Modern Algebra

John Perry

December 7, 2019

Contents

List of Symbols

$\equiv$	the congruence relation
$c \mid a$	c divides a
$\backslash$	set difference: $S\backslash T$ consists of the elements of S that do not appear in T
$\subseteq$	subset relation: $S \subseteq T$ if every element of S is in T
$\not\subseteq$	not a subset relation: $S \not\subseteq T$ if at least one element of S is not in T
$\subsetneq$	proper subset relation: $S \subsetneq T$ if every element of S is in T, but some elements of T are not in S
$\in$	element relation: for a set S, the sentence $s \in S$ states that s is an element of S
$\cap$	set intersection: $S \cap T$ consists of the elements that appear in both S and T
$\mathbb{N}^2$	the lattice of natural numbers: ordered pairs (a, b) such that a and b are both natural
$\mathbb{N}$	the set of natural numbers, $\{0, 1, 2, \ldots\}$
$\mathbb{N}^+$	the set of positive integers, $\{1, 2, \ldots\}$
$\phi(m)$	the number of integers between 1 and m that are relatively prime to m
$\mathbb{Q}$	the set of rational numbers, $\{a/b : a, b \in \mathbb{Z}$ and $b \neq 0\}$
$\mathbb{Q}[x]$	the set of polynomials with rational coefficients
$\langle r_1, \ldots, r_m \rangle$	the ideal generated by $r_1, \ldots, r_m$
$\mathbb{R}$	the set of real numbers
$\mid S \mid$	the size of the set S
$\cup$	set union: $S \cup T$ consists of the elements in at least one of S or T
$\times$	Cartesian product: $S \times T$ consists of ordered pairs (s, t) such that $s \in S$ and $t \in T$
$\mathbb{Z}$	the integers, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Z}_m$	the set of remainders after division by m, with arithmetic modulo m

$\mathbb{Z}_m^*$	the subset of $\mathbb{Z}_m$ whose elements are relatively prime to m
$\mathbb{Z}[x]$	the set of polynomials with integer coefficients

Chapter 1

Modular arithmetic

1.1 Sets and relations

The material in this section is fundamental! Much of mathematics depends entirely on the definitions.

Sets

A ***set*** is a collection of ***element***s. From now until the end of the text, S is a set, unless we write otherwise. We write $s \in S$ to say that "s is an element of S," and we write $t \notin S$ to say that "t is not an element of S." If S has finitely many elements, then the ***size*** of S is the number of elements. We write $|S|$ for the size of S.

- If we only consider one or two elements of S, then we'll start with $s \in S$, then consider $t \in S$. Similarly we may start with $a \in A$, then consider $b \in A$.

- Sometimes, when we consider several elements of S, we'll start with $s, t \in S$, then write subsequent elements in a "decorated" fashion, as $\hat{s}$, s', $\hat{t}$, t', and so forth. The hat and the apostrophe have no special meaning. They just means that we're looking at a different element of S. The apostrophe does *not* indicate a derivative! That said, s, $\hat{s}$, and s' will typically be related in some fashion.

- For a long sequence of elements, we'll write the first one as s_0, the next one as s_1, the one after that as s_2, and so forth. We call the number a *subscript* and sometimes we'll use a letter when we don't necessarily know *which* element of the sequence we mean; for instance, s_i is the ith element in $s_1, s_2, \ldots$.

Example 1.1. Suppose $S = \{1, 2, 3, 4, 5\}$. We see that1 $\in S$, but $0 \notin S$. For any $s, t \in S$ we know that $s + t \in \{2, 3, \ldots, 10\}$ and $s - t \in \{-4, -3, \ldots, 3, 4\}$.

The basic sets of school mathematics are[1]

- the ***positive numbers*** $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$;

[1] This is not really important, but $\mathbb{Z}$ is from the German word for "number" and $\mathbb{Q}$ is from the Italian word for "quotient."

- the ***natural numbers*** $\mathbb{N} = \{0, 1, 2, \ldots\}$;
- the ***integers*** $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$;
- the ***rational numbers*** $\mathbb{Q} = \{a/b : a, b \in \mathbb{Z} \text{ and } b \neq 0\}$;
- the ***real numbers*** $\mathbb{R}$, which we can describe intuitively as "any length of a line segment," written $\mathbb{R}$ for short.

Whenever every element a set S is also an element of a set T, we say that S is a *subset of* T, written $S \subseteq T$ for short. If S is not a subset of T, then we write $S \not\subseteq T$. We say that two sets are *equal* if each is a subset of the other, written $S = T$ for short. If we know that $S \subseteq T$ but $S \neq T$, we write $S \subsetneq T$.

Notice the difference between $S \subsetneq T$ and $S \not\subseteq T$; in the first case, S is a subset of T; it's just not equal to it; in the second, S is not even a subset of T.

