, and observe that $(-1)^u = 1$ (since $p \equiv 1 \bmod 4$, so that u is even). Corollary 1.15.4 follows. □

This corollary shows that $p \mid u!^2 + 1 = (u+i)(u-i)$ in $\mathbb{Z}[i]$.

If p was prime in $\mathbb{Z}[i]$, then this would entail that $p \mid u+i$ or $p \mid u-i$ (by the definition of "prime"), but this is not the case, since $\dfrac{u+i}{p} = \dfrac{u}{p} + \dfrac{1}{p}i \notin \mathbb{Z}[i]$ and $\dfrac{u+i}{p} = \dfrac{u}{p} - \dfrac{1}{p}i \notin \mathbb{Z}[i]$. So p (while prime in $\mathbb{Z}$) cannot be a prime in $\mathbb{Z}[i]$. Since $\mathbb{Z}[i]$ is a PID, this entails that p is not irreducible either (since prime = irreducible in a PID). In other words, we can factor p as $p = \alpha\beta$ for two Gaussian integers $\alpha, \beta \in \mathbb{Z}[i]$ that are both non-units.

How do we make anything useful out of this? We recall our favorite Euclidean norm N on $\mathbb{Z}[i]$. This is the map

$$\begin{aligned} N : \mathbb{Z}[i] &\to \mathbb{N}, \\ a + bi &\mapsto a^2 + b^2. \end{aligned}$$

It has some nice properties:

Proposition 1.15.5. For any $\alpha, \beta \in \mathbb{Z}[i]$, we have $N(\alpha\beta) = N(\alpha) N(\beta)$.

Proof. Straightforward by computation. Or observe that $N(z) = |z|^2$ and $|\alpha\beta| = |\alpha| |\beta|$. □

Corollary 1.15.6. If $\alpha \mid \beta$ in $\mathbb{Z}[i]$, then $N(\alpha) \mid N(\beta)$.

Proof. The assumption $\alpha \mid \beta$ means $\beta = \alpha\gamma$, hence $N(\beta) = N(\alpha\gamma) = N(\alpha) N(\gamma)$. □

Corollary 1.15.7. The units of $\mathbb{Z}[i]$ are exactly the elements $\alpha \in \mathbb{Z}[i]$ with norm $N(\alpha) = 1$, and these elements are precisely 1, i, -1, $-i$.

Proof. If $\alpha \in \mathbb{Z}[i]$ is a unit, then $\alpha \mid 1$, so the previous corollary yields $N(\alpha) \mid N(1) = 1$, and therefore $N(\alpha) = 1$ (since $N(\alpha)$ is a nonnegative integer). If $N(\alpha) = 1$, then α is one of 1, i, -1, $-i$, by a fairly simple case analysis. Finally, if α is one of 1, i, -1, $-i$, then α is a unit, since you can just exhibit its inverse. □

Now all is in place for proving Fermat's two-squares theorem:

Proof of Fermat's two-squares theorem. As we showed above, p is not irreducible in $\mathbb{Z}[i]$. But p is nonzero and not a unit (e.g., by the last corollary). So $p = \alpha\beta$ for two non-units $\alpha, \beta \in \mathbb{Z}[i]$ (by the definition of "irreducible"). Consider these α, β.

From $p = \alpha\beta$, we obtain $N(p) = N(\alpha\beta) = N(\alpha) N(\beta)$. So $N(\alpha) N(\beta) = N(p) = p^2$. Thus, $N(\alpha)$ and $N(\beta)$ are two nonnegative divisors of p^2 that multiply together to form p^2. Moreover, these two divisors are not 1 (by the last corollary), since α and β are not units. Thus, $N(\alpha)$ and $N(\beta)$ must be p and p. In particular, $N(\alpha) = p$. Writing α as $a + bi$, this means that $p = N(\alpha) = a^2 + b^2$, qed. □

To recap, this proof proceeded roughly as follows:

- We showed that there is some perfect square that is $\equiv -1 \bmod p$. (This square was $u!^2$, but we don't need the details.)
- We used this to obtain a product of two Gaussian integers that is divisible by p without either factor being divisible by p. (Specifically: $x^2 \equiv -1 \bmod p$, then $p \mid (x+i)(x-i)$ but $p \nmid x+i$ and $p \nmid x-i$.)

- We concluded that p cannot be prime in $\mathbb{Z}[i]$, hence cannot be irreducible, so $p = \alpha\beta$ for some non-units α, β.
- We showed that $N(\alpha) = p$ because there isn't much freedom left for $N(\alpha)$ and $N(\beta)$ when $p = \alpha\beta$.

Note that one thing we learned is that a prime in $\mathbb{Z}$ does not have to remain a prime in $\mathbb{Z}[i]$.

Is the above proof constructive? Does it give an algorithm for finding a, b that satisfy $p = a^2 + b^2$? Yes, though we need to read the proof correctly (including, most importantly, the proof of "not prime $\Longrightarrow$ not irreducible") to really bring out the algorithmic content. That said, the algorithm it gives for computing a and b is probably not the best one.

Fermat's two-squares theorem is just the tip of an iceberg, which was being explored for the last few centuries and is still less than fully mapped. The first step beyond it is to extend the theorem to non-primes:

Theorem 1.15.8. Let n be a positive integer with prime factorization $n = 2^a p_1^{b_1} p_2^{b_2} \cdots p_k^{b_k}$, where $p_1, p_2, \ldots, p_k$ are distinct primes > 2, and where $a, b_1, b_2, \ldots, b_k$ are nonnegative integers. Then:

(a) The number n can be written as a sum of two perfect squares if and only if the following condition holds: For each $i \in \{1, 2, \ldots, k\}$ satisfying $p_i \equiv 3 \bmod 4$, the exponent b_i is even.

(b) If this condition holds, then the number of ways to write n as a sum of two perfect squares (i.e., the number of pairs $(x, y) \in \mathbb{Z} \times \mathbb{Z}$ such that $n = x^2 + y^2$) is

$$\prod_{\substack{i \in \{1,2,\ldots,k\}; \\ p_i \equiv 1 \bmod 4}} (b_i + 1).$$

This can also be proved using $\mathbb{Z}[i]$. See my 2019 notes for the details (or [Dummit/Foote]).

More about decompositions of numbers into sums of perfect squares can be found in various texts referenced in the notes. Let me shift to a slight variation of the problem, where we replace $x^2 + y^2$ by $x^2 + 2y^2$ or $x^2 + 3y^2$ or $x^2 + xy + y^2$ or $x^2 - 2y^2$ or many other such expressions. A whole book has been written about such questions [Cox: *Primes of the form* $x^2 + ny^2$, 3rd edition 2022]. The simpler of the questions can be answered with similar methods as above, just using $\mathbb{Z}[2i]$ or $\mathbb{Z}\left[\sqrt{-3}\right]$ or similar rings instead of $\mathbb{Z}[i]$. These work well as long as these rings are PIDs. In harder situations, proofs have been seen using quadratic forms, elliptic curves, elliptic functions, ... Here is a potpourri of answers for the forms $x^2 + ny^2$ for certain values of n:

Theorem 1.15.9. Let p be a prime number.

(a) We can write p as $p = x^2 + y^2$ with $x, y \in \mathbb{Z}$ if and only if $p = 2$ or $p \equiv 1 \bmod 4$.

(b) We can write p as $p = x^2 + 2y^2$ with $x, y \in \mathbb{Z}$ if and only if $p \equiv 1, 3 \bmod 8$. (The comma means "or": i.e., we are saying "$p \equiv 1 \bmod 8$ or $p \equiv 3 \bmod 8$".)

(c) We can write p as $p = x^2 + 3y^2$ with $x, y \in \mathbb{Z}$ if and only if $p = 3$ or $p \equiv 1 \bmod 3$.

(d) We can write p as $p = x^2 + 4y^2$ with $x, y \in \mathbb{Z}$ if and only if $p \equiv 1 \bmod 4$.

(e) We can write p as $p = x^2 + 5y^2$ with $x, y \in \mathbb{Z}$ if and only if $p \equiv 1, 9 \bmod 20$.

(f) We can write p as $p = x^2 + 6y^2$ with $x, y \in \mathbb{Z}$ if and only if $p \equiv 1, 7 \bmod 24$.

(g) We can write p as $p = x^2 + 14y^2$ with $x, y \in \mathbb{Z}$ if and only if $p \equiv 1, 9, 15, 23, 25, 39 \bmod 56$ and there exists some integer z satisfying $\left(z^2 + 1\right)^2 \equiv 8 \bmod p$.

(h) We can write p as $p = x^2 + 27y^2$ with $x, y \in \mathbb{Z}$ if and only if we have $p \equiv 1 \bmod 3$ and there exists some integer z satisfying $z^3 \equiv 2 \bmod p$.

Cox's book discusses most of these. Part **(d)** is a homework exercise. Part **(b)** is an exercise in the text. Part **(c)** can still be done with similar methods, but is a bit trickier (it requires working not in $\mathbb{Z}\left[\sqrt{-3}\right]$ but in the slightly larger ring of **Eisenstein integers**). Part **(e)** is proved using genus theory of quadratic forms. Part **(f)** requires class field theory. Parts **(g)** and **(h)** are proved using elliptic functions. Conditions like the ones in part **(g)** and **(h)** are unavoidable for larger coefficients.

We can also ask about sums of more than two squares. Lagrange proved that every nonnegative integer can be written as a sum of **four** squares. These days, one of the shortest proof uses the so-called **Hurwitz quaternions**.

2. Modules

Roughly speaking, a ring is a system of "number-like objects" that can be "added" and "multiplied".

In contrast, a **module** over a given ring R is a system of "vector-like objects" that can be "added" and "scaled" (by elements of R). Thus, modules generalizes vector spaces.

2.1. Definitions and examples

Convention 2.1.1. We shall fix a ring R for the rest of this section.

2.1.1. Defining modules

Modules come in two forms: left modules and right modules. Let us define the left ones:

Definition 2.1.2. Let R be a ring. A **left R-module** (or a **left module over** R) means a set M equipped with

- a binary operation $+$ (that is, a map from $M \times M$ to M) that is called **addition**;
- an element 0_M of M that is called the **zero element** or the **zero vector** or just the **zero**, and will often just be called 0;
- a map from $R \times M$ to M that is called the **action of R on M**, and is written as multiplication (i.e., we denote the image of a pair (r, m) under this map as rm or $r \cdot m$)

such that the following **module axioms** hold:

- $(M, +, 0)$ is an abelian group.
- The **right distributivity law** holds: We have $(r + s)\, m = rm + sm$ for all $r, s \in R$ and $m \in M$.
- The **left distributivity law** holds: We have $r\,(m + n) = rm + rn$ for all $r \in R$ and $m, n \in M$.
- The **associativity law** holds: We have $(rs)\, m = r\,(sm)$ for all $r, s \in R$ and $m \in M$.
- We have $0_R m = 0_M$ for all $m \in M$.
- We have $r \cdot 0_M = 0_M$ for all $r \in R$.
- We have $1m = m$ for all $m \in M$.

When M is a left R-module, the elements of M are called **vectors**, while the elements of R are called **scalars**.

As the name "left R-module" suggests, there is a mirror notion of "**right R-modules**", in which the action is not a map $R \times M \to M$ but a map $M \times R \to M$, and its values are denoted by mr rather than rm. Correspondingly, the associativity law for a right R-module takes the form $m\,(rs) = (mr)\, s$.

When the ring R is commutative, any left R-module M becomes a right R-

module by setting

$$mr := rm \qquad \text{for all } r \in R \text{ and } m \in M,$$

and conversely, any right R-module becomes a left R-module by setting

$$rm := mr \qquad \text{for all } r \in R \text{ and } m \in M.$$

For example, the associativity law for a left R-module, $(rs)\,m = r\,(sm)$, becomes $m\,(rs) = (ms)\,r$. But the associativity law for a right R-module says that $m\,(rs) = (mr)\,s$. When R is commutative, the two laws are equivalent because we can apply the former law to s and r instead of r and s and recall that $sr = rs$. So left vs. right R-modules over a commutative ring R are "the same objects" except for notational differences. Not so when R is non-commutative.

If R is not commutative, then the left R-modules cannot be transformed into right R-modules. However, they can be transformed into right R^{op}-modules, where R^{op} is the **opposite ring** of R, which is the same ring as R but with the order of factors in its multiplication flipped. So you can translate between left modules and right modules at the cost of changing the base ring from R to R^{op}. In practice, this allows us to prove theorems about left modules and automatically conclude that analogous theorems are true for right modules, as long as these theorems are not particular to a specific ring R but allow for general R.

When R is commutative, we will often just speak of "R**-modules**" to mean left or right R-modules, as we desire (since the two concepts are equivalent).

When R is a field, the R-modules are also known as the R**-vector spaces**. These are precisely the vector spaces from linear algebra. However, this case is not representative of the complexity that R-modules can have when R is not a field. Vector spaces over a field always have bases and dimensions; in contrast, R-modules for general R rarely do so. The wilder the ring R is, the less well-behaved and the more diverse are the R-modules.

Another piece of terminology:

Definition 2.1.3. Let M be a left R-module, and let $r \in R$ be a scalar. Then, the map

$$\begin{aligned} M &\to M, \\ m &\mapsto rm \end{aligned}$$

is called **scaling by** r.

This map is a group morphism from the additive group $(M, +, 0)$ to itself. Scaling by 1 is the identity map on M, whereas scaling by 0 sends every vector to the zero vector.

2.1.2. Defining submodules

Definition 2.1.4. Let M be a left R-module. An **R-submodule** (or, to be more precise, a **left R-submodule**) of M means a subset N of M such that

- N is **closed under addition**, i.e., we have $a + b \in N$ for all $a, b \in N$;
- N is **closed under scaling**, i.e., we have $ra \in N$ for all $r \in R$ and $a \in N$;
- N **contains zero**, i.e., we have $0_M \in N$.

We will soon see that any such R-submodule is also closed under negation, so it becomes a left R-module in its own right.

All of this applies "mutatis mutandis" to right R-modules.

2.1.3. Examples

- Let R be a ring. Then, R itself becomes a left R-module: Just define the action to be the multiplication of R. Thus, the elements of R serve both as the scalars and as the vectors. Scaling is just multiplying.

  The R-submodules of this left R-module R are the subsets L of R that are closed under addition and contain 0 and satisfy $ra \in L$ for all $r \in R$ and $a \in L$. These subsets L are called **left ideals** of R. They differ from the ideals of R in that they don't need to satisfy $ar \in L$ for all $r \in R$ and $a \in L$. Correspondingly, many rings have a lot more left ideals than ideals (any ideal is a left ideal, but not vice versa). For example, if R is the matrix ring $\mathbb{Q}^{2\times 2}$, then R has only two ideals ($\{0\}$ and R itself), but lots of left ideals (e.g., the set $\begin{pmatrix} 0 & * \\ 0 & * \end{pmatrix}$).

  When R is commutative, the left ideals of R are just the ideals of R, so the notion of R-submodules generalizes that of ideals of a commutative ring.

- Let R be any ring, and let $n \in \mathbb{N}$. Then,

$$R^n := \{(a_1, a_2, \ldots, a_n) \mid \text{all } a_i \text{ belong to } R\}$$

  is a left R-module, where addition and scaling are defined entrywise, i.e., by setting

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1,\ a_2 + b_2,\ \ldots,\ a_n + b_n)$$

  and

$$r\,(a_1, a_2, \ldots, a_n) = (ra_1, ra_2, \ldots, ra_n)\,.$$

  The zero vector of this left R-module is $(0, 0, \ldots, 0)$.

- Let R be any ring, and let $n, m \in \mathbb{N}$. Consider the set $R^{n\times m}$ of all $n \times m$-matrices with entries in R. This set $R^{n\times m}$ is not a ring unless $n = m$, but it is always a left R-module, where addition and action are defined entrywise. For instance, for 2×2-matrices, the action is given by

$$r \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ra & rb \\ rc & rd \end{pmatrix},$$

  and similarly for the other sizes. The zero vector is the zero matrix.

  The set $R^{n\times m}$ is also a right R-module in a similar way.

  According to our above definition, this allows us to refer to matrices as "vectors". This is a rather general notion of a vector that relies not on what a vector is but on what we can do with it (add and scale and take the zero).

- Just as we defined the left R-module R^n (consisting of n-tuples) for any $n \in \mathbb{N}$, we can define a left R-module "R^∞" consisting of all infinite sequences of elements of R. The proper name of this module is $R^{\mathbb{N}}$. Explicitly, $R^{\mathbb{N}}$ is defined to be the left R-module

$$\{(a_0, a_1, a_2, \ldots) \mid \text{all } a_i \text{ belong to } R\},$$

  whose addition and action are defined entrywise.

  This left R-module $R^{\mathbb{N}}$ has an R-submodule

$$R^{(\mathbb{N})} = \left\{(a_0, a_1, a_2, \ldots) \in R^{\mathbb{N}} \mid \text{only finitely many } i \in \mathbb{N} \text{ satisfy } a_i \neq 0\right\}.$$

  You can check that this is indeed an R-submodule of $R^{\mathbb{N}}$. For instance, for $R = \mathbb{Q}$, we have

$$(1, 1, 1, \ldots) \in R^{\mathbb{N}} \setminus R^{(\mathbb{N})};$$

$$\left(3, 2, 0, 5, \underbrace{0, 0, 0, \ldots}_{\text{only zeroes here}}\right) \in R^{(\mathbb{N})};$$

$$\left(\underbrace{1, 0, 1, 0, 1, 0, 1, 0, \ldots}_{\text{1's and 0's taking turns}}\right) \in R^{\mathbb{N}} \setminus R^{(\mathbb{N})};$$

$$\left(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots\right) \in R^{\mathbb{N}} \setminus R^{(\mathbb{N})}.$$

Note that the zero vector of an R-module is uniquely determined by its addition, so we don't have to provide it.

2.1.4. Direct products

Fix a ring R.

Most of our above examples of R-modules involve tuples on which addition and action work entrywise. There is a general concept for this:

Definition 2.1.5. Let $n \in \mathbb{N}$, and let $M_1, M_2, \ldots, M_n$ be any n left R-modules. Then, the Cartesian product $M_1 \times M_2 \times \cdots \times M_n$ becomes a left R-module as well, where addition and action are defined entrywise: e.g., the action is given by

$$r \cdot (m_1, m_2, \ldots, m_n) = (rm_1, rm_2, \ldots, rm_n).$$

This left R-module $M_1 \times M_2 \times \cdots \times M_n$ is called the **direct product** of $M_1, M_2, \ldots, M_n$.

More generally, we can define direct products of arbitrary families of R-modules, just as for rings. The resulting products are called $\prod_{i \in I} M_i$. (See §3.3.1 in the notes for the details.)

A particular case is of particular importance:

Definition 2.1.6. Let M be any left R-module. Let $n \in \mathbb{N}$. Then, we set

$$M^n := \underbrace{M \times M \times \cdots \times M}_{n \text{ times}}.$$

In particular, for $M = R$, we recover $M^n = R^n$ from the above examples.

2.1.5. Restriction of scalars

In linear algebra, you often consider $\mathbb{R}$-vector spaces and $\mathbb{C}$-vector spaces. One thing that you probably have seen is this: Any $\mathbb{C}$-vector space V becomes an $\mathbb{R}$-vector space if we simply forget how to scale by non-real numbers. This is called **restriction**, since we are just restricting the action $\mathbb{C} \times V \to V$ to a map $\mathbb{R} \times V \to V$.

Generalizing this idea to arbitrary modules gives the following operation:

- If R is a subring of a ring S, then any left S-module M becomes a left R-module, by restricting the action $S \times M \to M$ to a map $R \times M \to M$. Roughly speaking, we simply forget how to scale by scalars in $S \setminus R$.

  In particular, S itself (being a left S-module) becomes a left R-module. In this case, the action of R on S is just a restriction of the multiplication of S.

- More generally, if R and S are any two rings, and if $f : R \to S$ is a ring morphism, then any left S-module M becomes a left R-module, where the action is given by

$$r \cdot m = f(r) \cdot m \qquad \text{for all } r \in R \text{ and } m \in M.$$

  This is called **restriction of scalars**.

  In particular, S itself becomes a left R-module (since S is a left S-module, by multiplication). Some examples:

  - Any quotient ring R/I of a ring R by some ideal I becomes a left R-module, because the canonical projection $\pi : R \to R/I$ (sending each r to $\overline{r} = r + I$) is a ring morphism. Explicitly, the action of R on R/I is given by

$$r \cdot \overline{u} = \pi(r) \cdot \overline{u} = \overline{r} \cdot \overline{u} = \overline{ru} \qquad \text{for all } r, u \in R.$$

  - Another, weirder particular case: I claim that the abelian group $\mathbb{Z}/5$ becomes a $\mathbb{Z}[i]$-module if we define the action by

$$(a + bi) \cdot m = \overline{a + 2b} \cdot m \qquad \text{for all } a + bi \in \mathbb{Z}[i] \text{ and } m \in \mathbb{Z}/5.$$

    To understand this properly, we notice that there is a ring morphism

$$\begin{aligned} f : \mathbb{Z}[i] &\to \mathbb{Z}/5, \\ a + bi &\mapsto \overline{a + 2b}, \end{aligned}$$

    which is a ring morphism because $\overline{2}^2 = -\overline{1}$ in $\mathbb{Z}/5$ (check this!). The above $\mathbb{Z}[i]$-module structure on $\mathbb{Z}/5$ is simply the one obtained by restriction of scalars from using this ring morphism f.

    There is also a second $\mathbb{Z}[i]$-module structure on $\mathbb{Z}/5$, given by

$$(a + bi) \cdot m = \overline{a - 2b} \cdot m \qquad \text{for all } a + bi \in \mathbb{Z}[i] \text{ and } m \in \mathbb{Z}/5.$$

    When we speak of the $\mathbb{Z}[i]$-module $\mathbb{Z}/5$, we must be specific which one we are talking about.

2.2. A couple generalities

Next, we shall show a few general properties of modules. Again, fix a ring R.

2.2.1. Negation and subtraction

Proposition 2.2.1. Let R be a ring. Let M be a left R-module. Then, $(-1)\,a = -a$ for each $a \in M$.

Proof. Let $a \in M$. Then, $1a = a$ (by the module axioms). Thus,

$$(-1)\,a + \underbrace{a}_{=1a} = (-1)\,a + 1a = \underbrace{((-1)+1)}_{=0}\,a \qquad \text{(by distributivity)}$$
$$= 0a = 0 \qquad \text{(by the module axioms)}\,.$$

In other words, $(-1)\,a$ is an additive inverse to a. But that just means it is $-a$. □

Proposition 2.2.2. Let R be a ring. Let M be a left R-module. Let $r \in R$ and $m \in M$. Then,

$$(-r)\,m = -(rm) = r\,(-m)$$

and

$$(-r)\,(-m) = rm.$$

Proof. LTTR. □

Proposition 2.2.3. Let R be a ring. Let M be a left R-module. Then, any R-submodule of M is a subgroup of the additive group $(M, +, 0)$.

Proof. Let N be an R-submodule of M. Then, N is closed under addition and under scaling and thus also under negation (since $(-1)\,a = -a$ shows that negation is the same as scaling by -1). Moreover, it contains zero. So it is a subgroup. □

Proposition 2.2.4. Let R be a ring. Let M be a left R-module. Then, an R-submodule of M is the same as a subgroup of the additive group $(M, +, 0)$ that is closed under scaling by every scalar $r \in R$.

Proof. Follows from the above. □

Proposition 2.2.5. Let R be a ring. Let M be a left R-module. Then, any R-submodule of M becomes a left R-module in its own right (just like a subring of a ring becomes a ring itself).

Proof. Follows from the above. □

We also have "distributivity laws for subtraction":

Proposition 2.2.6. Let R be a ring. Let M be a left R-module. Then:
(a) We have $(r-s)\,m = rm - sm$ for all $r, s \in R$ and $m \in M$.
(b) We have $r\,(m-n) = rm - rn$ for all $r \in R$ and $m, n \in M$.

Proof. LTTR. □

2.2.2. Finite sums

Finite sums $\sum_{s \in S} a_s$ of elements of an R-module are defined just as they are in a ring. Finite products, of course, cannot be defined, since there is no multiplication on an R-module. The generalized distributivity laws

$$(r_1 + r_2 + \cdots + r_n)\, a = r_1 a + r_2 a + \cdots + r_n a \qquad \text{and}$$
$$r\,(a_1 + a_2 + \cdots + a_n) = r a_1 + r a_2 + \cdots + r a_n$$

hold in every left R-module A.

Since the associativity axiom says $(rs)\, m = r\,(sm)$ for all $r, s \in R$ and $m \in M$, we can write both sides as rsm without parentheses.

2.2.3. Principal submodules

The simplest way to construct submodules of a module is the following:

Proposition 2.2.7. Let R be a ring. Let a be a central element of R. Let M be a left R-module. Then,

$$aM := \{am \mid m \in M\}$$

is an R-submodule of M.

In particular, $0M = \{0_M\}$ and $1M = M$ are R-submodules of M.

Proof. Closed under addition: $am + an = a\,(m + n)$.

Closed under scaling: $r \cdot am = ram = arm$ since a is central.

Contains zero: $0 = a0_M$. □

Every R-submodule N of M lies between $0M$ and $1M$, meaning that $0M \subseteq N \subseteq 1M$.

If $M = R$, then the above submodules aM are just the principal ideals aR.

2.3. Abelian groups as $\mathbb{Z}$-modules

Let us now understand $\mathbb{Z}$-modules in particular.

Recall how the product of two integers is defined: Multiplication of nonnegative integers is defined by

$$nm = \underbrace{m + m + \cdots + m}_{n \text{ times}}.$$

More precisely, this formula defines nm for all $m \in \mathbb{Z}$ and all $n \in \mathbb{N}$. Then, to define nm for negative n, we set

$$nm = -\underbrace{(m + m + \cdots + m)}_{-n \text{ times}}.$$

So, altogether, nm is defined by

$$nm = \begin{cases} \underbrace{m+m+\cdots+m}_{n \text{ times}}, & \text{if } n \geq 0; \\ -\underbrace{(m+m+\cdots+m)}_{-n \text{ times}}, & \text{if } n < 0. \end{cases}$$

The same definition can be adapted to any abelian group:

Proposition 2.3.1. Let A be an abelian group, written additively (i.e., the operation on A is denoted by $+$, and the neutral element by 0). Then, for any $n \in \mathbb{Z}$ and $a \in A$, we define

$$na = \begin{cases} \underbrace{a+a+\cdots+a}_{n \text{ times}}, & \text{if } n \geq 0; \\ -\underbrace{(a+a+\cdots+a)}_{-n \text{ times}}, & \text{if } n < 0. \end{cases}$$

Thus, we have defined a map

$$\begin{aligned} \mathbb{Z} \times A &\to A, \\ (n, a) &\mapsto na. \end{aligned}$$

We shall refer to this map as the **action of $\mathbb{Z}$ by repeated addition**.

(a) The group A becomes a $\mathbb{Z}$-module, where we take this map as the action of $\mathbb{Z}$ on A.

(b) This is the **only** $\mathbb{Z}$-module structure on A. That is, if A is **any** $\mathbb{Z}$-module, then the action of $\mathbb{Z}$ on A is given by the above formula for na (and therefore uniquely determined by the abelian group structure on A).

(c) The $\mathbb{Z}$-submodules of A are precisely the subgroups of A.

Proof. See the text (§3.4). □

The proposition reveals what $\mathbb{Z}$-modules really are: They are just abelian groups with a more convenient "user interface". The "scaling by repeated addition" structure is inherent in the group, and by making the group into a $\mathbb{Z}$-module, you are "exposing" it for easier use.

In contrast, for a typical ring R, the R-modules have much more structure than the underlying abelian groups. The R-action on an R-module M is rarely ever determined by the addition on M. That this happens for $R = \mathbb{Z}$ is an exception.

That said, $\mathbb{Z}$ is not the only exception. Another case where the R-module structure is uniquely determined by the addition is the case $R = \mathbb{Q}$. The $\mathbb{Q}$-modules are also known as the $\mathbb{Q}$-vector spaces (since $\mathbb{Q}$ is a field), and again

the action of $\mathbb{Q}$ on such a module is uniquely determined by its addition: If a is a vector in a $\mathbb{Q}$-module M, and if $\dfrac{n}{m}$ is a rational number, then

$$\frac{n}{m} \cdot a = \text{the unique vector } b \text{ such that } mb = na.$$

Thus, any abelian group becomes a $\mathbb{Q}$-module in at most one way (there is no freedom in choosing the action). However, not every abelian group can be made into a $\mathbb{Q}$-module in the first place! For example, $\mathbb{Z}/2$ cannot be made into a $\mathbb{Q}$-module, because if it did, then

$$\frac{1}{2} \cdot \underbrace{\left(2 \cdot \overline{1}\right)}_{=\overline{0}} = \frac{1}{2} \cdot \overline{0} = \overline{0}$$

would equal

$$\frac{1}{2} \cdot \left(2 \cdot \overline{1}\right) = \underbrace{\left(\frac{1}{2} \cdot 2\right)}_{=1} \cdot \overline{1} = 1 \cdot \overline{1} = \overline{1}.$$

Thus, we see that

- any abelian group can be turned into a $\mathbb{Z}$-module, and in a unique way;
- not every abelian group can be turned into a $\mathbb{Q}$-module, but when it can be, this is also unique.

There is actually a characterization of abelian groups that can be turned into $\mathbb{Q}$-modules: they are the **uniquely divisible** abelian groups, i.e., those abelian groups A such that for each positive integer n and each $a \in A$, there is a unique $b \in A$ such that $a = nb$.

What about $\mathbb{R}$-modules (aka $\mathbb{R}$-vector spaces)? Again, not every abelian group can be made into an $\mathbb{R}$-module (for instance, $\mathbb{Q}$ is not an $\mathbb{R}$-module). Moreover, uniqueness is not a given either: In fact, the action of $\mathbb{R}$ on an $\mathbb{R}$-module is never uniquely determined by the addition (unless the $\mathbb{R}$-module is trivial, i.e., just contains a single vector). Likewise, the action of $\mathbb{Z}[i]$ on a $\mathbb{Z}[i]$-module is not uniquely determined by the addition (as we already saw above).

2.4. Module morphisms

2.4.1. Definition

Ring morphisms are maps between rings that respect the defining features of a ring $(+, \cdot, 0, 1)$.

Module morphisms play a similar role for modules:

Definition 2.4.1. Let R be a ring. Let M and N be two left R-modules.

(a) A **left R-module morphism** (aka **left R-linear map**) from M to N means a map $f : M \to N$ tha

- **respects addition** – i.e., satisfies $f(a+b) = f(a) + f(b)$ for all $a, b \in M$;
- **respects scaling** – i.e., satisfies $f(ra) = rf(a)$ for all $r \in R$ and $a \in M$;
- **respects the zero** – i.e., satisfies $f(0_M) = 0_N$.

We can drop the word "left" and just say "**R-linear map**" or "**R-module morphism**".

(b) A **left R-module isomorphism** from M to N means an invertible R-module morphism $f : M \to N$ whose inverse $f^{-1} : N \to M$ is also an R-module morphism. (The latter part is actually redundant, as we will soon see.)

(c) The left R-modules M and N are said to be **isomorphic** if there exists a left R-module isomorphism $f : M \to N$. In this case, we write $M \cong N$.

(d) Everything is defined analogously for right R-modules.

2.4.2. Simple examples

Here are some examples of R-module morphisms:

- When F is a field, the F-module morphisms are just the F-linear maps you know from linear algebra.
- Let $k \in \mathbb{Z}$. The map $\mathbb{Z} \to \mathbb{Z}$, $a \mapsto ka$ is a $\mathbb{Z}$-module morphism.
- More generally: Let R be a ring. Let k be a **central** element of R. Let M be any left R-module. Then, the map

$$\begin{aligned} M &\to M, \\ a &\mapsto ka \end{aligned}$$

  is a left R-module morphism. (We note that k needs to be central in order for this map to respect scaling.)

- Let R be a ring. Let $n \in \mathbb{N}$. For any $i \in \{1, 2, \ldots, n\}$, the map

$$\begin{aligned} \pi_i : R^n &\to R, \\ (a_1, a_2, \ldots, a_n) &\mapsto a_i \end{aligned}$$

  (which sends each n-tuple to its i-th entry) is a left R-module morphism.

Similar things hold for direct products more generally: Let $M_1, M_2, \ldots, M_n$ be any n left R-modules. Then, for any $i \in \{1, 2, \ldots, n\}$, the map

$$\begin{aligned} \pi_i : M_1 \times M_2 \times \cdots \times M_n &\to M_i, \\ (a_1, a_2, \ldots, a_n) &\mapsto a_i \end{aligned}$$

is a left R-module morphism.

- If M and N are two left R-modules, then the map

$$\begin{aligned} M \times N &\to N \times M, \\ (m, n) &\mapsto (n, m) \end{aligned}$$

is a left R-module isomorphism.

The $\mathbb{Z}$-module morphisms (i.e., the $\mathbb{Z}$-linear maps) are just the morphisms of abelian groups:

Proposition 2.4.2. Let M and N be two $\mathbb{Z}$-modules. Then, the $\mathbb{Z}$-module morphisms from M to N are precisely the group morphisms from $(M, +, 0)$ to $(N, +, 0)$.

Proof. Easy. □

2.4.3. Ring morphisms as module morphisms

Here is one more source of R-module morphisms:

- Let R and S be two rings. Let $f : R \to S$ be a ring morphism.

  As we observed a couple lectures ago, this morphism f makes S into a left R-module by the rule

$$rs = f(r) \cdot s \qquad \text{for all } r \in R \text{ and } s \in S$$

  ("restriction of scalars").

  It is now easy to see that f is a left R-module morphism from R to S. For instance, it respects scaling because

$$f(ra) = rf(a) \qquad \text{for all } r \in R \text{ and } a \in R$$

  (since f is a ring morphism, so $f(ra) = f(r) \cdot f(a) = rf(a)$ by the definition of the action of R on S).

  Here is a specific example: There is a ring morphism

$$\begin{aligned} &f : \mathbb{C} \to \mathbb{C}, \\ &a + bi \mapsto a - bi \qquad (\text{for all } a, b \in \mathbb{R}). \end{aligned}$$

This morphism f is called **complex conjugation** (geometrically, it flips the plane upside down, i.e., reflects it across the x-axis); the image $f(z)$ of a $z \in \mathbb{C}$ is called $\overline{z}$.

Obviously, $\mathbb{C}$ is a $\mathbb{C}$-module, by multiplication. However, we can define a second $\mathbb{C}$-module structure on $\mathbb{C}$, which is given by restriction of scalars via the morphism $f : \mathbb{C} \to \mathbb{C}$. This second structure is explicitly given by

$$r \rightharpoonup s = \underbrace{f(r)}_{=\overline{r}} \cdot s = \overline{r} \cdot s \qquad \text{for any } r, s \in \mathbb{C},$$

where $\rightharpoonup$ is the symbol for the new action (the addition is the same as in the usual $\mathbb{C}$).

Let me denote this new $\mathbb{C}$-module by $\overline{\mathbb{C}}$. The map $f : \mathbb{C} \to \mathbb{C}$ is not $\mathbb{C}$-linear as a map from $\mathbb{C}$ to $\mathbb{C}$, but it is $\mathbb{C}$-linear as a map from $\mathbb{C}$ to $\overline{\mathbb{C}}$ (or vice versa).

We can play this game more generally: If M is any $\mathbb{C}$-module (= $\mathbb{C}$-vector space), then we can define a second $\mathbb{C}$-module structure on M by restricting the $\mathbb{C}$-module M via the complex conjugation map f. We call this second $\mathbb{C}$-module $\overline{M}$. As an additive group, it is just M, but its action is given by

$$r \rightharpoonup s = \overline{r} \cdot s \qquad \text{for any } r, s \in \mathbb{C}.$$

You can think of $\overline{M}$ as a "mirror version" of M; it consists of the same vectors as M but "sees the scalars through the looking glass".

If M and N are two $\mathbb{C}$-modules, then a map $g : M \to N$ is said to be **antilinear** (or **conjugate-linear**) if it is a $\mathbb{C}$-linear map from M to $\overline{N}$. Explicitly, this means that g has the following properties:

$$\begin{aligned} g(a+b) &= g(a) + g(b) && \text{for all } a, b \in M; \\ g(ra) &= \overline{r} g(a) && \text{for all } r \in \mathbb{C} \text{ and } a \in M; \\ g(0) &= 0. \end{aligned}$$

For example, the complex conjugation map f is an antilinear map from $\mathbb{C}$ to $\mathbb{C}$ (or a linear map from $\mathbb{C}$ to $\overline{\mathbb{C}}$).

For instance, the Hermitian dot product

$$\begin{aligned} \mathbb{C}^n \times \mathbb{C}^n &\to \mathbb{C}, \\ \left(\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \right) &\mapsto \sum_{k=1}^{n} \overline{v_k} w_k \end{aligned}$$

is antilinear in its first argument and linear in its second. Such maps are called **sesquilinear**, and include all Hermitian forms.

2.4.4. General properties of linearity

Fix a ring R. The following facts about R-module morphisms are analogues of known facts about group and ring morphisms:

Proposition 2.4.3. Any invertible left R-module morphism is a left R-module isomorphism.

Proposition 2.4.4. A composition of two left R-module morphisms is again a left R-module morphism.

Proposition 2.4.5. A composition of two left R-module isomorphisms is again a left R-module isomorphism.

Proposition 2.4.6. The inverse of a left R-module isomorphism is a left R-module isomorphism.