Example 1.2. Every natural number is an integer, allowing us to write $\mathbb{N} \subseteq \mathbb{Z}$. The negatives are not natural, so $\mathbb{N} \subsetneq \mathbb{Z}$.

Example 1.3. Every rational number of the form $a/1$ is in identified with the integer a, so every integer is a rational number, allowing us to write $\mathbb{Z} \subseteq \mathbb{Q}$. Some rationals, like $1/2$, are not integers, so $\mathbb{Q} \subsetneq \mathbb{Z}$ and so $\mathbb{Q} \neq \mathbb{R}$.

Example 1.4. Is every rational number a real number? Let $a/b \in \mathbb{Q}$, and take any line segment you like. Copy it b times and lay the b copies end-to-end; we will say that the resulting line segment has length "1 unit," so that the original segment had length "$1/b$ units." Now copy the original segment a times and lay the a copies end-to-end; the resulting line segment has length "a/b units." Thus, a/b is the length of a line segment, and hence it is a real number.

We placed no particular conditions on a/b, which means it was "arbitrary" in $\mathbb{Q}$. By working with an arbitrary element of $\mathbb{Q}$, we have shown that *every* rational number corresponds to the length of a line, allowing us to write $\mathbb{Q} \subseteq \mathbb{R}$.

Example 1.5. Is every real number a rational number? We won't answer this question quite yet, but propose a thought experiment instead.

Consider $\sqrt{2}$. You will show in Exercise 1.15 that $\sqrt{2}$ is the length of a line segment, so $\sqrt{2}$ is definitely a real number. Can we also write $\sqrt{2} = a/b$, where a and b are integers?

If so, *and* if we can do this for every real number, then $\mathbb{R} \subseteq \mathbb{Q}$. This is a bit hard at the moment, so we postpone it until Section 1.4.

We often define sets using ***set-builder notation***. For instance, we defined the rational numbers as

$$\mathbb{Q} = \{a/b : a, b \in \mathbb{Z} \text{ and } b \neq 0\} \ .$$

This reads as, "$\mathbb{Q}$ is the set of all a/b such that a and b are integers and b is not 0."

There are three common ways to build one set from two others.

- The ***union*** of S and T is the set of elements that are members of S or T,[2] written $S \cup T$. In set-builder notation,
$$S \cup T = \{x : x \in S \text{ or } x \in T\} \ .$$
- The ***intersection*** of S and T is the set of elements that are members of S and T, written $S \cap T$. In set-builder notation,
$$S \cap T = \{x : x \in S \text{ and } x \in T\} \ .$$
- The ***difference*** of S and T is the set of elements that are members of S but not of T,[3] written $S\backslash T$. In set-builder notation,
$$S\backslash T = \{x : x \in S \text{ and } x \notin T\} \ .$$

Example 1.6. Can we simplify the expression, $\mathbb{N} \cup \mathbb{Z}$?

To answer this, consider an arbitrary element $a \in \mathbb{N} \cup \mathbb{Z}$. By definition of union, $a \in \mathbb{N}$ or $a \in \mathbb{Z}$. If $a \in \mathbb{N}$, then $a \in \mathbb{Z}$, as well. Hence $\mathbb{N} \cup \mathbb{Z} \subseteq \mathbb{Z}$.

On the other hand, consider an arbitrary $b \in \mathbb{Z}$. By definition of union, $b \in \mathbb{N} \cup \mathbb{Z}$, as well. Hence $\mathbb{N} \cup \mathbb{Z} \supseteq \mathbb{Z}$.

We said above that two sets are equal if each is a subset of the other. We have now shown that $\mathbb{N} \cup \mathbb{Z} \subseteq \mathbb{Z}$ and $\mathbb{N} \cup \mathbb{Z} \supseteq \mathbb{Z}$, so $\mathbb{N} \cup \mathbb{Z} = \mathbb{Z}$.

Example 1.7. Can we simplify the expression, $\mathbb{N} \cap \mathbb{Z}$?

Again, consider an arbitrary element $a \in \mathbb{N} \cap \mathbb{Z}$. By definition of intersection, $a \in \mathbb{N}$ and $a \in \mathbb{Z}$. If $a \in \mathbb{N}$, then $a \in \mathbb{Z}$, as well. Hence $\mathbb{N} \subseteq \mathbb{N} \cap \mathbb{Z}$.

On the other hand, consider any $b \in \mathbb{Z}$ such that $b \notin \mathbb{N}$. By definition of intersection, $b \notin \mathbb{N}\cap\mathbb{Z}$, either. Hence $\mathbb{N} \supseteq \mathbb{N} \cap \mathbb{Z}$.