Proposition 2.4.7. The relation $\cong$ (between left R-modules) is an equivalence relation.

Again, there is an isomorphism principle: Any intrinsic property of an R-module M (i.e., any property that does not depend on what the elements of M "are") automatically holds for any R-module isomorphic to M.

Everything we say about left R-modules holds equally well for right R^{op}-modules, and thus for right R-modules if R is allowed to be arbitrary (since $R^{\mathrm{op}\,\mathrm{op}} = R$).

2.4.5. Kernels and images

Just like ring morphisms, module morphisms have kernels and images.

Again, we fix a ring R.

Definition 2.4.8. Let R be a ring. Let M and N be two left R-modules. Let $f : M \to N$ be a left R-module morphism. Then, the **kernel** (aka **nullspace**) of f (denoted $\operatorname{Ker} f$ or $\ker f$) is defined to be the subset

$$\operatorname{Ker} f := \{a \in M \mid f(a) = 0_N\}$$

of M.

Examples:

- Let R be a commutative ring. Let $b \in R$. Then, the map

$$\begin{aligned} R &\to R, \\ r &\mapsto br \end{aligned}$$

(multiplication by b) is an R-module morphism. Its kernel is

$$\{r \in R \mid br = 0\}.$$

If b is not zero and no zero-divisor, then this kernel is $\{0\}$.

- Both $\mathbb{Z}^3$ and $\mathbb{Z} \times (\mathbb{Z}/2)$ are abelian groups, thus $\mathbb{Z}$-modules. The map

$$\mathbb{Z}^3 \to \mathbb{Z} \times (\mathbb{Z}/2),$$
$$(a,b,c) \mapsto \left(a-b,\ \overline{b-c}\right)$$

is a $\mathbb{Z}$-module morphism. Its kernel is

$$\begin{aligned}
&\left\{(a,b,c) \in \mathbb{Z}^3 \mid \left(a-b,\ \overline{b-c}\right) = 0\right\} \\
&= \left\{(a,b,c) \in \mathbb{Z}^3 \mid a-b = 0 \text{ and } \overline{b-c} = 0\right\} \\
&= \left\{(a,b,c) \in \mathbb{Z}^3 \mid a = b \text{ and } b \equiv c \bmod 2\right\} \\
&= \{(a,a,a+2k) \mid a,k \in \mathbb{Z}\}.
\end{aligned}$$

Some basic facts from linear algebra still hold at the level of modules:

Theorem 2.4.9. Let R be a ring. Let M and N be two left R-modules. Let $f : M \to N$ be a left R-module morphism. Then:
(a) The kernel $\operatorname{Ker} f$ is an R-submodule of M.
(b) The image $\operatorname{Im} f = f(M)$ is an R-submodule of N.

Lemma 2.4.10. Let R be a ring. Let M and N be two left R-modules. Let $f : M \to N$ be a left R-module morphism. Then, f is injective if and only if $\operatorname{Ker} f = \{0_M\}$.

2.4.6. Quotient modules

Again, we fix a ring R.

Quotient modules are an analogue of quotient rings and quotient groups:

Definition 2.4.11. Let M be a left R-module. Let I be a left R-submodule of M. Thus, I is a subgroup of the additive group $(M,+,0)$, so we obtain a quotient group M/I, whose elements $a+I$ we shall write as $\overline{a}$ and call **residue classes** or **cosets**. Addition is given by

$$\overline{a} + \overline{b} = \overline{a+b} \qquad \text{for all } a,b \in M.$$

We make M/I into a left R-module by defining an action of R on M/I by setting

$$r\overline{a} = \overline{ra} \qquad \text{for all } r \in R \text{ and } m \in M.$$

This makes M/I into a left R-module with zero vector $\overline{0} = 0 + I$. It is called the **quotient R-module of M by I**, and is denoted M/I.

Theorem 2.4.12. This is indeed a left R-module. Moreover, the map

$$\begin{aligned} \pi : M &\to M/I, \\ a &\mapsto \overline{a} = a + I \end{aligned}$$

is a surjective R-module morphism. This map π is called the **canonical projection**.

Proof. Straightforward. □

Examples of quotient modules come from many places:

- Quotients of abelian groups (e.g., $\mathbb{Z}/n$ or $\mathbb{R}/\mathbb{Z} = S^1$ (a circle) or $\mathbb{R}^2/\mathbb{Z}^2 = T^2$ (a torus)) are just quotients of $\mathbb{Z}$-modules.
- Quotients of vector spaces are quotients of F-modules, where F is a field.

  For instance, consider the 3D vector space (i.e., $\mathbb{R}$-module) $\mathbb{R}^3$ over the ring $\mathbb{R}$. Typically, we view $\mathbb{R}^3$ as the usual geometric 3D space. Define a vector subspace (i.e., an $\mathbb{R}$-submodule) I of $\mathbb{R}^3$ by

  $$I = \left\{ (x,y,z) \in \mathbb{R}^3 \;\mid\; x+y+z = 0 \right\}.$$

  Geometrically, this is a plane through the origin.

  What can we say about the quotient $\mathbb{R}$-module $\mathbb{R}^3/I$? Its elements are residue classes of the form $\overline{(x,y,z)}$, where two vectors (x,y,z) and (x',y',z') belong to the same residue class if and only if their entry-wise difference $(x-x',\ y-y',\ z-z')$ belongs to I (that is, if and only if $(x-x')+(y-y')+(z-z') = 0$, or, equivalently, $x+y+z = x'+y'+z'$). For exmaple, the two residue classes $\overline{(3,0,0)}$ and $\overline{(1,1,1)}$ are identical. It is not hard to see that each element of $\mathbb{R}^3/I$ can be uniquely written in the form $\overline{(r,0,0)}$ for some $r \in \mathbb{R}$. This shows that $\mathbb{R}^3/I$ is 1-dimensional as a vector space.
- If R is any ring, and M is any left R-module, then the two obvious R-submodules $\{0_M\}$ and M lead to uninteresting quotient modules:

  $M/\{0_M\} \cong M$ and M/M is trivial (i.e., has only 1 element).
- Let R be a ring. As we recall, the left R-module

  $$R^{\mathbb{N}} = \{(a_0, a_1, a_2, \ldots) \;\mid\; \text{all } a_i \in R\}$$

  has an R-submodule

  $$R^{(\mathbb{N})} = \{(a_0, a_1, a_2, \ldots) \;\mid\; \text{all } a_i \in R, \text{ and almost all } a_i = 0\}$$

(where "almost all" means "all but finitely many", i.e., "we have $a_i = 0$ for all but finitely many i"). What is the quotient module $R^{\mathbb{N}}/R^{(\mathbb{N})}$? Its elements are residue classes of the form $\overline{(a_0, a_1, a_2, \ldots)}$, where two infinite sequences $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$ lie in the same residue class if and only if their entrywise difference $(a_0 - b_0,\ a_1 - b_1,\ a_2 - b_2,\ \ldots)$ lies in $R^{(\mathbb{N})}$ (that is, if all but finitely many i satisfy $a_i = b_i$). So we can view a residue class $\overline{(a_0, a_1, a_2, \ldots)}$ as "an infinite sequence that is defined up to finitely many places".

Such residue classes have a bunch of features known from analysis. In particular, if the limit $\lim\limits_{n\to\infty} a_n$ exists, then it depends only on the residue class $\overline{(a_0, a_1, a_2, \ldots)}$ rather than on the sequence $(a_0, a_1, a_2, \ldots)$.

For quotient rings, we have previously proved a universal property and a first isomorphism theorem. The same can be done for quotient modules. Let me just state the analogue of the universal property:

Theorem 2.4.13 (Universal property of quotient modules, elementwise form)**.** Let M be a left R-module. Let I be a left R-submodule of M.

Let N be a left R-module. Let $f : M \to N$ be a left R-module morphism. Assume that $f(I) = 0$ (that is, $f(i) = 0$ for each $i \in I$). Then, the map

$$\begin{aligned} f' : M/I &\to N, \\ \overline{a} &\mapsto f(a) \end{aligned}$$

is well-defined and is a left R-module morphism.

Proof. Analogous to the ring case. □

2.5. Spanning, linear independence, bases, free modules

Again, we fix a ring R.

2.5.1. Definitions

Definition 2.5.1. Let M be a left R-module. Let $m_1, m_2, \ldots, m_n$ be finitely many vectors in M. Then:

(a) A **linear combination** of $m_1, m_2, \ldots, m_n$ means a vector of the form

$$r_1 m_1 + r_2 m_2 + \cdots + r_n m_n \qquad \text{for all } r_1, r_2, \ldots, r_n \in R.$$

(b) The set of all linear combinations of $m_1, m_2, \ldots, m_n$ is called the **span** of $m_1, m_2, \ldots, m_n$, and is denoted by $\operatorname{span}(m_1, m_2, \ldots, m_n)$.

(c) If the span of $m_1, m_2, \ldots, m_n$ is M, then we say that the vectors $m_1, m_2, \ldots, m_n$ **span** M (or **generate** M).

(d) We say that the vectors $m_1, m_2, \ldots, m_n$ are **linearly independent** if the following holds: If $r_1, r_2, \ldots, r_n \in R$ satisfy

$$r_1 m_1 + r_2 m_2 + \cdots + r_n m_n = 0,$$

then $r_1 = r_2 = \cdots = r_n = 0$.

(e) We say that the n-tuple $(m_1, m_2, \ldots, m_n)$ is a **basis** of M if and only if the vectors $m_1, m_2, \ldots, m_n$ are linearly independent and span M.

(f) All of this terminology depends on R. If R is not clear from the context, you can make it explicit: say "R-linear combination", "R-span", and so on.

These features can be defined not just for a finite list $(m_1, m_2, \ldots, m_n)$ of vectors, but for any family $(m_i)_{i \in I}$ of vectors. There is one complication: We do not allow "truly infinite" linear combinations like $1m_0 + 1m_1 + 1m_2 + \cdots$. So, even if your family $(m_i)_{i \in I}$ is infinite (i.e., if the set I is infinite), we only allow linear combinations where all but finitely many coefficients are 0. For instance, with an infinite sequence of vectors $(m_0, m_1, m_2, \ldots)$, you can take linear combinations such as $m_3 + m_5 - 2m_7$ but not $0m_0 + 1m_1 + 2m_2 + \cdots$.

Likewise, linear independence for a family $(m_i)_{i \in I}$ is defined in terms of finite sums only.

Definition 2.5.2. Let M be a left R-module. Let $(m_i)_{i \in I}$ be a family of vectors in M (with I being any set). Then:

(a) A **linear combination** of $(m_i)_{i \in I}$ means a vector of the form

$$\sum_{i \in I} r_i m_i$$

for some family $(r_i)_{i \in I}$ of scalars (i.e., for a choice of $r_i \in R$ for each $i \in I$) with the property that

$$\text{all but finitely many } i \in I \text{ satisfy } r_i = 0.$$

Here, the sum $\sum_{i \in I} r_i m_i$ is an infinite sum, but all but finitely many of its addends are 0, and thus we can make sense of this sum simply by throwing away the 0 addends and summing the rest (for example, $3 + 2 + 0 + 0 + 0 + \cdots = 3 + 2$).

(b) The set of all linear combinations of $(m_i)_{i \in I}$ is called the **span** of $(m_i)_{i \in I}$, and is denoted by $\operatorname{span}(m_i)_{i \in I}$.

(c) If the span of $(m_i)_{i \in I}$ is M, then we say that the family $(m_i)_{i \in I}$ **spans** M (or **generates** M).

(d) We say that the family $(m_i)_{i \in I}$ is **linearly independent** if the following holds: If some family $(r_i)_{i \in I}$ of scalars $r_i \in R$ satisfies

$$\text{all but finitely many } i \in I \text{ satisfy } r_i = 0$$

and

$$\sum_{i \in I} r_i m_i = 0,$$

then $r_i = 0$ for all $i \in I$.

(e) We say that the family $(m_i)_{i \in I}$ is a **basis** of M if and only if it is linearly independent and spans M.

(f) All of this terminology depends on R. If R is not clear from the context, you can make it explicit: say "R-linear combination", "R-span", and so on.

2.5.2. Spans are submodules

As in linear algebra, we can generate submodules of a given module M by taking spans of vectors in M:

Proposition 2.5.3. Let M be a left R-module. Let $(m_i)_{i \in I}$ be a family of vectors in M. Then, the span of this family is an R-submodule of M.

Proof. "Closed under addition":

$$\sum_{i \in I} a_i m_i + \sum_{i \in I} b_i m_i = \sum_{i \in I} (a_i + b_i) m_i.$$

(You have to check finiteness, but that's simply because a union of two finite sets is finite. See 2023 Lecture 20 for details.)

The other axioms are similar. □

2.5.3. Coordinates

The notions of linear independence and spanning can be described equivalently as follows:

Proposition 2.5.4. Let M be a left R-module. Let $(m_i)_{i \in I}$ be a family of vectors in M. Then:

(a) The family $(m_i)_{i \in I}$ spans M if and only if each vector $v \in M$ can be written as an R-linear combination of $(m_i)_{i \in I}$ in **at least one** way.

(b) The family $(m_i)_{i \in I}$ is linearly independent if and only if each vector $v \in M$ can be written as an R-linear combination of $(m_i)_{i \in I}$ in **at most one** way (i.e., there is **at most one** family $(r_i)_{i \in I}$ of scalars such that $v = \sum_{i \in I} r_i m_i$ and such that all but finitely many $i \in I$ satisfy $r_i = 0$).

(c) The family $(m_i)_{i \in I}$ is a basis of M if and only if each vector $v \in M$ can be written as an R-linear combination of $(m_i)_{i \in I}$ in **exactly one** way (i.e., there is **exactly one** family $(r_i)_{i \in I}$ of scalars such that $v = \sum_{i \in I} r_i m_i$ and such that all but finitely many $i \in I$ satisfy $r_i = 0$).

Part **(c)** of this proposition shows that a basis of an R-module M can be used as a "coordinate system" for M. The scalars r_i that represent v as $v = \sum_{i \in I} r_i m_i$ are called the **coordinates** of v with respect to this basis.

For the proof of the proposition, see Winter 2023 Lecture 20.

2.5.4. Free modules

In linear algebra, there is a celebrated result saying:

Theorem 2.5.5. If F is a field, then every F-module (i.e., every F-vector space) has a basis.

Proof. See [Treil] or [Keith Conrad, dimension.pdf]. The case when the F-module is finitely generated (i.e., spanned by a finite list of vectors) is relatively easy and done in most linear algebra texts. Note that the general version is not constructive at all. For instance, it yields that $\mathbb{R}$ has a basis as a $\mathbb{Q}$-vector space, but no one can actually construct such a basis. (This is known as a **Hamel basis**.) □

In comparison, not every R-module over a ring R has a basis. Modules with bases are rather rare and have their own name:

Definition 2.5.6. (a) A left R-module M is said to be **free** if it has a basis.

(b) Let $n \in \mathbb{N}$. A left R-module M is said to be **free of rank** n if it has a basis of size n (that is, a basis consisting of n vectors).

Note that not every free R-module has a rank in this sense, since its basis could be infinite.

Confusing curiosity: A free R-module can have several ranks at the same time. This happens for some (rather weird) noncommutative rings R, and also for the trivial ring R. It never happens when R is a nontrivial commutative ring, but the proof is not quite easy.

Surprisingly, even though not all R-modules are free in general, free R-modules appear over and over in mathematics. Let us give some examples.

We begin with examples that make sense for any ring R. We fix an arbitrary ring R.

- Consider the left R-module

$$R^2 = \{(a, b) \mid a, b \in R\}.$$

This R-module R^2 is free of rank 2, since the list

$$((1,0), (0,1))$$

is a basis of R^2. This is a basis because:

  - It spans R^2, since each $(a,b) \in R^2$ is $a \cdot (1,0) + b \cdot (0,1)$.
  - It is linearly independent, since if $a \cdot (1,0) + b \cdot (0,1) = 0$, then $0 = a \cdot (1,0) + b \cdot (0,1) = (a,b)$, so $a = b = 0$.

- Likewise, the left R-module R^3 has basis

$$((1,0,0),\ (0,1,0),\ (0,0,1)).$$

- More generally: For any $n \in \mathbb{N}$, the left R-module R^n has a basis

$$(e_1, e_2, \ldots, e_n),$$

  where e_i is the n-tuple $(0,0,\ldots,0,1,0,0,\ldots,0)$ with the 1 at the i-th position. This basis is called the **standard basis** of R^n. So R^n is free of rank n.

- In particular, R^1 is free of rank 1. Since $R^1 \cong R$ as left R-modules, this shows that R is free of rank 1.

  Also, R^0 is free of rank 0.

- More generally: If I is a set, then the set

$$R^I = \prod_{i \in I} R = \left\{ (r_i)_{i \in I} \ \mid \ \text{all } r_i \text{ belong to } R \right\}$$

  is a left R-module (with entrywise addition and action). If I is finite, then this R-module is free (indeed, if I is an n-element set, then $R^I \cong R^n$). If I is infinite, then R^I is usually not free. For instance, the $\mathbb{Z}$-module

$$\mathbb{Z}^{\mathbb{N}} = \{\text{all infinite sequences of integers}\}$$

  is not free. (This is not obvious, but can be proved.) Of course, when R is a field, then R^I is free, by the above theorem saying that any vector space is free.

  However, the left R-module R^I has a very important submodule that is always free. Namely, we define

$$R^{(I)} = \left\{ (r_i)_{i \in I} \in R^I \ \mid \ \text{all but finitely many } i \in I \text{ satisfy } r_i = 0 \right\},$$

  a subset of R^I. This subset $R^{(I)}$ is a left R-submodule of R^I (easy to check), and is actually free, with a **standard basis** $(e_i)_{i \in I}$, where e_i is the family with a 1 in its i-th position and 0's everywhere else (a generalization of the standard basis $(e_1, e_2, \ldots, e_n)$ of R^n).

- Let $n, m \in \mathbb{N}$. Then, the set $R^{n \times m}$ of $n \times m$-matrices is a left R-module. It is free of rank nm, and in fact it has a basis $\left(E_{i,j}\right)_{(i,j) \in \{1,2,\ldots,n\} \times \{1,2,\ldots,m\}}$ consisting of the **elementary matrices** $E_{i,j}$. Each $E_{i,j}$ is the $n \times m$-matrix which has a 1 in its (i,j)-th cell and 0's in all other cells.

- Let $n \in \mathbb{N}$. Then, the set $R^{n\times n}_{\text{symm}}$ of all symmetric $n \times n$-matrices is a left R-module. It is free of rank $\dfrac{n\left(n+1\right)}{2}$, with a basis consisting of the diagonal elementary matrices $E_{i,i}$ and the symmetrized off-diagonal elementary matrices $E_{i,j} + E_{j,i}$ for $i < j$.

Let us now look at $\mathbb{Z}$-modules. As we know, they are just abelian groups in fancy clothes, but let us see which of them are free (as $\mathbb{Z}$-modules):

- Consider the $\mathbb{Z}$-submodule

$$U := \left\{(a,b,c) \in \mathbb{Z}^3 \ \mid \ a+b+c = 0\right\} \text{ of } \mathbb{Z}^3.$$

  Is U free? Can we find a basis for U ? Yes:

$$\left((-1,1,0)\,, \quad (-1,0,1)\right).$$

  What about more general versions of U ? So subsets of $\mathbb{Z}^n$ carved out by (systems of) linear equations?

  In this example, we can find this basis by Gaussian elimination, as in linear algebra. But in more general situations, Gaussian elimination can fail, since it can incur denominators (and thus take us out of $\mathbb{Z}$).

  Nevertheless, it can be shown that any $\mathbb{Z}$-submodule of $\mathbb{Z}^k$ (for $k \in \mathbb{N}$) is free. More on this later.

- The $\mathbb{Z}$-module $\mathbb{Z}/2$ is not free. More generally: Any free $\mathbb{Z}$-module is either trivial or infinite. So $\mathbb{Z}/n$ is only free as a $\mathbb{Z}$-module if n is 1, 0 or -1.

- The $\mathbb{Z}$-module $\mathbb{Q}$ is not free. In a nutshell, this is because one vector is not enough to span $\mathbb{Q}$, but two vectors are already linearly dependent.

- Consider the $\mathbb{Z}$-submodule

$$V := \left\{(a,b) \in \mathbb{Z}^2 \ \mid \ a \equiv b \bmod 2\right\} \text{ of } \mathbb{Z}^2.$$

  This $\mathbb{Z}$-module is free. Can you find a basis?

$$\begin{aligned} &\left((1,1)\,, \ \ (2,0)\right) \text{ is a basis;} \\ &\left((1,1)\,, \ \ (1,-1)\right) \text{ is a basis.} \end{aligned}$$

  Linear independence can be checked over $\mathbb{Q}$ (why?). Spanning for $\left((1,1)\,, \ \ (2,0)\right)$ can be checked by observing that each $(a,b) \in V$ satisfies

$$(a,b) = b\,(1,1) + \underbrace{\frac{a-b}{2}}_{\substack{\in\mathbb{Z} \\ \text{since } a\equiv b \bmod 2}} (2,0) = \left(b + 2\cdot\frac{a-b}{2},\ b\right).$$

- More examples: see §3.7.3 in the text.

Some more generalities about free modules:

Theorem 2.5.7. Let M be a left R-module. Let $n \in \mathbb{N}$. Then, the left R-module M is free of rank n if and only if $M \cong R^n$ as left R-modules.

More concretely, any basis of M gives an isomorphism $f : R^n \to M$:

Theorem 2.5.8. Let M be a left R-module. Let $m_1, m_2, \ldots, m_n$ be n vectors in M. Consider the map

$$\begin{aligned} f : R^n &\to M, \\ (r_1, r_2, \ldots, r_n) &\mapsto r_1 m_1 + r_2 m_2 + \cdots + r_n m_n. \end{aligned}$$

Then:

(a) This map f is always a left R-module morphism.

(b) This map f is injective if and only if the vectors $m_1, m_2, \ldots, m_n$ are R-linearly independent.

(c) This map f is surjective if and only if the vectors $m_1, m_2, \ldots, m_n$ span M.

(d) This map f is bijective (i.e., an isomorphism) if and only if $(m_1, m_2, \ldots, m_n)$ is a basis of M.

This can be generalized from R^n to $R^{(I)}$ for arbitrary sets I:

Theorem 2.5.9. Let M be a left R-module. Let $(m_i)_{i \in I}$ be a family of vectors in I. Consider the map

$$\begin{aligned} f : R^{(I)} &\to M, \\ (r_i)_{i \in I} &\mapsto \sum_{i \in I} r_i m_i. \end{aligned}$$

Then:

(a) This map f is always a left R-module morphism.

(b) This map f is injective if and only if the family $(m_i)_{i \in I}$ is R-linearly independent.

(c) This map f is surjective if and only if the family $(m_i)_{i \in I}$ spans M.

(d) This map f is bijective (i.e., an isomorphism) if and only if $(m_i)_{i \in I}$ is a basis of M.

2.6. The universal property of a free module

As before, fix a ring R.

Recall that R-linear maps (= R-module morphisms) respect addition, scaling and zero. Thus, they respect any linear combinations:

Proposition 2.6.1. Let M and P be two left R-modules. Let $f : M \to P$ be an R-linear map. Let $(m_i)_{i \in I}$ be a family of vectors in M, and let $(r_i)_{i \in I} \in R^{(I)}$ be a family of scalars. Then,

$$f\left(\sum_{i \in I} r_i m_i\right) = \sum_{i \in I} r_i f\left(m_i\right).$$

Proof. If I is finite, induct on $|I|$. If I is infinite, throw away the zeroes. □

Now, we shall state the **universal property of free modules**. This property gives an easy way to construct a linear map f from a free R-module M: It just says that we can specify the values $f\left(m_i\right)$ of f on a given basis $(m_i)_{i \in I}$ of M, and then all other values of f are uniquely determined. In formal words:

Theorem 2.6.2 (Universal property of free modules)**.** Let M be a free left R-module with basis $(m_i)_{i \in I}$. Let P be a further left R-module (free or not). Let $p_i \in P$ be a vector for each $i \in I$. Then, there exists a **unique** R-linear map $f : M \to P$ such that

$$\text{each } i \in I \text{ satisfies } f\left(m_i\right) = p_i.$$

Explicitly, this map is given by

$$f\left(\sum_{i \in I} r_i m_i\right) = \sum_{i \in I} r_i p_i \qquad \text{for all } (r_i)_{i \in I} \in R^{(I)}.$$

Proof. See the notes. (Straightforward.) (Lecture 21 in Winter 2023) □

The uniqueness part of the universal property (i.e., the part claiming that f is unique) is true under a weaker assumption: It suffices that $(m_i)_{i \in I}$ spans M; we don't need it to be a basis for that. So we get the following fact:

Theorem 2.6.3 (Linear maps are determined on a spanning set)**.** Let M be a left R-module. Let $(m_i)_{i \in I}$ be a family of vectors in M that spans M. Let $f, g : M \to P$ be two R-linear maps from M to a further R-module P such that

$$\text{each } i \in I \text{ satisfies } f\left(m_i\right) = g\left(m_i\right).$$

Then, $f = g$.

Proof. See the notes. (Straightforward.) (Lecture 21 in Winter 2023) □

The universal property of free R-modules is why we can represent linear maps between free R-modules as matrices. So if you have a free left R-module M with basis $(m_1, m_2, \ldots, m_n)$ and a free left R-module M' with basis $\left(m'_1, m'_2, \ldots, m'_{n'}\right)$, then a linear map $f : M \to M'$ can be represented by the $n' \times n$-matrix whose j-th column is given by the coordinates of $f\left(m_j\right)$ with respect to the basis $\left(m'_1, m'_2, \ldots, m'_{n'}\right)$ of M'.

2.7. Bilinear maps

Let R be a commutative ring.

The addition map

$$\begin{aligned} \text{add} : R \times R &\to R, \\ (a, b) &\mapsto a + b \end{aligned}$$

is R-linear (where the domain is the direct product of two copies of R). But the multiplication map

$$\begin{aligned} \text{mul} : R \times R &\to R, \\ (a, b) &\mapsto ab \end{aligned}$$

is not. Nevertheless, it has "some linearity" in it: Namely, if we fix one argument, then mul is linear in the other. That is:

- For any $a \in R$, the map

$$\begin{aligned} R &\to R, \\ b &\mapsto ab \end{aligned}$$

  is R-linear.

- For any $b \in R$, the map

$$\begin{aligned} R &\to R, \\ a &\mapsto ab \end{aligned}$$

  is R-linear.

Maps with these properties are called **bilinear**:

Definition 2.7.1. Let R be a commutative ring. Let M, N and P be three R-modules. A map $f : M \times N \to P$ is said to be R-**bilinear** (or just **bilinear**) if it satisfies the following two conditions:

1. For any $n \in N$, the map

$$\begin{aligned} f : M &\to P, \\ m &\mapsto f(m, n) \end{aligned}$$

is R-linear. Explicitly, this is saying that for any $n \in N$, we have

$$\begin{aligned} f(m_1 + m_2,\ n) &= f(m_1, n) + f(m_2, n) \qquad \text{for all } m_1, m_2 \in M; \\ f(rm, n) &= rf(m, n) \qquad \text{for all } r \in R \text{ and } m \in M; \\ f(0, n) &= 0. \end{aligned}$$

This is called "f is **linear in its first argument**".

2. For any $m \in M$, the map

$$\begin{aligned} f : N &\to P, \\ n &\mapsto f(m, n) \end{aligned}$$

is R-linear. Explicitly, this is saying that for any $m \in M$, we have

$$\begin{aligned} f(m,\ n_1 + n_2) &= f(m, n_1) + f(m, n_2) \qquad \text{for all } n_1, n_2 \in N; \\ f(m, rn) &= rf(m, n) \qquad \text{for all } r \in R \text{ and } n \in N; \\ f(m, 0) &= 0. \end{aligned}$$

This is called "f is **linear in its second argument**".

Here are some examples of R-bilinear maps:

- The multiplication map $\text{mul} : R \times R \to R$ is bilinear.
- For any $n \in \mathbb{N}$, the map

$$\begin{aligned} R^n \times R^n &\to R, \\ ((a_1, a_2, \ldots, a_n),\ (b_1, b_2, \ldots, b_n)) &\mapsto a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \end{aligned}$$

– known as the **dot product** or the **standard scalar product** – is bilinear.

- For any $n \in \mathbb{N}$, the map

$$\begin{aligned} \mathbb{C}^n \times \mathbb{C}^n &\to \mathbb{C}, \\ ((a_1, a_2, \ldots, a_n),\ (b_1, b_2, \ldots, b_n)) &\mapsto a_1 \overline{b_1} + a_2 \overline{b_2} + \cdots + a_n \overline{b_n} \end{aligned}$$

(where $\overline{z}$ denotes the complex conjugate of z) is not bilinear as a map from $\mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$, but it is bilinear as a map from $\mathbb{C}^n \times \overline{\mathbb{C}}^n \to \mathbb{C}$ (remember: $\overline{M}$ is M with the scaling twisted by conjugation, meaning that $z \rightharpoonup v = \overline{z} \cdot v$). Such maps are called **sesquilinear**.

- The cross product map
$$R^3 \times R^3 \to R^3,$$
$$\left((a,b,c)\,,\ \left(a',b',c'\right)\right) \mapsto \left(bc'-cb',\ ca'-ac',\ ab'-ba'\right)$$
is bilinear.
- The determinant map
$$\det : R^2 \times R^2 \to R,$$
$$((a,b)\,,\ (c,d)) \mapsto ad-bc$$
is R-bilinear.
- The Hadamard product
$$R^{n\times m} \times R^{n\times m} \to R^{n\times m},$$
$$\left(\left(a_{i,j}\right)_{i,j}\,,\ \left(b_{i,j}\right)_{i,j}\right) \mapsto \left(a_{i,j}b_{i,j}\right)_{i,j}$$
is R-bilinear.
- For any R-module M, the action
$$R \times M \to M,$$
$$(r,m) \mapsto rm$$
is an R-bilinear map. (Note that commutativity of R is needed.)

Recall the universal property of free modules we proved above. That property allows us to define a linear map from a free module by specifying its images on the basis vectors. The same can be done for bilinear maps:

Theorem 2.7.2 (Universal property of free modules wrt bilinear maps)**.** Let M be a free left R-module with basis $(m_i)_{i\in I}$. Let N be a free left R-module with basis $\left(n_j\right)_{j\in J}$. Let P be a further left R-module (free or not). Let $p_{i,j} \in P$ be a vector for each pair $(i,j) \in I \times J$. Then, there exists a **unique** R-bilinear map $f : M \times N \to P$ such that

$$\text{each } i \in I \text{ satisfies } f\left(m_i, n_j\right) = p_{i,j}.$$

Explicitly, this map is given by

$$f\left(\sum_{i\in I} r_i m_i,\ \sum_{j\in J} s_j n_j\right) = \sum_{(i,j)\in I\times J} r_i s_j p_{i,j} \qquad \text{for all } \left(r_i\right)_{i\in I}, \left(s_j\right)_{j\in J}.$$

Proof. See the notes (Theorem 3.9.2). □

2.8. Multilinear maps

Multilinear maps are a generalization of linear and bilinear maps:

Definition 2.8.1. Let R be a commutative ring. Let $M_1, M_2, \ldots, M_n$ be finitely many R-modules. Let P be any R-module. A map $f : M_1 \times M_2 \times \cdots \times M_n \to P$ is said to be **R-multilinear** (or just **multilinear)** if it satisfies the following condition:

- For any $i \in \{1, 2, \ldots, n\}$ and any $m_1, m_2, \ldots, m_{i-1}, m_{i+1}, \ldots, m_n$ in the respective modules (meaning that $m_k \in M_k$ for each $k \neq i$), the map

$$\begin{aligned} M_i &\to P, \\ m_i &\mapsto f(m_1, m_2, \ldots, m_n) \end{aligned}$$

  is R-linear. In other words, if we fix all arguments of f other than the i-th argument, then f is R-linear as a function of the i-th argument. This is called "f is **linear in the i-th argument**".

So "bilinear" means "multilinear for $n = 2$", whereas "linear" means "multilinear for $n = 1$".

The simplest examples of a multilinear map are

- the determinant map

$$\begin{aligned} \det : \underbrace{R^n \times R^n \times \cdots \times R^n}_{n \text{ times}} &\to R, \\ (v_1, v_2, \ldots, v_n) &\mapsto \det(v_1, v_2, \ldots, v_n); \end{aligned}$$

- the product map

$$\begin{aligned} \operatorname{prod}_n : R^n &\to R, \\ (a_1, a_2, \ldots, a_n) &\mapsto a_1 a_2 \cdots a_n. \end{aligned}$$

- the triple product

$$\begin{aligned} \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 &\to \mathbb{R}, \\ (u, v, w) &\mapsto u \underbrace{\cdot}_{\text{dot product}} (v \times w). \end{aligned}$$

There is a universal property for free modules again.

2.9. Algebras over commutative rings

Convention 2.9.1. In this section, we fix a **commutative** ring R.

2.9.1. Definition

We know rings and we know R-modules. The former have addition and multiplication; the latter have addition and scaling. What happens if we combine these features, to obtain an object that has addition, multiplication and scaling?

This kind of object turns out to be really useful. Here is the precise definition (we impose an extra condition to keep the multiplication and the scaling in harmony):

Definition 2.9.2. An **R-algebra** is a set A endowed with

- two binary operations (i.e., maps from $A \times A$ to A) called **addition** and **multiplication** and denoted by $+$ and $\cdot$;
- a map $\cdot$ from $R \times A$ to A that is called **action** of R and A (not the same as multiplication, despite being called $\cdot$ as well);
- two elements of A called **zero** and **unity** and denoted by 0 and 1,

such that the following axioms (the **algebra axioms**) hold:

- The addition, the multiplication, the zero and the unity satisfy all the ring axioms (so that A is a ring).
- The addition, the action and the zero satisfy all the module axioms (so that A is an R-module).
- **Scale-invariance of multiplication:** We have

$$r\left(ab\right) = \left(ra\right)b = a\left(rb\right) \qquad \text{for all } r \in R \text{ and } a, b \in A.$$

Here (and in the following), we omit the $\cdot$ signs.

Thus, an R-algebra is an R-module that is also a ring at the same time, with the same addition, and satisfying the "scale-invariance" axiom.

The "scale-invariance" axiom can be replaced by requiring that the multiplication map

$$\begin{aligned} A \times A &\to A, \\ (a, b) &\mapsto ab \end{aligned}$$

is R-bilinear. Thus, an R-algebra is "just" an R-module with an R-bilinear multiplication as well as a unity.

You can also restate the "scale-invariance" axiom as "scalars commute with vectors". (More precisely: For each $r \in R$, the "proxy of r in A" – that is, the element $r \cdot 1_A$ – is central in A.)

Some examples of R-algebras include:

- The commutative ring R is itself an R-algebra. Both multiplication and action are just the multiplication of R.
- The zero ring $\{0\}$ is an R-algebra.
- The matrix ring $R^{n \times n}$ is an R-algebra (usually not commutative).
- Any quotient ring R/I of R is a commutative R-algebra.
- The ring $\mathbb{C}$ is an $\mathbb{R}$-algebra.
- More generally: If a ring R is a subring of a **commutative** ring S, then S becomes an R-module and an R-algebra.
- Even more generally: If R and S are two **commutative** rings, and if $f : R \to S$ is a ring morphism, then S becomes an R-algebra. We recall that the R-module structure is given by restriction of scalars, i.e.,

  $$r \cdot s = f(r) \cdot s \qquad \text{for all } r \in R \text{ and } s \in S.$$

  This R-algebra structure on S is said to be **induced** by the morphism f.

  The above example of R/I is an instance of this.
- Even more generally: If R and S are two commutative rings, and if $f : R \to S$ is a ring morphism, then any S-algebra becomes an R-algebra via the "restriction of scalars" rule

  $$r \cdot a = f(r) \cdot a \qquad \text{for all } r \in R \text{ and } a \in A.$$

- The quaternion ring $\mathbb{H}$ is an $\mathbb{R}$-algebra, but **not** a $\mathbb{C}$-algebra (despite $\mathbb{C}$ being a subring of $\mathbb{H}$), because the "scale-invariance of multiplication" axiom is violated. That axiom would require

  $$r(ab) = (ra)b = a(rb) \qquad \text{for all } r \in \mathbb{C} \text{ and } a, b \in \mathbb{H}.$$

  But this fails for $r = i$ and $a = j$ and $b = 1$ for example, since $ij \neq ji$.
- The polynomial ring $R[x]$ (to be defined soon) is an R-algebra.
- More examples: see §3.11.2 in the text.

2.9.2. $\mathbb{Z}$-algebras = rings

Just like any abelian group automatically becomes a $\mathbb{Z}$-module, any ring automatically becomes a $\mathbb{Z}$-algebra:

Proposition 2.9.3. Let A be any ring. Then, A is an abelian group (with respect to $+$), thus a $\mathbb{Z}$-module. Together with the given structure on A, this turns A into a $\mathbb{Z}$-algebra.

Proof. Easy. □

2.9.3. The underlying structures

Every R-algebra A has an underlying ring (i.e., the ring you obtain if you forget the action of R on A) and an underlying R-module (i.e., the R-module you obtain if you forget the multiplication and the unity). We will just refer to them as "the ring A" and "the R-module A".

Thus, when A and B are two R-algebras, then a ring morphism from A to B means a morphism of the underlying rings, whereas an R-module morphism from A to B means a morphism of the underlying R-modules.

Do not mistake the underlying ring (A) for the base ring (R).