We said above that two sets are equal if each is a subset of the other. We have now shown that $\mathbb{N} \cup \mathbb{Z} \subseteq \mathbb{Z}$ and $\mathbb{N} \cup \mathbb{Z} \supseteq \mathbb{Z}$, so $\mathbb{N} \cup \mathbb{Z} = \mathbb{Z}$.

Example 1.8. Can we simplify the expression, $\mathbb{Z}\backslash\mathbb{N}$?

Once again, consider an arbitrary element $a \in \mathbb{Z}\backslash\mathbb{N}$. By definition of set difference, $a \in \mathbb{Z}$ but $a \notin \mathbb{N}$. The only numbers that satisfy that are negative numbers, so $\mathbb{Z}\backslash\mathbb{N} = \{-1, -2, -3, \ldots\}$.

You will generalize these results in Exercise 1.16.

[2] In common English, the word “or” typically means “either-or,” or “exclusive-or.” For example, most people would understand the phrase, “Would you like cake or pie?” to mean choosing one or the other. In mathematics, however, the word “or” is an “inclusive-or,” so that a mathematician understands that the only correct answer to, “Would you like cake or pie?” is “Yes.”

[3] As the set-builder notation shows, “but” and “and” have the same logical meaning in mathematics, and one can usually interchange them. In common English, “but” and “and” are not typically interchangeable: “You can have cake and not pie” sounds wrong in far too many ways.

Relations

From here until the end of the section, T is also a set.

The ***Cartesian product*** of S and T is the set of all ordered pairs whose first entry is an element of S and whose second entry is an element of T. Written symbolically,

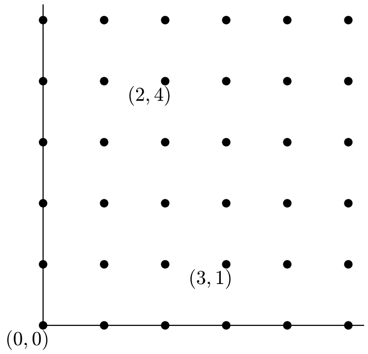
$$S \times T = \{(s, t) : s \in S, t \in T\} \ .$$

Example 1.9. The Cartesian product of $\mathbb{N}$ with itself is

$$\mathbb{N} \times \mathbb{N} = \{(0,0), (0,1), (0,2), \ldots, (1,0), (1,1), (1,2), \ldots\} \ .$$

In a case like this we will write $\mathbb{N}^2$ instead of $\mathbb{N} \times \mathbb{N}$. We call $\mathbb{N}^2$ the *lattice* of natural numbers.

An interesting aspect of the lattice of natural numbers is that you can plot its elements rather easily using the first quadrant of the real plane, typically drawn on a grid:



The only points allowed on the lattice are the ones marked with dots. Unlike an ordinary graph, *no points lie between them!* We have diagrammed the lattice points (0, 0), (2, 4), and (3, 1).

A ***relation*** *R between S and T* is a subset of $S \times T$. Given a relation R, we say that $s \in S$ and $t \in T$ are related if $(s, t) \in R$. However, it is more common to write relations differently: we typically use symbols such as $\leq$, $\sim$, $\equiv$, and write $s \leq t$ or $s \sim t$ or $s \equiv t$.

If $S = T$ then we call R a *relation on S*.

Example 1.10. Consider the relation on $\mathbb{N}$

$$R = \left\{(a, b) \in \mathbb{N}^2 : b - a \in \mathbb{N}^+\right\} \ .$$

Elements of this set include

$$(0,1), \quad (0,2), \quad (2,4), \quad (8,15)$$

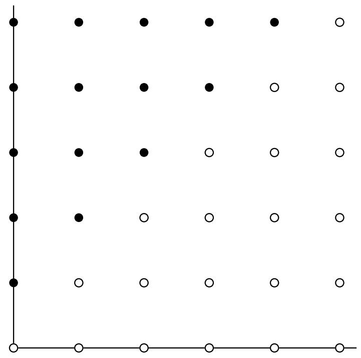
but *not* elements of the form

$$(0,0), \quad (2,0), \quad (-3,7) \ .$$

You are more likely to think of this as the *less-than* relation:

$$0 < 1, \quad 0 < 2, \quad \ldots \quad 1 < 2, \quad 2 < 3, \quad \ldots \quad 2 < 3, \quad 3 < 4, \quad \ldots \ .$$

We can diagram this relation on the lattice: dots indicate points (a, b) that belong to the relation because $b - a \in \mathbb{N}^+$ (that is, $a < b$); circles indicate points (a, b) that don't belong to the relation because $b - a \notin \mathbb{N}^+$.



Example 1.11. Here's another relation on $\mathbb{N}$:

$$R = \{(0,0), (1,1), (2,2), (3,3), \ldots\} \ .$$

What familiar relation are you looking at? How would you diagram it?