2.9.4. Commutative R-algebras

Definition 2.9.4. An R-algebra is said to be **commutative** if its underlying ring is commutative.

2.9.5. Subalgebras

Subalgebras are to algebras what subrings are to rings:

Definition 2.9.5. Let A be an R-algebra. An **R-subalgebra** of A means a subset of A that is simultaneously a subring and an R-submodule of A.

Every R-subalgebra of an R-algebra A becomes an R-algebra in its own right automatically.

2.9.6. R-algebra morphisms

Definition 2.9.6. Let A and B be two R-algebras.

(a) An **R-algebra morphism** (or, short, **algebra morphism**) from A to B means a map $f : A \to B$ that is both a ring morphism and an R-module morphism.

(b) An **R-algebra isomorphism** from A to B means an invertible R-algebra morphism $f : A \to B$ whose inverse $f^{-1} : B \to A$ is an R-algebra morphism as well.

(c) The R-algebras A and B are said to be **isomorphic** (written $A \cong B$) if there is an R-algebra isomorphism from A to B.

All the fundamental properties of ring (iso)morphisms have analogues for algebras instead of rings (see the notes for details).

Furthermore, $\mathbb{Z}$-algebra morphisms are nothing but ring morphisms.

2.9.7. Direct products

Definition 2.9.7. Direct products of R-algebras are defined just as for rings and for R-modules: All structures are entrywise.

2.10. Defining algebras: the case of $\mathbb{H}$

An R-algebra carries more information than a ring, but sometimes this extra information makes it easier to define: Instead of starting with a ring and putting an R-module structure on it, you can start with an R-module and put a ring structure (i.e., is an R-bilinear multiplication with a unity) on it. When you do so, you can use the existing R-module structure as "scaffolding" for defining the ring structure.

We shall now give an example how this works.

Recall the ring $\mathbb{H}$ of Hamilton quaternions, which were "defined" (in Math 331) to be "numbers" of the form $a + bi + cj + dk$ with $a, b, c, d \in \mathbb{R}$ and with multiplication rules

$$i^2 = j^2 = k^2 = -1, \qquad ij = -ji = k, \qquad jk = -kj = i, \qquad ki = -ik = j.$$

It is clear how to calculate in $\mathbb{H}$ using these rules. But it is not clear that this $\mathbb{H}$ exists in the first place.

This is not a vacuous question. For instance, if we replace the rule $k^2 = -1$ by $k^2 = 1$ in the above definition, then we get

$$j^2 \underbrace{k^2}_{=1} = j^2 = -1$$

and thus

$$-1 = j^2k^2 = j \underbrace{jk}_{=i} k = j \underbrace{ik}_{=-j} = j(-j) = -j^2 = -(-1) = 1,$$

showing that our nicely defined ring collapses to the trivial ring ($-1 = 1$ implies $0 = 2$ and thus $0 = 1$ by dividing by 2).

So this is a danger that always exists when you invent new "numbers" and declare new rules. These "numbers" technically exist, but the ring they form might be trivial, or at least much smaller than you expected, and in particular there is no guarantee that your "old" numbers are injectively embedded in it. (The simplest example for this is division by 0: Introduce $\infty = \dfrac{1}{0}$ and you get $0 = 1$.)

So we need to show that this does not happen when we define $\mathbb{H}$.

One safe way to define $\mathbb{H}$ is as follows: We define a quaternion to be a 4-tuple (a, b, c, d) of real numbers (stand-in for $a + bi + cj + dk$), and define the $\mathbb{R}$-algebra operations on the ring of these quaternions by

$$(a, b, c, d) + (a', b', c', d') = (a + a',\ b + b',\ c + c',\ d + d')$$

and

$$\begin{aligned}(a,b,c,d)\left(a',b',c',d'\right) = (\ & aa'-bb'-cc'-dd', \\ & ab'+ba'+cd'-dc', \\ & ac'-bd'+ca'+db', \\ & ad'+bc'-cb'+da'\ )\end{aligned}$$

and

$$r\,(a,b,c,d) = (ra,rb,rc,rd) \qquad \text{for } r \in \mathbb{R}.$$

This is a valid definition, but you have to check that all the ring axioms (and module axioms, and scale-invariance) hold. In particular, associativity of $\cdot$ is a lot of work.

This definition of $\mathbb{H}$ does its job well, but as we just said, it is laborious to justify and somewhat inflexible if we try to generalize it.

A simpler and slicker way to define $\mathbb{H}$ proceeds as follows: First define $\mathbb{H}$ as an $\mathbb{R}$-module (= $\mathbb{R}$-vector space); this is easy: just say $\mathbb{H} = \mathbb{R}^4$ (free $\mathbb{R}$-module). Then, build the multiplication on top of it, using $\mathbb{R}$-bilinearity (since the multiplication in an R-algebra must always be an R-bilinear map). Recall from last time that an $\mathbb{R}$-bilinear map on a free $\mathbb{R}$-module needs to be only specified on a basis. Thus, instead of defining the product of two quaternions, we will only have to define the product of two quaternions from a given basis (which we will take to be $(1,i,j,k)$).

Let us do this. We define $\mathbb{H}$ to be the $\mathbb{R}$-module $\mathbb{R}^4$, which is a free $\mathbb{R}$-module of rank 4. Thus, we define a quaternion to be a 4-tuple (a,b,c,d) of real numbers. The addition and the scaling of quaternions are thus defined entrywise.

We denote the standard basis (e_1,e_2,e_3,e_4) of $\mathbb{R}^4$ by (e,i,j,k). (The e will be the 1, but we don't know yet that it is the unity.)

Now, define the multiplication of $\mathbb{H}$ to be the $\mathbb{R}$-bilinear map $\mu : \mathbb{H} \times \mathbb{H} \to \mathbb{H}$ that satisfies

$$\begin{array}{llll} \mu(e,e)=e, & \mu(e,i)=i, & \mu(e,j)=j, & \mu(e,k)=k, \\ \mu(i,e)=i, & \mu(i,i)=-e, & \mu(i,j)=k, & \mu(i,k)=-j, \\ \mu(j,e)=j, & \mu(j,i)=-k, & \mu(j,j)=-e, & \mu(j,k)=i, \\ \mu(k,e)=k, & \mu(k,i)=j, & \mu(k,j)=-i, & \mu(k,k)=-e. \end{array}$$

By the universal property of free modules wrt bilinear maps, there really is a unique $\mathbb{R}$-bilinear map $\mu : \mathbb{H} \times \mathbb{H} \to \mathbb{H}$; thus we have defined our μ.

We claim that the $\mathbb{R}$-module $\mathbb{H}$ becomes an $\mathbb{R}$-algebra (and thus a ring) if endowed with the multiplication μ and the unity e. For this, we need to show the algebra axioms. Most of them are clear from our construction (a free $\mathbb{R}$-module always satisfies the module axioms, and the bilinearity of μ guarantees a bunch of the other axioms). The only two axioms we need to check are:

1. The map μ is associative (i.e., we have $\mu\left(\mu\left(a,b\right),c\right) = \mu\left(a,\mu\left(b,c\right)\right)$ for all $a,b,c \in \mathbb{H}$).

2. The element e is a neutral element for μ (that is, we have $\mu\left(a,e\right) = \mu\left(e,a\right) = a$ for all $a \in \mathbb{H}$).

Good news: Both of these axioms need only to be checked on the basis! In other words,

1. to check associativity of μ, it suffices to show that $\mu\left(\mu\left(a,b\right),c\right) = \mu\left(a,\mu\left(b,c\right)\right)$ holds for all $a,b,c \in \{e,i,j,k\}$. (Straightforward.)

2. to check neutrality of e, it suffices to show that $\mu\left(a,e\right) = \mu\left(e,a\right) = a$ holds for all $a \in \{e,i,j,k\}$. (Easy.)

This is because μ is R-bilinear, so that these properties are inherited by R-linear combinations. (See the notes for details.)

3. Monoid algebras and polynomials

Convention 3.0.1. For this entire chapter, we fix a **commutative** ring R.

In the previous section, we have learned the "quick" way to define an R-algebra: Define an R-module first, then define its multiplication μ as a certain R-bilinear map which you can specify on the basis elements (if you have a basis). Then, associativity and neutrality of the unity can also be proved just by verifying them on the basis. The keyword for this method is "by linearity".

We shall now apply this strategy to construct an important class of algebras: the **monoid algebras**, including the **group algebras** and the **polynomial rings**.

3.1. Monoid algebras

3.1.1. Definition

Recall the notion of a **monoid**: Roughly speaking, it is a "group without inverses". That is, a **monoid** is a triple $(M,\cdot,1)$, where M is a set, $\cdot$ is an associative binary operation on M, and 1 is a neutral element for $\cdot$. We call $m \cdot n$ the **product** of m and n, and we write mn for it. The monoid M is **abelian** if $mn = nm$ for all $m,n \in M$. Given a monoid $(M,\cdot,1)$, the binary operation $\cdot$ is called the **operation** of M, and the element 1 is called the **neutral element** of M. When the operation is denoted by $\cdot$ and the neutral element by 1, we say that the monoid is **written multiplicatively** (or is a **multiplicative monoid**). When the operation is denoted by $+$ and the neutral element by 0, we say that the monoid is **written additively** (or is an **additive monoid**).

If M is a monoid written multiplicatively, then we can define the **monoid algebra** $R[M]$. Informally, this is the R-algebra obtained by "throwing" the elements of M "into" the ring R. Its elements are "formal R-linear combinations of elements of M", that is, expressions of the form

$$r_1 m_1 + r_2 m_2 + \cdots + r_k m_k$$

with $k \in \mathbb{N}$ and $m_1, m_2, \ldots, m_k \in M$ and $r_1, r_2, \ldots, r_k \in M$. These expressions are multiplied by distributivity and using the multiplications of R and M: that is,

$$\begin{aligned} &\left(r_1 m_1 + r_2 m_2 + \cdots + r_k m_k\right)\left(s_1 n_1 + s_2 n_2 + \cdots + s_\ell n_\ell\right) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{\ell} \underbrace{r_i s_j}_{\substack{\text{product} \\ \text{in } R}} \underbrace{m_i n_j}_{\substack{\text{product} \\ \text{in } M}} . \end{aligned}$$

In order to make this rigorous, let us recall some concepts:

If M is any set, then R^M is the R-module

$$\left\{ \left(r_m\right)_{m \in M} \mid r_m \in R \text{ for each } m \in M \right\},$$

whereas $R^{(M)}$ is its R-submodule

$$\left\{ \left(r_m\right)_{m \in M} \in R^M \mid r_m = 0 \text{ for all but finitely many } m \in M \right\}.$$

If M is finite, then $R^{(M)} = R^M$.

The R-module $R^{(M)}$ is free, and its **standard basis** $\left(e_m\right)_{m \in M}$ is defined as follows: Each e_m is the family whose m-th entry is 1 while all its other entries are 0.

Now we can define the monoid algebra $R[M]$ rigorously:

Definition 3.1.1. Let M be a monoid, written multiplicatively (so that $\cdot$ denotes its operation, and 1 its neutral element).

The **monoid algebra** of M over R (also known as the **monoid ring** of M over R) is the R-algebra $R[M]$ defined as follows:

As an R-module, it is the free R-module

$$R^{(M)} = \left\{ \left(r_m\right)_{m \in M} \in R^M \mid r_m = 0 \text{ for all but finitely many } m \in M \right\}.$$

Its multiplication is defined to be the unique R-bilinear map $\mu : R^{(M)} \times R^{(M)} \to R^{(M)}$ that satisfies

$$\mu\left(e_m, e_n\right) = e_{mn} \qquad \text{for all } m, n \in M,$$

where $\left(e_m\right)_{m \in M}$ is the standard basis of $R^{(M)}$. The unity of this R-algebra is e_1.

Theorem 3.1.2. This is indeed a well-defined R-algebra.

Proof. By linearity (and associativity of M). See Theorem 4.1.2 in the text for details. □

Definition 3.1.3. If G is a group, then its monoid algebra $R[G]$ is called the **group algebra** (or **group ring**) of G over R.

3.1.2. Examples

Example 3.1.4. Consider the cyclic group C_2 of order 2. We write it multiplicatively as $C_2 = \{1, u\}$ where $u^2 = 1$. (Written additively, it is just $\mathbb{Z}/2$, but we want it multiplicative.)

What does its group algebra (= monoid algebra) $\mathbb{Q}[C_2]$ look like?

As a $\mathbb{Q}$-module, it is

$$\begin{aligned}\mathbb{Q}^{(C_2)} &= \mathbb{Q}^{(\{1,u\})} = \mathbb{Q}^{\{1,u\}} \\ &= \left\{ (r_m)_{m\in\{1,u\}} \ \mid \ r_m \in \mathbb{Q} \text{ for each } m \in \{1,u\} \right\}.\end{aligned}$$

A family of this form $(r_m)_{m\in\{1,u\}}$ consists of just two entries: r_1 and r_m. By abuse of notation, we can thus identify such a family with the pair (r_1, r_u). Thus,

$$\mathbb{Q}^{(C_2)} = \{(r_1, r_u) \ \mid \ r_1, r_u \in \mathbb{Q}\} = \mathbb{Q}^2.$$

The addition and the action of the group algebra $\mathbb{Q}[C_2]$ are entrywise. What about its multiplication?

Its standard basis is $(e_m)_{m\in\{1,u\}} = (e_1, e_u)$, where $e_1 = (1,0)$ and $e_u = (0,1)$. The multiplication of the group algebra $\mathbb{Q}[C_2]$ is given by

$$\mu(e_m, e_n) = e_{mn} \qquad \text{for all } m, n \in C_2.$$

In other words,

$$e_m e_n = e_{mn} \qquad \text{for all } m, n \in C_2.$$

Explicitly,

$$\begin{aligned} &e_1e_1 = e_{1\cdot 1} = e_1, \qquad && e_1e_u = e_{1u} = e_u, \\ &e_ue_1 = e_{u1} = e_u, \qquad && e_ue_u = e_{uu} = e_1 \qquad \left(\text{since } uu = u^2 = 1\right). \end{aligned}$$

Since (e_1, e_u) is a basis of $\mathbb{Q}[C_2]$, we can write each element of $\mathbb{Q}[C_2]$ uniquely as $ae_1 + be_u$ with $a, b \in \mathbb{Q}$. To multiply two such elements, we use $\mathbb{Q}$-bilinearity of μ:

$$\begin{aligned}&(ae_1 + be_u)(ce_1 + de_u) \\ &= ace_1e_1 + bce_ue_1 + ade_1e_u + bde_ue_u \\ &= ace_1 + bce_u + ade_u + bde_1 \\ &= (ac + bd)e_1 + (bc + ad)e_u.\end{aligned}$$

Using the pair notation for these elements, $ae_1 + be_u$ is simply (a, b), so this multiplication rule rewrites as

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

This looks a lot like the multiplication rule for complex numbers (as pairs of real numbers), which says

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

Thus, we can think of $\mathbb{Q}[C_2]$ as a "twin brother" of $\mathbb{C}$, except that we are using $\mathbb{Q}$ instead of $\mathbb{R}$ as the base ring (but we could just as well have used $\mathbb{R}$ or any other commutative ring). So we should instead think of $\mathbb{R}[C_2]$ as a "twin brother" of $\mathbb{C}$.

However, this "twin brother" behaves rather differently from $\mathbb{C}$ in many ways. For example, $\mathbb{C}$ is a field, but $\mathbb{R}[C_2]$ is not a field. In fact, $\mathbb{R}[C_2]$ is not even an integral domain, because

$$(1, 1)(1, -1) = (0,\ 0) = 0_{\mathbb{R}[C_2]}.$$

Or, in terms of linear combinations of basis vectors:

$$(e_1 + e_u)(e_1 - e_u) = 0.$$

Actually, we can say more:

$$\mathbb{Q}[C_2] \cong \mathbb{Q} \times \mathbb{Q} \qquad \text{as } \mathbb{Q}\text{-algebras.}$$

Indeed, there is a $\mathbb{Q}$-algebra isomorphism

$$\begin{gathered} \mathbb{Q}[C_2] \to \mathbb{Q} \times \mathbb{Q}, \\ ae_1 + be_u = (a, b) \mapsto (a + b,\ a - b) \end{gathered}$$

(the simplest case of the discrete Fourier transform). Its inverse is

$$\begin{gathered} \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}[C_2], \\ (c, d) \mapsto \left(\frac{c + d}{2},\ \frac{c - d}{2} \right). \end{gathered}$$

See the notes for a more conceptual way to find this. Note that we can redo the above computations with $\mathbb{R}$ instead of $\mathbb{Q}$, but we cannot do them with $\mathbb{Z}$ instead of $\mathbb{Q}$. The group algebra $\mathbb{Z}[C_2]$ is not isomorphic to $\mathbb{Z} \times \mathbb{Z}$, and in fact it is not isomorphic to any nontrivial direct product, despite not being an integral domain.

Example 3.1.5. Consider the cyclic group $C_3 = \{1, u, v\}$ of order 3 with $u^3 = 1$ and $v = u^2$. Its group algebra $\mathbb{Q}[C_3]$ is then $\mathbb{Q}^3$ as a $\mathbb{Q}$-vector space (upon identifying each family $(r_m)_{m \in C_3}$ with the triple (r_1, r_u, r_v)). Its multiplication rule is given by

$$\begin{aligned}&\left(ae_1 + be_u + ce_v\right)\left(a'e_1 + b'e_u + c'e_v\right)\\ &= \left(aa' + bc' + cb'\right)e_1 + \left(ab' + ba' + cc'\right)e_u + \left(ac' + bb' + ca'\right)e_v,\end{aligned}$$

aka

$$\begin{aligned}&(a, b, c)\left(a', b', c'\right)\\ &= \left(aa' + bc' + cb',\ ab' + ba' + cc',\ ac' + bb' + ca'\right).\end{aligned}$$

Again, this is not an integral domain, since

$$(1, 1, 1) \cdot (1, -1, 0) = (0, 0, 0).$$

Actually, you can show that

$$\mathbb{Q}[C_3] \cong \mathbb{Q} \times S$$

for some $\mathbb{Q}$-algebra S that is 2-dimensional as a $\mathbb{Q}$-vector space.

More generally, for any finite group G, the group algebra $\mathbb{Q}[G]$ has a central idempotent

$$z := \frac{\sum_{g \in G} e_g}{|G|}$$

(prove this!), and the principal ideal $z\mathbb{Q}[G]$ is $\cong \mathbb{Q}$ as a $\mathbb{Q}$-algebra, so that using one old exercise (HW#4 Exercise 1 **(d)**), we conclude that

$$\mathbb{Q}[G] \cong \mathbb{Q} \times S$$

for some subalgebra $S = (1 - z)\mathbb{Q}[G]$. Sometimes (but not always), you can split the S further into direct products. If $G = C_2$, then $S \cong \mathbb{Q}$, but in general S is more complicated.