Many relations belong to an important class of relations called ***equivalence relations***. Equivalence relations satisfy three important properties. To describe them, we need a set S and a relation $\sim$ on S.

- The *reflexive* property states that $s \sim s$ for every $s \in S$.
- The *symmetric* property states that for every $s, t \in S$ if $s \sim t$, then $t \sim s$.
- The *transitive* property states that for every $s, t, u \in S$ if $s \sim t$ and $t \sim u$, then $s \sim u$.

Example 1.12. Look back at Example 1.10. It does *not* satisfy the reflexive property, because $0 \not< 0$, or, $(0,0) \notin R$. It is therefore not an equivalence relation. It also does not satisfy the symmetric property, because $0 < 1$ but $1 \not< 0$, or, $(0,1) \in R$ but $(1,0) \notin R$. On the other hand, it does satisfy the transitive property, because if $a < b$ and $b < c$ then $a < c$, or, if $(a,b), (b,c) \in R$ then $(a,c) \in R$.

Example 1.13. Look back at Example 1.11. You hopefully noticed that R is really the equality relation: every element of R has the form (a,b) where $a = b$. We rely on this to show that R is an equivalence relation.

- Let $a \in \mathbb{N}$. We see that $(a,a) \in R$, so R is symmetric.
- Let $a, b \in \mathbb{N}$. If $(a,b) \in R$, then $a = b$, but we know that $b = a$, so $(b,a) \in R$.
- Let $a, b, c \in \mathbb{N}$. If $(a,b), (b,c) \in R$, then $a = b$ and $b = c$, so $a = c$, in which case $(a,c) \in R$.

Exercises

Exercise 1.14. For the sets $S = \{1, 2, 3\}$ and $T = \{2, 3, 4\}$, compute the following sets.

(a) $S \cup T$

(b) $S \cap T$

(c) $S \backslash T$

(d) $S \times T$

Exercise 1.15. Use the intuitive definition of real number as the "length of a line segment" to explain how we know that $\sqrt{2}$ is a real number. The Pythagorean Theorem will help.

Exercise 1.16. Suppose $S \subseteq T$. Explain why $S \cup T = T$ and $S \cap T = S$.
Hint: Use Examples 1.6 and 1.7 as your guide.

Exercise 1.17. Consider the relation $R = \left\{(a, b) \in \mathbb{N}^2 : ab = 12\right\}$. List the elements of R, and diagram them on a lattice. Do you see a pattern?

Exercise 1.18. Consider the relation $R = \left\{(a, b) \in \mathbb{N}^2 : b - a \notin \mathbb{N}\right\}$. List ten elements of R, and diagram them on a lattice. Do you see a pattern? What familiar symbol does this relation represent?

Exercise 1.19. Define a relation on $\mathbb{Q}$ in the following way. For any $a/b, c/d \in \mathbb{Q}$, we say that $a/b \sim c/d$ if $ad = bc$.

(a) Show that $4/6 \sim 6/9$.

(b) Show that $-2/5 \sim 10/-25$.

(c) Show that $a/b \sim a/b$.

(d) Show that if $a/b \sim c/d$, then $c/d \sim a/b$.

(e) Show that if $a/b \sim c/d$ and $c/d \sim e/f$, then $a/b \sim e/f$.

(f) Is $\sim$ an equivalence relation?

(g) You have already used this equivalence relation many, many times before. Where?

Remember to use the *definition* of the relation in each part! If you aren't rewriting fractions as multiplication, then you aren't doing it right!

Exercise 1.20. The *Euclidean distance* between $(a, b), (c, d) \in \mathbb{N}^2$ is the value determined by the ordinary distance formula, $\sqrt{(a - c)^2 + (b - d)^2}$. Unfortunately, this is not usually a natural number.

It is possible to compute distance a different way, so that it is a natural number. In this case we consider the distance from (a, b) to (c, d) to be $|(a - c)| + |(b - d)|$. We'll call this the *sidewalk* distance between two points, because it indicates the number of sidewalks you'd have to travel from point (a, b) to point (c, d) when walking through a city laid out with perpendicular streets.

(a) Compute the sidewalk distance between $(0,0)$ and $(3,4)$.

(b) Compute the sidewalk distance between $(1,5)$ and $(7,2)$.

(c) Make a lattice diagram of all the points whose sidewalk distance to $(4,5)$ is at most 4.

(d) Make a lattice diagram of all the points whose Euclidean distance to $(4,5)$ is at most 4.

(e) Comment on how the answers to (c) and (d) are different.

Sage supplement

This section shows how to perform some elementary operations in Sage.

Some sets are already defined in Sage. For instance, you can type `NN`, `ZZ`, and `QQ` to obtain the sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$. If you type them into Sage, it will display a funny name:

```
sage: NN
Non negative integer semiring
sage: ZZ
Integer Ring
sage: QQ
Rational Field
```

Don't worry too much about the names; we explain later what the names "ring" and "field" mean.