If you work with $\mathbb{C}$ instead of $\mathbb{Q}$, then for any finite abelian group G you can actually split

$$\mathbb{C}[G] \cong \underbrace{\mathbb{C} \times \mathbb{C} \times \cdots \times \mathbb{C}}_{|G| \text{ times}}$$

(the discrete Fourier transform). This relies on roots of unity ($e^{2\pi ik/n}$), so it does require the base ring to be $\mathbb{C}$ (or something else that has these roots of unity).

For non-abelian groups G, even roots of unity don't let you decompose the algebra $\mathbb{C}[G]$ into a direct product of $\mathbb{C}$s. No surprise here: If G is not abelian, then $\mathbb{C}[G]$ will not be commutative.

3.1.3. General properties of monoid algebras

Here are some general properties of and conventions about monoid algebras.

Proposition 3.1.6. Let M be an **abelian** monoid. Then, the monoid ring $R[M]$ is commutative.

Proof. By linearity. (Proposition 4.1.9 in the text.) □

Proposition 3.1.7. Let M be a monoid with neutral element 1. Then, the map

$$\begin{aligned} R &\to R[M], \\ r &\mapsto r \cdot e_1 \end{aligned}$$

is an injective R-algebra morphism.

Proof. Easy. (Morphicity relies on $e_1 e_1 = e_1$.) □

Convention 3.1.8. If M is a monoid, then we shall identify each $r \in R$ with $r \cdot e_1 \in R[M]$. This identification is harmless (i.e., does not lead to any false conclusions), since the map in the previous proposition is an injective R-algebra morphism. Thus, R turns into an R-subalgebra of $R[M]$.

An element of the form $r \cdot e_1 \in R[M]$ are called **constant**.

Proposition 3.1.9. Let M be a monoid. Then, the map

$$\begin{aligned} M &\to R[M], \\ m &\mapsto e_m \end{aligned}$$

is a monoid morphism from M to $(R[M], \cdot, 1)$.

Proof. This is just saying that $e_m e_n = e_{mn}$ for all $m, n \in M$, and that $e_1 = 1_{R[M]}$. Both are clear from the definition of $R[M]$. □

The last two propositions show that the monoid algebra $R[M]$ contains "a copy of" R and "a copy of" M. However, the "copy of M" is collapsed if R is a trivial ring.

Convention 3.1.10. Let M be a monoid. Then, the elements e_m of the standard basis $(e_m)_{m \in M}$ of $R[M]$ will just be written as m if no confusion can arise.

For example, if M is the cyclic group $C_3 = \{1, u, v\}$ as in the above example, then we write the element $ae_1 + be_u + ce_v$ as $a1 + bu + cv = a + bu + cv$.

For another example, if M is the cyclic group $C_2 = \{1, u\}$ as above, then $ae_1 + be_u$ becomes $a + bu$.

3.2. Polynomial rings

3.2.1. Univariate polynomials

We can now effortlessly define univariate polynomials: They are just elements of certain monoid algebras. Which ones?

Recall that R is a commutative ring, whereas $\mathbb{N} = \{0, 1, 2, \ldots\}$.

Definition 3.2.1. Let C be the free monoid with a single generator x. This is the monoid whose elements are countably many distinct symbols called

$$x^0, \quad x^1, \quad x^2, \quad x^3, \quad \ldots$$

with operation given by

$$x^i \cdot x^j = x^{i+j} \qquad \text{for all } i, j \in \mathbb{N}.$$

Of course, this monoid is just the well-known additive monoid $(\mathbb{N}, +, 0)$ rewritten in a multiplicative form (with each element $i \in \mathbb{N}$ renamed as x^i).

The neutral element of this monoid C is x^0. We set $x := x^1$.

The elements of C are called **monomials** in the variable x. The specific element x is called the **indeterminate**.

Now, the **univariate polynomial ring** $R[x]$ over R is defined to be the monoid algebra $R[C]$. Following the conventions above, we simply write m for each standard basis vector e_m. That is, we write x^i for e_{x^i}. Thus, $R[x]$ is a free R-module with basis

$$\left(x^0, x^1, x^2, \ldots\right) = \left(1, x, x^2, x^3, \ldots\right).$$

Hence, any $p \in R[x]$ can be written as a finite R-linear combination of powers of x. That is, p can be written as

$$p = a_0 x^0 + a_1 x^1 + a_2 x^2 + \cdots + a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

for some $n \in \mathbb{N}$ and some $a_0, a_1, \ldots, a_n \in R$. This representation is unique up to trailing zeroes (i.e., up to adding extra terms of the form $0x^{n+1}$ and $0x^{n+2}$ and so on).

Elements of $R[x]$ are called **polynomials** in x over R.

Note that the ring $R[x]$ is commutative, since it is the monoid ring of the abelian monoid C.

Example 3.2.2. (a) This is a polynomial: $1 + 3x^2 + 9x^5 \in \mathbb{Z}[x]$ (also in $\mathbb{Q}[x]$ and so on).

(b) This is not a polynomial: $1 + x + x^2 + x^3 + \cdots$, unless R is trivial. Infinite sums like this are called **formal power series** and lie not in $R^{(C)}$ but

in R^C. They also form a ring, but this is a story for another class (e.g., Math 531 Algebraic Combinatorics).

So we have defined **univariate** polynomials (i.e., polynomials in one variable). We can similarly define **multivariate** polynomials (i.e., polynomials in many variables). For simplicity, I restrict myself to the case of finitely many variables.

3.2.2. Multivariate polynomials

Definition 3.2.3. Let $n \in \mathbb{N}$. Let $C^{(n)}$ be the free abelian monoid with n generators $x_1, x_2, \ldots, x_n$. This is the monoid whose generators are the distinct symbols

$$x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \qquad \text{with } i_1, i_2, \ldots, i_n \in \mathbb{N}$$

and with operation given by

$$\left(x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}\right)\left(x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n}\right) = x_1^{i_1+j_1} x_2^{i_2+j_2} \cdots x_n^{i_n+j_n}.$$

This monoid is just the additive monoid $\mathbb{N}^n$ (with entrywise addition), written multiplicatively (with each n-tuple $(i_1, i_2, \ldots, i_n)$ renamed as $x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}$).

The elements of $C^{(n)}$ are called **monomials**.

For each $i \in \{1, 2, \ldots, n\}$, we define the monomial x_i to be

$$x_i = x_1^0 x_2^0 \cdots x_{i-1}^0 x_i^1 x_{i+1}^0 x_{i+2}^0 \cdots x_n^0.$$

These specific monomials x_i are called the **indeterminates**.

Now, the monoid algebra $R\left[C^{(n)}\right]$ is denoted $R[x_1, x_2, \ldots, x_n]$, and is called the **polynomial ring in n variables $x_1, x_2, \ldots, x_n$ over R**. Its elements are called **polynomials** in $x_1, x_2, \ldots, x_n$. A typical polynomial looks like this:

$$\sum_{(i_1, i_2, \ldots, i_n) \in \mathbb{N}^n} r_{i_1, i_2, \ldots, i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}$$

with $r_{i_1, i_2, \ldots, i_n} \in R$ such that all but finitely many of these coefficients $r_{i_1, i_2, \ldots, i_n}$ are 0. You can also write it as

$$\sum_{m \in C^{(n)}} r_m m.$$

For example,

$$4x_1^2 + 3x_2 x_3 - x_5 + 6$$

is a polynomial in $x_1, x_2, \ldots, x_n$ whenever $n \geq 5$ and for any base ring R. Likewise, $\frac{1}{2}x_1 + 7x_2x_3 - \pi x_1 x_3^2$ is a polynomial in x_1, x_2, x_3 over $\mathbb{R}$ or $\mathbb{C}$.

Note that the univariate polynomial ring $R[x]$ is the particular case of the multivariate polynomial ring $R[x_1, x_2, \ldots, x_n]$ for $n = 1$, once you rename the x_1 as x.

So our polynomials are pretty much formal objects. They are NOT functions of an argument in R. In fact, we have yet to see how polynomials can be evaluated. This will be done next time.

3.2.3. Constant polynomials

A **constant polynomial** is a constant element of the monoid ring $R\left[C^{(n)}\right]$ (or $R[C]$ in the univariate case): i.e., a scalar multiple of the monomial $1 = x_1^0 x_2^0 \cdots x_n^0$. We can identify the constant polynomials with the elements of R, thus making R into a subring of any polynomial ring over R.

3.2.4. Coefficients

By their definition, polynomials are R-linear combinations of monomials. We now introduce a notation for their coefficients:

Definition 3.2.4. Let $p \in R[x_1, x_2, \ldots, x_n]$ be a polynomial. Let $m = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$ be a monomial. Then, the **coefficient** of m in p is the element $[m]\, p$ defined as follows: If we write p as

$$p = \sum_{(i_1, i_2, \ldots, i_n) \in \mathbb{N}^n} p_{i_1, i_2, \ldots, i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \qquad \left(\text{with } p_{i_1, i_2, \ldots, i_n} \in R\right),$$

then

$$[m]\, p := p_{a_1, a_2, \ldots, a_n}.$$

For example,

$$\left[x^3\right]\left((1+x)^5\right) = \binom{5}{3} = 10 \qquad \text{and}$$
$$\left[x^6\right]\left((1+x)^5\right) = 0$$

(since $(1+x)^5 = x^5 + 5x^4 + 10x^3 + 10x^2 + 5x + 1$). For another example, working in $R[x_1, x_2]$ and renaming the variables x_1, x_2 as x, y, we have

$$\left[x^2 y^3\right]\left((x+y)^5\right) = 10$$

and

$$[xy]\left((x+y)^5\right) = 0$$

(since $(x+y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$).

3.2.5. Symbols for indeterminates

In our above definition of a multivariate polynomial ring $R[x_1, x_2, \ldots, x_n]$, we have "hardcoded" the indeterminates to be called $x_1, x_2, \ldots, x_n$. Often, you want a more flexible definition, where you can call the variables whatever you want, and you want the resulting polynomial rings to "know" how the variables are called. For example, you want the polynomial rings $R[x, y]$ and $R[y, z]$ to be isomorphic, but not the same ring (and also different from $R[x_1, x_2]$).

This necessitates some minor changes to our definition of multivariate polynomials. Namely, instead of using the monoid

$$C^{(n)} = \left\{ x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \mid (i_1, i_2, \ldots, i_n) \in \mathbb{N}^n \right\},$$

we now use the monoid

$$C^{(S)} = \left\{ \prod_{s \in S} s^{i_s} \mid i_s \in \mathbb{N} \text{ for each } s \in S \right\}$$

where S is our chosen (finite) set of indeterminates (e.g., $S = \{x, y\}$ or $S = \{y, z\}$ or $S = \{\alpha, \omega, \mathcal{F}, \Re\}$). A monomial in this monoid $C^{(S)}$ is a "formal" product of the form $\prod_{s \in S} s^{i_s}$, really a map from S to $\mathbb{N}$ that sends each $s \in S$ to i_s. These monomials are multiplied by the rule

$$\left(\prod_{s \in S} s^{i_s} \right) \left(\prod_{s \in S} s^{j_s} \right) = \prod_{s \in S} s^{i_s + j_s}.$$

We shall refer to the monoid ring $R\left[C^{(S)}\right]$ as a **multivariate polynomial ring with named variables**, and just call it $R[S]$.

We will be cavalier about this all, and pretend that this naming issue is a non-issue.

3.3. Univariate polynomials

Let us now take a closer look at univariate polynomial rings, as these have the best properties of all the polynomial rings.

3.3.1. Degrees and coefficients

Recall: If $p = \sum_{j \in \mathbb{N}} p_j x^j \in R[x]$ with $p_j \in R$, then $\left[x^i\right] p = p_i$ for each $i \in \mathbb{N}$.

Definition 3.3.1. Let $p \in R[x]$ be a univariate polynomial.

(a) If $p \neq 0$, then the **degree** of p is defined to be the largest $i \in \mathbb{N}$ such that $\left[x^i\right] p \neq 0$. The degree of the zero polynomial $0 \in R[x]$ is defined to be $-\infty$.

The degree of p is called $\deg p$.

(b) If $p \neq 0$, then the **leading coefficient** of p is defined to be $\left[x^{\deg p}\right] p \in R$.

(c) The polynomial p is said to be **monic** if its leading coefficient is 1.

For example, the polynomial

$$5x^3 + 2x + 1 \in \mathbb{Q}[x]$$

has degree 3 and leading coefficient 5, thus is not monic. The polynomial

$$\overline{5}x^3 + \overline{2}x + \overline{1} \in (\mathbb{Z}/n)[x] \qquad \text{(for a given integer } n > 0\text{)}$$

has

- degree 3 if $n \neq 5, 1$;
- degree 1 if $n = 5$;
- degree $-\infty$ if $n = 1$ (since it is just 0 in this case).

Remark 3.3.2. Let $n \in \mathbb{N}$. Then,

$$\begin{aligned}
&\{f \in R[x] \mid \deg f \leq n\} \\
&= \left\{f \in R[x] \mid f = a_0 x^0 + a_1 x^1 + \cdots + a_n x^n \text{ for some } a_i \in R\right\} \\
&= \operatorname{span}\left(x^0, x^1, \ldots, x^n\right).
\end{aligned}$$

In particular, this is an R-submodule of $R[x]$.

Corollary 3.3.3. Let $p, q \in R[x]$. Then,

$$\begin{aligned}
&\deg(p+q) \leq \max\{\deg p, \deg q\} \qquad \text{and} \\
&\deg(p-q) \leq \max\{\deg p, \deg q\}.
\end{aligned}$$

Remark 3.3.4. The polynomials of degree ≤ 0 are just the constant polynomials.

Proposition 3.3.5. Let $p, q \in R[x]$. Then:
(a) We have $\deg(pq) \leq \deg p + \deg q$.
(b) We have $\deg(pq) = \deg p + \deg q$ if $p \neq 0$ and the leading coefficient of p is a unit.
(c) We have $\deg(pq) = \deg p + \deg q$ if R is an integral domain.
(d) If $n, m \in \mathbb{N}$ satisfy $n \geq \deg p$ and $m \geq \deg q$, then

$$\left[x^{n+m}\right](pq) = \left[x^n\right](p) \cdot \left[x^m\right](q).$$

(e) If $pq = 0$ and $p \neq 0$ and if the leading coefficient of p is a unit, then $q = 0$.

Corollary 3.3.6. If R is an integral domain, then so is the polynomial ring $R[x]$.

Remark 3.3.7. If R is not an integral domain, then you can get shenanigans with degrees. For instance, if $R = \mathbb{Z}/4$, then

$$\left(\overline{1} + \overline{2}x\right)^2 = \overline{1} + \overline{4}x + \overline{4}x^2 = \overline{1} \qquad \left(\text{since } \overline{1} = \overline{0}\right).$$

So the degree of a polynomial can decrease when it is squared!

All the above results have pretty easy proofs. See Proposition 4.3.5 in the notes.

3.3.2. Division with remainder

Just like integers, univariate polynomials can be divided with remainder, as long as the polynomial you are dividing by has an invertible (= unit) leading coefficient:

Theorem 3.3.8. Let $a, b \in R[x]$ be two polynomials such that b is nonzero and the leading coefficient of b is a unit.
(a) Then, there is a **unique** pair (q, r) of polynomials in $R[x]$ such that

$$a = qb + r \qquad \text{and} \qquad \deg r < \deg b.$$

(b) Moreover, this pair satisfies $\deg q \leq \deg a - \deg b$.

Proof. See Theorem 4.3.7 in the text.

Example: $a = 4x^4 + 2x^3 - 3x + 5$ and $b = x^2 - x + 7$. Then, we want q, r with

$$a = qb + \underbrace{r}_{\deg<2}, \qquad \text{that is,}$$
$$4x^4 + 2x^3 - 3x + 5 = q \cdot \left(x^2 - x + 7\right) + r.$$

To make the $4x^4$-terms agree, we want q to have degree ≤ 2 and we want $\left[x^2\right] q = 4$. Thus, we get $q = 4x^2 + q'$ (note: q' is not the derivative q), so the above equality becomes

$$4x^4 + 2x^3 - 3x + 5 = \left(4x^2 + q'\right) \cdot \left(x^2 - x + 7\right) + r, \qquad \text{that is,}$$
$$4x^4 + 2x^3 - 3x + 5 = 4x^2\left(x^2 - x + 7\right) + q'\left(x^2 - x + 7\right) + r, \qquad \text{that is,}$$
$$4x^4 + 2x^3 - 3x + 5 - 4x^2\left(x^2 - x + 7\right) = q'\left(x^2 - x + 7\right) + r, \qquad \text{that is,}$$
$$6x^3 - 28x^2 - 3x + 5 = q'\left(x^2 - x + 7\right) + r.$$

To make the $6x^3$ terms agree here, we want q' to have degree ≤ 1 and we want $\left[x^1\right] q' = 6$. Thus, we get $q' = 6x^1 + q''$ and

$$6x^3 - 28x^2 - 3x + 5 = \left(6x^1 + q''\right)\left(x^2 - x + 7\right) + r, \qquad \text{that is,}$$
$$-22x^2 - 45x + 5 = q''\left(x^2 - x + 7\right) + r.$$

To make the $-22x^2$ terms agree here, we want q'' to have degree ≤ 0 and we want $\left[x^0\right] q'' = -22$. Thus we get $q'' = -22$ and

$$-22x^2 - 45x + 5 = -22\left(x^2 - x + 7\right) + r,$$

so that

$$r = -22x^2 - 45x + 5 - \left(-22\left(x^2 - x + 7\right)\right) = 159 - 67x.$$

This is a valid remainder, since it has degree $< \deg b$. Moreover, substituting back in, we find $q = 4x^2 + 6x^1 + (-22)$. □

The polynomials q and r in the above theorem are called the **quotient** and the **remainder** obtained when dividing a by b. Note that if $\deg a < \deg b$, then $q = 0$ and $r = a$. Also note that $b \mid a$ if and only if $r = 0$.

Don't forget the condition "the leading coefficient of b is a unit". This condition is automatically satisfied if b is monic or if R is a field, but not in general (see homework set #6 for some examples).

3.3.3. Evaluation of polynomials

Polynomials can not only be added, scaled, multiplied etc., but they can also be **evaluated**, meaning that we can substitute things into them. Let us define how this evaluation works, first for univariate polynomials:

Definition 3.3.9. Let $p \in R[x]$ be a univariate polynomial. Let A be any R-algebra. Let $a \in A$.

We define the element $p(a)$ aka $p[a]$ of A as follows: Write p as

$$p = \sum_{i \in \mathbb{N}} p_i x^i \qquad (\text{where } p_i \in R),$$

and set

$$p(a) := \sum_{i \in \mathbb{N}} p_i a^i.$$

This element $p(a)$ aka $p[a]$ is called the **evaluation** of p at a; we say that it is obtained by **substituting** a for x in p.

A few comments:

- The A here can be any R-algebra, not just R itself. For example, A can be $R^{n \times n}$ (a case known from linear algebra) or $R[x]$ (in which case you are evaluating a polynomial at another polynomial – this is called composition of polynomials). So a polynomial is not a function – it allows for a-priori unbounded possibilities of domain.
- We cannot do this with formal power series (i.e., infinite sums like $1 + x + x^2 + x^3 + \cdots$).
- Note that $p[x] = p$.
- **Warning:** The notation $p(a)$ can be ambiguous. For example, what is $x(x+1)$? Is it the product $x \cdot (x+1)$ or the evaluation $x[x+1]$? So be careful with it, and fall back to $p[a]$ if necessary.
- Evaluation can act weird. For instance, let $R = \mathbb{Z}/2$ and $p = x^2 + x \in R[x]$. Then, evaluating p at the elements of R yields

$$p\left(\overline{0}\right) = \overline{0}^2 + \overline{0} = \overline{0};$$
$$p\left(\overline{1}\right) = \overline{1}^2 + \overline{1} = \overline{2} = \overline{0}.$$

That is, $p(r) = 0$ for all $r \in R$. But evaluating p at the 2×2-matrix

$\begin{pmatrix} \overline{0} & \overline{1} \\ \overline{1} & \overline{0} \end{pmatrix} \in R^{2\times 2}$ yields

$$p\left[\begin{pmatrix} \overline{0} & \overline{1} \\ \overline{1} & \overline{0} \end{pmatrix}\right] = \begin{pmatrix} \overline{0} & \overline{1} \\ \overline{1} & \overline{0} \end{pmatrix}^2 + \begin{pmatrix} \overline{0} & \overline{1} \\ \overline{1} & \overline{0} \end{pmatrix} = \begin{pmatrix} \overline{1} & \overline{1} \\ \overline{1} & \overline{1} \end{pmatrix} \neq 0_{R^{2\times 2}}.$$

So a polynomial can be nonzero even if all its values on scalars (= elements of R) are zero.

Given an R-algebra A and an element $a \in A$, the operation of evaluating polynomials $p \in R[x]$ at a behaves nicely:

Theorem 3.3.10. Let A be an R-algebra. Let $a \in A$. Then, the map

$$\begin{aligned} R[x] &\to A, \\ p &\mapsto p[a] \end{aligned}$$

is an R-algebra morphism. Explicitly, this is saying that:

$$\begin{aligned} (p+q)[a] &= p[a] + q[a] && \text{for all } p, q \in R[x]; \\ (pq)[a] &= p[a] \cdot q[a] && \text{for all } p, q \in R[x]; \\ (\lambda p)[a] &= \lambda \cdot p[a] && \text{for all } \lambda \in R \text{ and } p \in R[x]; \\ 0[a] &= 0; \\ 1[a] &= 1, \end{aligned}$$

The proof of this is easy, but it becomes even easier using the following lemma:

Lemma 3.3.11. Let A and B be two R-algebras. Let $f : A \to B$ be an R-linear map. Let $(m_i)_{i\in I}$ be a family of vectors in A that spans A. If we have

$$f\left(m_i m_j\right) = f\left(m_i\right) f\left(m_j\right) \qquad \text{for all } i, j \in I,$$

then

$$f(ab) = f(a) f(b) \qquad \text{for all } a, b \in A.$$

Proof. By linearity. □

Proof of the theorem. Easy using the lemma. □

We can similarly evaluate multivariate polynomials, but now we need to require that the "inputs" commute. Otherwise, the true equality $xy = yx$ in the polynomial ring $R[x, y]$ would become the not-necessarily-true equality $ab = ba$ in every R-algebra A.

Definition 3.3.12. Let $n \in \mathbb{N}$. Let $p \in R[x_1, x_2, \ldots, x_n]$ be a multivariate polynomial. Let A be any R-algebra. Let $a_1, a_2, \ldots, a_n$ be n elements of A that mutually commute (i.e., that satisfy $a_i a_j = a_j a_i$ for all i, j).

We define the element $p(a_1, a_2, \ldots, a_n)$ aka $p[a_1, a_2, \ldots, a_n]$ of A as follows: Write the polynomial p as

$$p = \sum_{(i_1, i_2, \ldots, i_n) \in \mathbb{N}^n} p_{i_1, i_2, \ldots, i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \qquad \left(\text{with } p_{i_1, i_2, \ldots, i_n} \in R\right),$$

and set

$$p(a_1, a_2, \ldots, a_n) := \sum_{(i_1, i_2, \ldots, i_n) \in \mathbb{N}^n} p_{i_1, i_2, \ldots, i_n} a_1^{i_1} a_2^{i_2} \cdots a_n^{i_n}.$$

This element $p(a_1, a_2, \ldots, a_n)$ is called the **evaluation** of p at $a_1, a_2, \ldots, a_n$, and we say that it is obtained by **substituting** $a_1, a_2, \ldots, a_n$ for $x_1, x_2, \ldots, x_n$ in p.

There is an analogue of the above theorem:

Theorem 3.3.13. Let $n \in \mathbb{N}$. Let A be any R-algebra. Let $a_1, a_2, \ldots, a_n$ be n elements of A that mutually commute (i.e., that satisfy $a_i a_j = a_j a_i$ for all i, j). Then, the map

$$\begin{aligned} R[x_1, x_2, \ldots, x_n] &\to A, \\ p &\mapsto p[a_1, a_2, \ldots, a_n] \end{aligned}$$

is an R-algebra morphism.

Proof. Somewhat analogous to the above; see the text for details (Theorem 4.2.11). Note that you need

$$\left(a_1^{i_1} a_2^{i_2} \cdots a_n^{i_n}\right)\left(a_1^{j_1} a_2^{j_2} \cdots a_n^{j_n}\right) = a_1^{i_1+j_1} a_2^{i_2+j_2} \cdots a_n^{i_n+j_n},$$

and that's where the commutativity of the a_is comes useful. □

3.3.4. Roots

Our notion of roots is quite liberal:

Definition 3.3.14. Let A be an R-algebra. Let $p \in R[x]$ be a polynomial. An element $a \in A$ is called a **root** of p if $p[a] = 0$.

For example, the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathbb{Q}^{2\times 2}$ is a root of the polynomial $x^2 - 1$, since

$$\left(x^2 - 1\right)\left[\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right] = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2 - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 0_{\mathbb{Q}^{2\times 2}}.$$

For another example, the polynomial $x^2 + 1 \in \mathbb{Z}[x]$ has infinitely many roots in the quaternion ring $\mathbb{H}$, since any "purely imaginary" quaternion $ai + bj + ck$ satisfies

$$(ai + bj + ck)^2 = -\left(a^2 + b^2 + c^2\right)$$

and thus is a root of $x^2 + 1$ whenever $a^2 + b^2 + c^2 = 1$.

Let us say a few first things about roots in R:

Proposition 3.3.15. Let f be a polynomial in $R[x]$. Let $a \in R$. Then, a is a root of f if and only if $x - a \mid f$ in $R[x]$.

Proof. See §4.3.3 in the text. In a nutshell: Divide f by $x - a$ with remainder, and observe that $x - a \mid f$ if and only if the remainder is 0. □

The following theorem is often known as the **easy half of the FTA (Fundamental Theorem of Algebra)**:

Theorem 3.3.16 (easy half of FTA)**.** Let R be an integral domain. Let $n \in \mathbb{N}$. Then, any nonzero polynomial $f \in R[x]$ of degree $\leq n$ has at most n roots in R. (We are not counting multiplicities here.)

Proof. See §4.3.3 in the text. In a nutshell: Each root a allows you to divide f by $x - a$, which lowers the degree of f by 1. All roots distinct from a remain roots when you do this, because R is an integral domain. Obviously, you cannot do this more than n times, since the degree of a nonzero polynomial cannot be negative. □

Remark 3.3.17. The "full" FTA says that a nonzero polynomial of degree n over $\mathbb{C}$ has exactly n roots, counted with multiplicities. This is not really a theorem of algebra, since it relies on analytic properties of $\mathbb{C}$. Some references to proofs can be found in the notes. For us in abstract algebra, the "full" FTA is usually not needed.

3.3.5. $F[x]$ is a Euclidean domain

The division-with-remainder theorem for polynomials looks very much like the definition of a Euclidean ring. However, it has the annoying condition "the leading coefficient of b is a unit", which does not allow it to be used as widely as the definition of a Euclidean ring would require. However, when R is a field, every leading coefficient is a unit, so the condition becomes unnecessary. Thus, we obtain:

Theorem 3.3.18. Let F be a field. Then, the polynomial ring $F[x]$ is a Euclidean domain with Euclidean norm

$$N : F[x] \to \mathbb{N},$$
$$p \mapsto \max\{\deg p, 0\} = \begin{cases} \deg p, & \text{if } p \neq 0; \\ 0, & \text{if } p = 0. \end{cases}$$

Proof. As just explained. □

This allows us to apply the machinery of Euclidean domains to $F[x]$. Thus, univariate polynomials over a field have gcds and lcms, which can be computed by the Euclidean algorithm, and Bezout's theorem (about $\gcd(a,b) = ua + vb$) holds for them. Also, this entails that $F[x]$ is a PID. See §4.3.6 in the text.

Note that even multivariate polynomial rings $F[x_1, x_2, \ldots, x_n]$ over a field are UFDs, so they have gcds and lcms, though they are not Euclidean domains! This is harder to prove and we won't get to it.

3.3.6. Lagrange interpolation

First, a corollary from the easy half of the FTA:

Corollary 3.3.19 (uniqueness of the interpolating polynomial)**.** Let R be an integral domain. Let $a_0, a_1, \ldots, a_n$ be $n+1$ distinct elements of R. Let $f, g \in R[x]$ be two polynomials of degree $\leq n$. Assume that

$$f[a_i] = g[a_i] \qquad \text{for all } i \in \{0, 1, \ldots, n\}.$$

Then, $f = g$.

Proof. The polynomial $f - g$ has degree $\leq n$, but it is 0 on all the $a_0, a_1, \ldots, a_n$. So $f - g$ has more roots than its degree. If $f - g$ was nonzero, this would contradict the easy half of the FTA. So $f - g$ is zero, and thus $f = g$. □

This corollary says that a polynomial of degree $\leq n$ over an integral domain is uniquely determined by its values at $n+1$ distinct elements $a_0, a_1, \ldots, a_n$ of

R. In other words, if you specify the values $f[a_0]$, $f[a_1]$, ..., $f[a_n]$, then the polynomial f is uniquely determined (assuming it has degree $\leq n$). But does such a polynomial f always exist (whatever choice of values)?

If R is a field, then the answer is "yes":

Theorem 3.3.20 (Lagrange interpolation)**.** Let F be a field. Let $n \in \mathbb{N}$.

Let $a_0, a_1, \ldots, a_n$ be $n+1$ distinct elements of F. Let $b_0, b_1, \ldots, b_n$ be $n+1$ elements of F. Then:

(a) There is a **unique** polynomial $p \in F[x]$ such that $\deg p \leq n$ and

$$p[a_i] = b_i \qquad \text{for all } i \in \{0, 1, \ldots, n\}.$$

(b) This polynomial p is explicitly given by

$$p = \sum_{j=0}^{n} b_j \frac{\prod\limits_{k \neq j} (x - a_k)}{\prod\limits_{k \neq j} (a_j - a_k)}.$$

Proof. See Theorem 4.3.26 in the text for details. In a nutshell: Define p by the formula in part **(b)**. Then, $\deg p \leq n$ and

$$p[a_i] = \sum_{j=0}^{n} b_j \underbrace{\frac{\prod\limits_{k \neq j} (a_i - a_k)}{\prod\limits_{k \neq j} (a_j - a_k)}}_{\substack{=0 \text{ when } j \neq i \\ \text{(because when } j \neq i\text{, then the} \\ \text{product in the numerator has a } k=i \\ \text{factor, which is } a_i - a_i = 0)}} = b_i \frac{\prod\limits_{k \neq i} (a_i - a_k)}{\prod\limits_{k \neq i} (a_i - a_k)} = b_i.$$

So p does fit the bill of part **(a)**. Remains to prove the uniqueness. But the above corollary does it for us. □

The theorem allows us to construct (or reconstruct) a polynomial of degree $\leq n$ over a field F from knowing $n+1$ of its values (at distinct inputs). This is called **Lagrange interpolation**. This is particularly useful when F is a finite field such as $\mathbb{Z}/p$ for a prime p. Here are two applications:

- **Shamir's Secret Sharing Scheme:** How do you "divide up" a piece of secret information among n "keepers" such that any k of the n "keepers" can reconstruct it but any $k-1$ cannot?

  Assume you have a secret **a**, and you want to distribute it among n people ("keepers") in such a way that any k keepers can reconstruct **a** (working together), but any $k-1$ cannot infer anything nontrivial about **a**.

How would you do this? Shamir's Secret Sharing Scheme does it as follows:

Fix a prime p such that $p > n$ and $p > 2^N$, where N is the size of **a** in bits.

Label the n keepers $1, 2, \ldots, n$.

Encode the secret **a** as a residue class $\alpha \in \mathbb{Z}/p$. (Note that the encoding algorithm and the p are not secret.)

Pick $k-1$ uniformly random elements $\beta_1, \beta_2, \ldots, \beta_{k-1}$ of $\mathbb{Z}/p$.

Let f be the polynomial

$$\beta_{k-1}x^{k-1} + \beta_{k-2}x^{k-2} + \cdots + \beta_1 x + \alpha \in (\mathbb{Z}/p)[x]$$

of degree $\leq k-1$.

Give one value of f to each keeper: namely, keeper i gets $f\left[\overline{i}\right] \in \mathbb{Z}/p$.

Lagrange interpolation allows any k keepers to reconstruct f and thus α and thus **a**.

For any $k-1$ keepers, the data they have is consistent with any possible value of α, since fixing the value of α is the same as fixing $f\left[\overline{0}\right]$, and that would provide just enough values for Lagrange interpolation.

- **Error-correcting codes:** How do you transmit data so that occasional errors in the transmission (noise, etc.) do not prevent the receiver from reconstructing the correct data?

  This is the main question of **coding theory** (including both storage media, like hard drives and RAM, and communication media, like telephone and radio).

  Here is a little taste of the subject:

  Imagine you want to send a message to a recipient via messenger pigeons. Each pigeon can carry an element of $\mathbb{Z}/p$ for a given prime p (more realistically, a bitstring of size n, but this can be reencoded into $\mathbb{Z}/p$). Your message is a tuple of n elements of $\mathbb{Z}/p$, so you could fit it onto n pigeons. But that's fragile: If any pigeon gets lost, then the receiver cannot recover your message.

  You can ameliorate this by adding redundancy into the system: Duplicating each message 3 times (so send $3n$ pigeons), then you can stomach the loss of 2 pigeons. But this is a bad deal: 3 times as many pigeons, but only 2 pigeons of loss tolerance.

  You want something better.

  One simple trick is "check-sums": If your message is the n-tuple $(a_1, a_2, \ldots, a_n)$, then you can send n pigeons with the entries a_i separately, and then an

extra "checksum" pigeon with the sum $a_1 + a_2 + \cdots + a_n \in \mathbb{Z}/p$. This allows the receiver to recover from the loss of any single pigeon:

$$a_i = (a_1 + a_2 + \cdots + a_n) - a_1 - a_2 - \cdots - a_{i-1} - a_{i+1} - \cdots - a_n.$$

Similarly, you can recover from 2 missing pigeons by sending two "checksum" pigeons, for example,

$$a_1 + a_2 + \cdots + a_n,$$
$$1a_1 + 2a_2 + \cdots + na_n$$

(if $n < p$). Things, however, get more complicated with more missing pigeons.

However, polynomials and Lagrange interpolation give a uniform solution. We encode our intended message into a polynomial

$$f := a_1 x^0 + a_2 x^1 + a_3 x^2 + \cdots + a_n x^{n-1} \in (\mathbb{Z}/p)[x],$$

and give each pigeon a value of this polynomial (say, pigeon i gets $f\left(\overline{i}\right)$). Any n values determine f uniquely, so the receiver can make do with receiving any n pigeons. So, if you send $n + r$ pigeons, then the receiver can recover from the loss of any r of them.

Moreover, this allows for some version of error correction (i.e., the receiver can recover from the corruption of any $\left\lfloor \frac{r}{2} \right\rfloor$ pigeons, which means that they arrive but transport wrong values).

This is an **error-correcting code** known as the **Reed–Solomon code** (going back to Reed and Solomon in 1960).

Several more codes are known; some textbooks are referenced in the notes (or Lecture 26 from 2023).

- There are theoretical applications, too.

  For example: Let p be a prime number. Consider the polynomial

  $$x^p - x \in (\mathbb{Z}/p)[x].$$

  All elements of $\mathbb{Z}/p$ are roots of this polynomial (by Fermat's Little Theorem). So our degree-p polynomial $x^p - x$ has p roots. No surprise.

  Now, let us tweak it a bit. Namely, consider the more sophisticated polynomial

  $$\begin{aligned} f &:= (x^p - x) - \prod_{u \in \mathbb{Z}/p} (x - u) \\ &= (x^p - x) - \underbrace{\left(x - \overline{0}\right)}_{=x} \left(x - \overline{1}\right) \left(x - \overline{2}\right) \cdots \left(x - \overline{p-1}\right). \end{aligned}$$

This polynomial f also has p roots (since any element of $\mathbb{Z}/p$ substituted for x in $\prod_{u\in\mathbb{Z}/p}(x-u)$ will yield 0). But its degree is $< p$ (since the x^p leading terms cancel when you take the difference). Since $\mathbb{Z}/p$ is an integral domain, this would contradict the easy half of the FTA if f was nonzero. So f is zero. Thus we have proved:

Proposition 3.3.21. Let p be a prime number. Then,

$$x^p - x = \prod_{u\in\mathbb{Z}/p}(x-u) \qquad \text{in the polynomial ring } (\mathbb{Z}/p)[x].$$

- Much can be gotten out of this proposition by comparing coefficients of powers of x. In particular, we can compare coefficients in front of x. Thus we find

$$\begin{aligned} -1 &= \left(-\overline{1}\right)\left(-\overline{2}\right)\cdots\left(-\overline{p-1}\right) = \prod_{u\in(\mathbb{Z}/p)^\times} u \\ &= \overline{1}\cdot\overline{2}\cdot\cdots\cdot\overline{p-1} = \overline{1\cdot 2\cdot\cdots\cdot(p-1)} = \overline{(p-1)!}. \end{aligned}$$

  Thus we recover Wilson's theorem $(p-1)! \equiv -1 \bmod p$.

- Another use of the easy half of the FTA is the following fact (see Proposition 4.3.20 in the text):

  Let p be a prime. Let $k \in \{0,1,\ldots,p-2\}$. Then, the sum

$$0^k + 1^k + \cdots + (p-1)^k = \sum_{j=0}^{p-1} j^k$$

  is divisible by p.

  Proposition 3.3.22.