You can define a set using either braces `{}` or the `set()` command.

```
sage: { 3, 5, 7 }
set([3, 5, 7])
sage: set( [ 7, 5, 3, 7, 7, 3, 5 ] )
set([3, 5, 7])
```

Notice how elements are automatically ordered, and no element can appear more than once.

It is also possible to define a set using something akin to set-builder notation, making use of `for` and `if` statements:

```
sage: { i^2 for i in { 3, 5, 7 } }
set([9, 49, 25])
sage: { i^2 for i in { 3, 4, 5, 6, 7 } if is_even( i ) }
set([16, 36])
```

Notice how all three numbers were squared in the first assignment, and how only the even numbers were taken and squared in the second assignment.

Another useful command for generating a set is `range()`.

- With just one integer between the parentheses, it returns a list of all the numbers that are between 0 and the specified number, including 0 but not the specified number.
- With two integers between the parentheses, it returns a list of all the numbers between the two, including the first but not the last.

```
sage: range( 8 )
[0, 1, 2, 3, 4, 5, 6, 7]
sage: range( 3, 8 )
[3, 4, 5, 6, 7]
```

You can assign a name to an object using the = operator. You can assign a name to any object in this way. You can even assign several objects at a time.

```
sage: a, b = 3, 4
sage: a + b
7
sage: S = { a, b, 7, 3 }
sage: S
set([3, 4, 7])
```

Notice that Sage does not display any messages after a successful assignment.

You can test whether two objects are equal using the == operator. This is *not* the same as the = operator; comparison uses two equality signs instead of one. Be careful when doing this; if you type only one sign, you may accidentally overwrite an object. In other cases, Sage will report an error. What happens depends on the context.

```
sage: 3 + 3 == 6
True
sage: 3 + 3 = 6
Error in lines 1-1
Traceback (most recent call last):
  File "/cocalc/lib/python2.7/site-packages/
smc_sagews/sage_server.py", line 1188, in execute
flags=compile_flags) in namespace, locals
  File "<string>", line 1
SyntaxError:  can’t assign to operator
```

Simple set operations are possible: use ***methods*** to accomplish them. Another name for a method is a “dot command”, because you access them by typing an identifier, followed by a dot, follow by the method name. The example below demonstrates the use of `.intersection`, `.union`, and `.difference`.

```
sage: S = { i^2 for i in range(20) if is_even(i) }
sage: T = { 4*i for i in range(100) }
sage: S.intersection(T)
set([0, 64, 4, 16, 256, 144, 196, 324, 36, 100])
sage: S.union(T)
set([0, 256, 392, 4, 8, ...  252])
sage: S.difference(T)
set([])
```

That last output is how Sage indicates that it has computed an empty set.

You can often discover commands available for an object by typing the object's name, adding a period, then pressing the "tab" key. If you do this with S, the version of Sage I am using will return 17 commands:

> add, clear, copy, difference, difference_update, discard, intersection, intersection_update, isdisjoint, issubset, issuperset, pop, remove, symmetric_difference, symmetric_difference_update, union, update

You should see commands for union, intersection, and difference. To learn more about a command, you can continue by typing the command after the period, followed a question mark, then executing the line, Sage will give you some help on the command. If I do this with pop, the version of Sage I am using will display the following:

```
sage: S.pop?
File:
Docstring :  Remove and return an arbitrary set element.
Raises KeyError if the set is empty.
```

This gives you an idea of what happens when you pop an element from a set.

You may have noticed that set objects lack a method for a Cartesian product. A command to compute Cartesian products actually exists; it just isn't a method. Try this:

```
sage: CP = cartesian_product((S,T))
sage: CP
The Cartesian product of ({0, 64, 4, 100, 324, 144, 256, 16,
196, 36}, {0, 256, 4, 8, ...  136, 252})
```

Notice that we used double parentheses; the cartesian_product command expects as input *one* argument, which is a pair or list of sets. Here we used parentheses to give a pair.

It may not seem especially useful to have a "Cartesian product" that displays itself only as a "Cartesian product" and not as a set of points, but trust us when we say that it is *very* useful.[4] In any case, we can use a set builder to transform CP into a set of ordered pairs: (that's an x between the S and the T below)

[4] Explaining why is beyond the scope of these notes.

```
sage: SxT = { P for P in CP }
sage: SxT
set([(100, 36), (100, 324), (4, 264), ...  (16, 328), (36,
76), (324, 352)])
```

(We omitted a *lot* of output this time.)

It is possible to define new commands in Sage. This is not a textbook on programming; we assume that if you are reading this, then you have some experience with programming, so we won't delve into the details of how various control structures work, but the following will define a command that computes the sidewalk distance between two points of the lattice, described in Exercise 1.20.

```
sage: def sidewalk_dist(P, Q):
        return abs(P[0] - Q[0]) + abs(P[1] - Q[1])
```

The `def` keyword defines a new command, in this case named `sidewalk_dist`. Parentheses always follow, and contain arguments that `sidewalk_dist` requires. In this case, it requires two arguments, `P` and `Q`.[5] The first line ends with a colon, and subsequent lines are indented; these two signals indicate that the indented lines depend on the line that ends with a colon. You will see this in all Sage's control structures.

The procedure then computes $|p_0 - q_0| + |p_1 - q_1|$, where p_0 is the first entry of P and p_1 is its second entry; we use `abs(...)` to compute the absolute value of whatever is in parentheses. The `return` command indicates that whatever follows on that line is the result of `sidewalk_dist`.

Once we define the command, we can use it as follows.

```
sage: sidewalk_dist((3,5),(7,2))
7
```

What just happened inside the computer? First it assigned the values $(3, 5)$ to `P` and $(7, 2)$ to `Q`. It then computed

- `abs( P[0] - Q[0] )` $= |3 - 7| = 4$;
- `abs( P[1] - Q[1] )` $= |5 - 2| = 3$;
- the sum of these numbers, 7;

and then returned the result.

Exercises

Exercise 1.21. Use Sage to verify your answers in Exercise 1.14.

Exercise 1.22. Use Sage to verify your answers to Exercise 1.20.

[5] Sage is like Python 2 or Perl, and unlike Python 3, in that it does not allow you to specify an argument's type of an argument. It is very much unlike C or Java, where you *must* specify the type.

1.2 Integer division

Division is a useful tool, but it is also rather strange. To understand why, we first need to make some ideas precise.

A ***function*** *from a set S to a set T* is a relation F between S and T such that if $(a, b), (a, c) \in F$ then $b = c$. That is, each “input” to F can have only one “output”. It is customary to write $F(a) = b$ instead of $(a, b) \in F$.

Operations and properties

An ***operation*** *on a set S* is a function from S^2 to S. It is customary to give an operation a symbol such as $\diamond$, so that instead of saying $((s, t), u)$ is in the operation, we say $s \diamond t = u$.

Example 1.23. Given two integers, we add them to obtain a third integer. Addition of integers is thus a function from $\mathbb{Z}^2$ to $\mathbb{Z}$. For instance, if we start with 2 and -5, we add them to obtain -3. This corresponds to the point $((2, -5), -3)$. However, we usually write $2 + (-5) = -3$ instead.

The most useful operations satisfy certain properties that we often take for granted:

- An operation on S is ***closed*** if for every $s, t \in S$ the operation’s result is some $u \in S$; that is, $s \diamond t = u$.
- An operation on S is ***commutative*** if for every $s, t \in S$ the order of the elements doesn’t matter; that is, $s \diamond t = t \diamond s$.
- An operation on S is ***associative*** if for every *three* elements $s, t, u \in S$ it doesn’t matter which *two* elements you first apply the operation to; that is, $(s \diamond t) \diamond u = s \diamond (t \diamond u)$.
- An operation on S has an ***identity*** if we can find some $z \in S$ such that for every $s \in S$ $s \diamond z = s$ and $z \diamond s = S$. In this case, we call z an identity of S.
- An operation on S is ***invertible*** if
  - it has an identity z; *and*
  - for any $s \in S$ we can find $t \in S$ such that $s \diamond t = z$ and $t \diamond s = z$.

  In this case, we call t the ***inverse*** of s.

Example 1.24.

- Addition is an operation on $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.
  - In fact, it is closed on all four sets.
  - It is also both commutative and associative on all four sets.
  - It has an identity on all four sets, written 0.
  - It is invertible on all four sets: given x, we write its inverse as $-x$. (Read that as *the opposite of x*, not as *negative x*. After all, $-(-5)$ is not negative.)

- Subtraction is an operation on $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.
    - However, it is not closed on $\mathbb{N}$, since $3 - 4 \notin \mathbb{N}$. It is closed on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.
    - Subtraction is generally neither commutative nor associative.
    - Subtraction does *not* have an identity! Even though $3 - 0 = 3$, we have $0 - 3 \neq 3$. The identity property requires *both* arrangements.
    - Since subtraction has no identity, the queston of whether it is invertible *makes no sense.*
- Multiplication is an operation on $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.
    - In fact, it is closed on all four sets.
    - It is both commutative and associative on all four sets.
    - It has an identity on all four sets. What is it?
    - It is *not* invertible on *any* of the sets, as 0 has no inverse! Even if we exclude 0, the first two sets remain un-invertible, since 2 has no *multiplicative* inverse in either $\mathbb{N}$ or $\mathbb{Z}$. However, multiplication is invertible on the nonzero rationals, $\mathbb{Q}\setminus\{0\}$, and the nonzero reals, $\mathbb{R}\setminus\{0\}$.
- Division is an operation on $\mathbb{Q}$ and $\mathbb{R}$. On $\mathbb{Q}$, for instance, you know that $a/b \div c/d = ad/bc$ so long as $b, c \neq 0$. However, things are not so simple on $\mathbb{N}$ and $\mathbb{Z}$… and that complication is the point of the rest of this chapter.