3.4. Intermezzo: Quotients of R-algebras

A ring can be quotiented by an ideal. An R-module can be quotiented by a submodule.

So you shouldn't be surprised that an R-algebra can be quotiented by something as well. This "something" has to be an ideal that also happens to be an R-submodule at the same time. But it turns out that any ideal of an R-algebra

is an R-submodule. Indeed, if I is an ideal of an R-algebra A, then I is closed under scaling, since

$$ri = (r \cdot 1_A)\, i \in I \qquad \text{since } I \text{ is an ideal.}$$

Thus, we can quotient an R-algebra by any ideal:

Theorem 3.4.1. Let A be an R-algebra. Let I be an ideal of A. Then:
(a) The ideal I is also an R-submodule of A.
(b) The quotient ring A/I and the quotient R-module A/I fit together to form an R-algebra.
(c) The canonical projection $\pi : A \to A/I$ is an R-algebra morphism.

Proof. Straightforward. See the notes/text. □

Quotient R-algebras have a universal property:

Theorem 3.4.2 (Universal property of quotient algebras, elementwise form)**.** Let A be a R-algebra. Let I be an ideal of A.
Let B be an R-algebra. Let $f : A \to B$ be an R-algebra morphism. Assume that $f(I) = 0$ (that is, $f(i) = 0$ for each $i \in I$). Then, the map

$$\begin{aligned} f' : A/I &\to B, \\ \overline{a} &\mapsto f(a) \end{aligned}$$

is well-defined and is an R-algebra morphism.

Proof. Analogous to the ring case. □

3.5. Adjoining roots

3.5.1. A notation

Convention 3.5.1. Let S be any commutative ring, and let $a \in S$. Then, the quotient ring S/aS will be called S/a.

This generalizes the notation $\mathbb{Z}/n$ for $\mathbb{Z}/n\mathbb{Z}$.

We briefly recall that we think of a quotient ring R/I as "what becomes of R if we equate all elements of I with zero". Thus, S/a is "what becomes of S if we equate all multiples of a with zero". Of course, this is tantamount to equating just a with zero and drawing the obvious consequences (0·anything= 0).

3.5.2. Why adjoin roots

We now come to one of the most important applications of polynomials to algebra: a way to "adjoin" roots of a polynomial to a given commutative ring (i.e., to "create" roots out of thin air).

The classical example is the construction of complex numbers as "real numbers plus a root i of x^2+1". Cardano did essentially this in the 16th century (more in the text) by sheer power of fancy: He essentially said "let's pretend that there is some new number i such that $i^2=-1$ and play with it". This is a dangerous thing to do, because you could just as well introduce a new number ∞ such that $0\cdot\infty=1$, and then quickly "learn" that all your old numbers are equal ($1+1=0\cdot\infty+0\cdot\infty=(0+0)\cdot\infty=0\cdot\infty=1$ for example). So nowadays we prefer to define complex numbers as pairs of real numbers instead.

3.5.3. How to adjoin roots

Nevertheless, the ability to invent new numbers satisfying some desired equalities is a good power to have, and it would be nice if we could tell when such an invention is harmless (as opposed to collapsing existing numbers). So let us try to make it rigorous. What does it mean to introduce a new number?

The simplest case is when we want to introduce a new number x that satisfies no relations (other than the ring axioms). This just means we work in the polynomial ring $\mathbb{R}[x]$. Our "new number" here is just the indeterminate x.

Now, if we want our "new number" to satisfy some relations – let's say $x^2+x=15$ – then we can move over to the quotient ring $\mathbb{R}[x]/\left(x^2+x-15\right)$. This has the effect of equating x^2+x-15 to 0, thus equating x^2+x to 15. Our "new number" is then the residue class $\overline{x}\in\mathbb{R}[x]/\left(x^2+x-15\right)$.

In particular, Cardano's complex numbers are therefore the elements of $\mathbb{R}[x]/\left(x^2+1\right)$, with the imaginary unit i being the residue class $\overline{x}$.

This method generalizes to any commutative ring R instead of $\mathbb{R}$, and to an arbitrary polynomial $b\in R[x]$ that we want to "equate to 0". In general, if we start with a commutative ring R and a polynomial $b\in R[x]$, then the quotient ring $R[x]/b$ has an element $\overline{x}$ (the residue class of x) that is a root of b (since $b\left[\overline{x}\right]=\overline{b[x]}=\overline{b}=\overline{0}$ (because b is in the ideal we're quotienting by)). So we have "created" a root of b. This quotient ring $R[x]/b$ is not just a ring, but actually a commutative R-algebra.

Alas, as we said, this ring $R[x]/b$ might "collapse" some existing elements of R, in the sense that distinct elements of R could become equal in $R[x]/b$. So we cannot generally say that $R[x]/b$ is an "extension of R by a root of b"; in general, it is merely an R-algebra that contains a root of b. We will soon see some criteria for when this kind of "collapse" happens and when it doesn't.

3.5.4. Some examples

Let us first see this construction in some concrete cases.

As we said, Cardano's complex numbers are the elements of $\mathbb{R}[x] / \left(x^2+1\right)$, while modern complex numbers are pairs of real numbers $(a,b) = a + bi \in \mathbb{C}$. We hope that these two rings (actually $\mathbb{R}$-algebras) are isomorphic. This is indeed the case:

Proposition 3.5.2. We have

$$\mathbb{R}[x] / \left(x^2+1\right) \cong \mathbb{C} \qquad \text{as } \mathbb{R}\text{-algebras.}$$

Concretely: There is an $\mathbb{R}$-algebra isomorphism

$$\begin{aligned} \mathbb{R}[x] / \left(x^2+1\right) &\to \mathbb{C}, \\ \overline{p} &\mapsto p[i]. \end{aligned}$$

Proof. Here is a six-step procedure to prove this claim (and generally claims like this):

1. Give a putative definition of the alleged isomorphism.
2. Prove that this definition actually defines a map ("the map is well-defined").
3. Prove that this map is an $\mathbb{R}$-algebra morphism.
4. Prove that this map is injective.
5. Prove that this map is surjective.
6. Conclude that this map is an $\mathbb{R}$-algebra isomorphism.

Let us say a few words about how these six steps look like in our case.

Step 1 has already been done: Our map is defined to be the map

$$\begin{aligned} \mathbb{R}[x] / \left(x^2+1\right) &\to \mathbb{C}, \\ \overline{p} &\mapsto p[i]. \end{aligned}$$

Step 2: We must show that if $\overline{p} = \overline{q}$, then $p[i] = q[i]$.

Let $\overline{p} = \overline{q}$. Then, $p \equiv q \bmod \left(x^2+1\right)\mathbb{R}[x]$, which means that $p = q + \left(x^2+1\right) r$ for some polynomial $r \in \mathbb{R}[x]$. Then, with this r, we get

$$\begin{aligned} p[i] &= \left(q + \left(x^2+1\right) r\right)[i] \\ &= q[i] + \underbrace{\left(x^2+1\right)[i]}_{=i^2+1=0} \cdot r[i] \\ &= q[i]. \end{aligned}$$

So Step 2 is finished.

Step 3: We must show that the map

$$\begin{aligned} \mathbb{R}\left[x\right] / \left(x^2+1\right) &\to \mathbb{C}, \\ \overline{p} &\mapsto p\left[i\right]. \end{aligned}$$

is an $\mathbb{R}$-algebra morphism. This is a general property of evaluation morphisms.

(Actually, Steps 2 and 3 could be done in one swoop using the universal property of quotient R-algebras.)

Step 4 (injectivity): Our map is $\mathbb{R}$-linear. Hence, in order to prove that it is injective, it suffices to show that its kernel is $\{0\}$. In other words, we must show that if a polynomial $p \in \mathbb{R}\left[x\right]$ satisfies $p\left[i\right] = 0$, then $\overline{p} = 0$.

Let $p \in \mathbb{R}\left[x\right]$ be a polynomial such that $p\left[i\right] = 0$. By division with remainder, we can write p as $p = q \cdot \left(x^2+1\right) + r$ for some polynomials q, r with $\deg r < \deg\left(x^2+1\right)$. Of course, $\deg r < \deg\left(x^2+1\right) = 2$ means that r is linear, i.e., that $r = ax + b$ for some constants $a, b \in \mathbb{R}$. From $p = q \cdot \left(x^2+1\right) + r$, we obtain $\overline{p} = \overline{r}$, so that $p\left[i\right] = r\left[i\right]$ because our map is well-defined.

From $r = ax + b$, we obtain $r\left[i\right] = ai + b = (b, a)$ (viewing $\mathbb{C}$ as $\mathbb{R} \times \mathbb{R}$). Thus, from $r\left[i\right] = p\left[i\right] = 0$, we obtain $(b, a) = 0$, so that $a = b = 0$ and therefore $r = 0$. Hence, $\overline{p} = \overline{r} = 0$, qed.

(Essentially, we have argued here that every element of $\mathbb{R}\left[x\right] / \left(x^2+1\right)$ can be written as $\overline{r}$ for some **linear** polynomial r.)

Step 5 (surjectivity) is easy: We must show that every $z \in \mathbb{C}$ can be written as $p\left[i\right]$ for some $p \in \mathbb{R}\left[x\right]$. To do this, just write z as $a + bi$ with $a, b \in \mathbb{R}$, and take $p = a + bx$.

Step 6 is automatic. So the proposition is proved. □

Similarly:

Proposition 3.5.3. We have

$$\mathbb{Z}\left[x\right] / \left(x^2+1\right) \cong \mathbb{Z}\left[i\right] \qquad \text{as } \mathbb{Z}\text{-algebras.}$$

Concretely: There is a $\mathbb{Z}$-algebra isomorphism

$$\begin{aligned} \mathbb{Z}\left[x\right] / \left(x^2+1\right) &\to \mathbb{Z}\left[i\right], \\ \overline{p} &\mapsto p\left[i\right]. \end{aligned}$$

Proposition 3.5.4. We have

$$\mathbb{Q}\left[x\right] / \left(x^2+1\right) \cong \mathbb{Q}\left[i\right] \qquad \text{as } \mathbb{Q}\text{-algebras.}$$

Concretely: There is a $\mathbb{Q}$-algebra isomorphism

$$\begin{aligned}\mathbb{Q}[x] / \left(x^2+1\right) &\to \mathbb{Q}[i], \\ \overline{p} &\mapsto p[i].\end{aligned}$$

Proposition 3.5.5. We have

$$\mathbb{Q}[x] / \left(x^2-5\right) \cong \mathbb{Q}\left[\sqrt{5}\right] \qquad \text{as } \mathbb{Q}\text{-algebras.}$$

Concretely: There is a $\mathbb{Q}$-algebra isomorphism

$$\begin{aligned}\mathbb{Q}[x] / \left(x^2-5\right) &\to \mathbb{Q}\left[\sqrt{5}\right], \\ \overline{p} &\mapsto p[i].\end{aligned}$$

Proof. Similar, but use the irrationality of $\sqrt{5}$ to show that $a + b\sqrt{5} = 0$ entails $a = b = 0$ (when $a, b \in \mathbb{Q}$). □

Proposition 3.5.6. We have

$$\begin{aligned}\mathbb{Q}[x] / \left(x^2-1\right) &\cong \mathbb{Q}[C_2] && \text{(the group algebra of the cyclic group } C_2\text{)} \\ &\cong \left\{ \begin{pmatrix} a & b \\ b & a \end{pmatrix} \mid a, b \in \mathbb{Q} \right\} && \left(\text{a subring of } \mathbb{Q}^{2\times 2}\right) \\ &\cong \mathbb{Q} \times \mathbb{Q} && \text{(a direct product of two } \mathbb{Q}\text{s)}.\end{aligned}$$

All the above examples have a commonality: The quotient ring always had the form $R[x]/b$ where b is a non-constant polynomial whose leading coefficient is a unit (actually, 1 in our examples). This ensures that the division with remainder we used in our proof goes through. The non-constantness of b ensures that the resulting quotient ring $R[x]/b$ contains a copy of R as a subring. (To be proved below.)

If the leading coefficient is not a unit or b is constant, then the quotient ring $R[x]/b$ might contain only a collapsed version of R as a subring:

Proposition 3.5.7. For any integer m, we have $\mathbb{Z}[x]/m \cong (\mathbb{Z}/m)[x]$.
In particular, $\mathbb{Z}[x]/1 \cong (\mathbb{Z}/1)[x]$ is a trivial ring.

Here is a subtler example:

Proposition 3.5.8. Fix a nonzero integer m. Then,

$$\mathbb{Z}[x]/(mx-1) \cong \mathbb{Z}\left[\frac{1}{m}\right] = \left\{\frac{a}{m^i} \mid a \in \mathbb{Z} \text{ and } i \in \mathbb{N}\right\}.$$

This is not obvious (there is a proof sketch in the text)!

Let us summarize: If we have a commutative ring R and a polynomial $b \in R[x]$, then the quotient ring $R[x]/b$ is like "R with a root of b thrown in". The residue class $\overline{x}$ is a new root of b in $R[x]/b$. This ring $R[x]/b$ does **not always** contain a copy of R (for example, $1 \neq 3$ in $\mathbb{Z}$ but $1 = 3$ in $\mathbb{Z}[x]/2$), but it is always a commutative R-algebra, and thus contains the subring $\{\overline{r} \mid r \in R\}$ which is isomorphic to a quotient of R.

Cardano was lucky: In his case, the ring $\mathbb{R}[x]/(x^2+1)$ really does contain a copy of $\mathbb{R}$, and its elements can be encoded as pairs (a,b) of two real numbers. This means that, as an $\mathbb{R}$-vector space, it has a basis $(\overline{1}, \overline{x})$. In general, as an R-module, $R[x]/b$ will not always be free.

As a ring, $R[x]/b$ is not always as nice as R. For example, $\mathbb{Q}$ is a field, but $\mathbb{Q}[x]/(x^2-1)$ is not even an integral domain (it is $\cong \mathbb{Q} \times \mathbb{Q}$).

This method of creating roots of polynomials b by passing from R to $R[x]/b$ is called **root adjunction**, or **adjoining a root of b to R**.

So we might want some criteria for when root adjunction behaves nicely: When do we get a field? When do we get a free R-module? When do we get a copy of R inside $R[x]/b$?

Before answering these questions (by some sufficient criteria), let me state some basics:

Proposition 3.5.9. Let $b \in R[x]$ be a polynomial.

(a) The projection map

$$\pi : R[x] \to R[x]/b, \quad p \mapsto \overline{p}$$

is an $R[x]$-algebra morphism, hence an R-algebra morphism.

(b) Its restriction

$$\pi|_R : R \to R[x]/b, \quad r \mapsto \overline{r}$$

is an R-algebra morphism.

(c) For any $p \in R[x]$, we have $p[\overline{x}] = \overline{p}$ in $R[x]/b$.

(d) The element $\overline{x} \in R[x]/b$ is a root of b.

Proof. Easy. (Proposition 4.5.7 in the text.) □

Now, here are the promised criteria for niceness of $R[x]/b$:

Theorem 3.5.10. Let $m \in \mathbb{N}$. Let $b \in R[x]$ be a polynomial of degree m whose leading coefficient $[x^m]\, b$ is a unit of R. Then:

(a) Each element of $R[x]/b$ can be uniquely written in the form

$$a_0\overline{x^0} + a_1\overline{x^1} + \cdots + a_{m-1}\overline{x^{m-1}} \qquad \text{with } a_0, a_1, \ldots, a_{m-1} \in R.$$

(b) The m vectors $\overline{x^0}, \overline{x^1}, \ldots, \overline{x^{m-1}}$ form a basis of the R-module $R[x]/b$. In particular, this R-module is free of rank m.

(c) If $m > 0$, then the R-algebra morphism

$$\begin{aligned} R &\to R[x]/b, \\ r &\mapsto \overline{r} \end{aligned}$$

is injective, and thus R can be viewed as an R-subalgebra of $R[x]/b$ (by identifying each $r \in R$ with $\overline{r} \in R[x]/b$).

(d) Thus, under the assumption that $m > 0$, there exists a commutative ring that contains R as a subring and that contains a root of b.

Proof. Again, see the text (Theorem 4.5.9). Part **(a)** reduces to division with remainder, and the other parts follow from it. □

OK, so we know how to adjoin a root of a polynomial to a commutative ring, and we can ensure that the resulting ring will contain a copy of R as a subring if our polynomial is non-constant and has its leading coefficient be a unit. This covers most cases we care about, such as Cardano's $\mathbb{R}[x] / \left(x^2+1\right)$. When the leading coefficient is not a unit, we don't get a basis any more, but the quotient can still be well-behaved depending on other things.

When does a root adjunction to a field produce a field? It does for $\mathbb{Q}[x] / \left(x^2+1\right)$ but it does not for $\mathbb{Q}[x] / \left(x^2-1\right)$. More complicated examples are $\mathbb{Q}[x] / \left(x^4+3x^3+2x^2+x+5\right)$ (a field) and $\mathbb{Q}[x] / \left(x^4+2x^3+4x^2+3x+2\right)$ (not a field). In general:

Theorem 3.5.11. Let F be a field. Let $b \in F[x]$ be a nonzero polynomial. Then, $F[x]/b$ is a field if and only if b is irreducible (= not a product of two non-constant polynomials, and not itself constant).

Proof. This is the polynomial analogue of the fact that $\mathbb{Z}/n$ is a field if and only if n is prime (or minus a prime, or 0 or 1 or -1). More generally: If R is any PID, then a quotient ring R/b by a nonzero element $b \in R$ is a field if and only if b is prime in R. (If R is Euclidean, you can prove this exactly as we proved the analogous property of $\mathbb{Z}/n$.) □

Thus, we can adjoin a root of any irreducible polynomial to a field and get a field.

If we do this multiple times, we can obtain a field in which our polynomial factors into linear factors:

Corollary 3.5.12. Let F be a field. Let $b \in F[x]$ be a monic polynomial. Then, there exists a field G that contains F as a subfield and such that b can be factored as

$$b = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n) \qquad \text{for some } \lambda_1, \lambda_2, \ldots, \lambda_n \in G.$$

Proof. Use root adjunction repeatedly. Make sure to only apply root adjunction to irreducible polynomials (so if b is not irreducible, just factor it and adjoin a root of one of the factors). □

Such a field G is called a **splitting field** for b. So we have shown that any monic polynomial over a field has a splitting field. Thus, for an algebraist, roots of a polynomial can always be summoned at will. This means that the hard part of the FTA is rarely ever necessary in algebra: You can just get your roots by root adjunction.

This also allows us to construct finite fields different from $\mathbb{Z}/p$. For example, you can get a finite field of size $2^5 = 32$ by adjoining a root of an irreducible polynomial of degree 5 to $\mathbb{Z}/2$. (For this, you need to find such a polynomial, but this can be done by brute force.)

In the text, I show that for any prime power p^m, there is an irreducible polynomial of degree m over $\mathbb{Z}/p$, and this allows you to produce a finite field of size p^m by root adjunction.