A Division Algorithm

What is division, and how do we accomplish it? The basic idea is that, given a set of n elements, we would like to divide it into subsets of d elements. We may not be able to do this perfectly, in which case we'll be greedy and take as many subsets as we can, and identify the number of objects left over as the *remainder*.

Example 1.25. Given a set of $n = 51$ elements, we can divide it into 8 sets of $d = 6$ elements, with a remainder of 3.

How do we actually do this? To make it concrete, suppose we have a pile 51 cookies, and we want to divide them among 6 students. One way to do this is via a sequence of 8 rounds:

- on the first round, hand each student a cookie: "one for you, one for you, one for you…";
- on the second round, hand each student a cookie;
- …
- on the eighth round, hand each student a cookie.

At this point we have only three cookies left, and we can't divide them fairly among the students without breaking them.

Be careful here. The definition of an operation is that it takes *two* inputs and returns *one* output (to use our words above, “from S^2 to S”). This worked fine for addition, subtraction, multiplication, and division on $\mathbb{Q}$ and $\mathbb{R}$.

If division were an operation on $\mathbb{N}$, we would map from $\mathbb{N}^2$ to $\mathbb{N}$. But we just pointed out that division on $\mathbb{N}$ has *two* results. It maps from $\mathbb{N}^2$ (the dividend n and the divisor d) to $\mathbb{N}^2$ (the quotient q and the remainder r). So our first observation about division is that it is *not* an operation, but merely a *function* on $\mathbb{N}^2$!

How can we carry out division? One rather simplistic way is via the following algorithm.

Algorithm 1.1 Simplistic Division Algorithm

```
input
    • $n \in \mathbb{N}$
    • $d \in \mathbb{N}^+$
output
    • $q, r \in \mathbb{N}$ such that $n = qd + r$ and $r < d$
do
    1. let $r = n$, $q = 0$
    2. while $r \geq d$
        (a) increment $q$ by 1
        (b) decrement $r$ by $d$
    3. return $q$ and $r$
```

Let’s see how this algorithm produces the result of the previous example.

Example 1.26. We want to divide $n = 51$ cookies among $d = 6$ students. Step 1 of the algorithm assigns $r = 51$ and $q = 0$.

We proceed to step 2. A “while” statement means that as long as the condition is true, we perform the steps indented underneath it. Since $r = 51$ and $d = 6$, we have $r \geq d$. In step 2(a), we increment q by 1, obtaining $q = 1$. In step 2(b), we decrement r by d, obtaining $r = 45$.

We remain in step 2 because $r \geq d$. Increment q to 2 and decrement r to 39.

We remain in step 2 because $r \geq d$. Increment q to 3 and decrement r to 33.

…

The algorithm continues until q rises to 8 and r falls to 3. At this point, $r < d$, so the “while” statement’s condition is false, and the algorithm ends.

The repetition in Step 2 corresponds to the repetition in our cookie analogy. In the first pass through Step 2, we give each student a cookie. Each student has received $q = 1$ cookies, and the number of cookies remaining decreases to $r = 45$. In the second pass through Step 2, we give each student another cookie. Each student has received $q = 2$ cookies, and the number of cookies remaining decreases to $r = 39$. … Eventually, each student has received $q = 8$ cookies, and the number of cookies remaining has decreased to $r = 3$.

Whenever we describe a new algorithm, we have to verify two important properties.

1. Termination: The algorithm eventually produces *some* result.

2. Correctness: The algorithm's result is the *claimed* result.

Unfortunately, we do not yet have enough theory to explain why Algorithm 1.1 terminates correctly. We need to consider the natural numbers a little more carefully.

The Well-Ordering Property and some of its consequences

You are familiar with the natural ordering of numbers; in Example 1.10 we saw that for any $a, b \in \mathbb{N}$

$$a < b \quad \text{if and only if} \quad b - a \in \mathbb{N}^+ .$$

This characterization works for any $a, b \in \mathbb{Z}$, as well.

An interesting property of $\mathbb{Z}$ is that it has no smallest element. After all, for any $z \in \mathbb{Z}$ we know that $z - 1 \in \mathbb{Z}$ and then

$$z - 1 < z \quad \text{because} \quad z - (z - 1) = 1 \in \mathbb{N}^+.$$

What about $\mathbb{Z}$'s subsets? Many of its sets do not have a smallest element. Consider $S = \{-3, -4, \ldots\}$; the same argument we applied to $\mathbb{Z}$ applies to S.

On the other hand, $\mathbb{N}$ has 0 as a smallest element: for any nonzero $n \in \mathbb{N}$, we have $n - 0 = n \in \mathbb{N}^+$, so $0 < n$. Let's highlight this as an important and useful fact.

Lemma 1.27. *Under the natural ordering, zero is the smallest element of* $\mathbb{N}$.

What about $\mathbb{N}$'s subsets? Intuitively, this seems to be true, but it is not so easy to prove this deductively. In fact, without assuming anything at all, *one cannot prove that all of* $\mathbb{N}$*'s subsets have smallest elements.* So we take it "on faith" to be true.[6]

Axiom (The Well-Ordering Property)**.** *Every subset of* $\mathbb{N}$ *has a least element.*

The Well-Ordering Property gives us a very useful technique that we will use repeatedly. A non-increasing sequence of negative numbers like

$$-5, -9, -12, -14, \ldots$$

might grow smaller and smaller for ever — or it might "stabilize" at a number and never go below it. For instance, if the sequence is

$$-5, -9, -12, -14, -21, -24, -30, -30, -30, -30, \ldots$$

[6] This is something we try to avoid in mathematics if at all possible, but in some cases it's unavoidable. This is one of those cases.

then the sequence has "stabilitized."

Without being able to check every element of the sequence, we can't say whether a sequence of integers stabilizes. But a sequence of natural numbers is different; if you see

$$14, 12, 9, 5, \ldots$$

then you feel fairly confident that the sequence will in fact stabilize. The Well-Ordering Property allows us to prove that this is in fact the case.

Theorem 1.28. *Every non-increasing sequence of natural numbers $a_1, a_2, \ldots$ eventually stabilizes at a least element.*

Proof. Let $a_1, a_2, \ldots$ be a non-increasing sequence of natural numbers. Let $A = \{a_1, a_2, \ldots\}$. Every $a_i \in \mathbb{N}$, so $A \subseteq \mathbb{N}$. By the Well-Ordering property, A has a least element; call it $\hat{a}$. By definition of A, there exists i such that $\hat{a} = a_i$.

We claim that the sequence stabilizes at $\hat{a}$. By hypothesis, the sequence is non-increasing, so $\hat{a} = a_i \geq a_{i+1} \geq a_{i+2} \geq \cdots$. By the transitive property, $\hat{a} \geq a_j$ for every $j \geq i$. On the other hand, $\hat{a}$ is the least element of A, and by definition of A, $a_j \in A$ for every $j \geq i$, so we also have $\hat{a} \leq a_j$. Now, if

$$\hat{a} \geq a_j \quad \text{and} \quad \hat{a} \leq a_j$$

then in fact

$$\hat{a} = a_j \ .$$

(You will prove this in Exercise 1.33.) This is true for all the $j \geq i$, so the sequence has $\hat{a} = a_{i+1} = a_{i+2} = \cdots$. In other words, the sequence has stabilized at $\hat{a}$. □

Theorem 1.28 gives us the information we need to prove that our simplistic division algorithm both produces output a result and produces the claimed result.

Corollary 1.29. *Algorithm 1.1 terminates correctly.*

Proof. First we show that the algorithm terminates. If the algorithm does not execute step 2 at all, then it certainly terminates, so suppose it continues to step 2. Enumerate each value of r computed in step 2(b) as r_1, r_2, and so forth. By definition, $r_{i+1} = r_i - d$. Rewrite this as $r_i - r_{i+1} = d$; since $d \in \mathbb{N}^+$, we have $r_i > r_{i+1}$. The sequence of r's is thus a non-increasing sequence, and by Theorem 1.28 it stabilizes at a least element, say r_k. If $r_k \geq d$, the algorithm would perform step 2 again, creating a smaller r_{k+1}, contradicting our observation that the sequence of r's has a least element. Hence $r_k < d$, in which case the algorithm has terminated.

Now we show that the algorithm's final q and r are correct. As before, enumerate the r's of step 2(b) as $r_1, r_2, \ldots, r_k$. Notice that the final $r = r_k$, so the algorithm repeated step 2 exactly k times. Put $r_0 = n$. We consider the criteria slightly out of order.

- Is $0 \leq r < d$?
    - Certainly $r < d$; otherwise, the algorithm wouldn't have terminated.

  - If $r \notin \mathbb{N}$, then $r < 0$. Since its initial value in step 1 is $r_0 = n$, a natural number, the algorithm must have performed step 2 at least once. In particular, it performed step 2 on r_{k-1}, so

    $$r_k = r_{k-1} - d \quad \Longrightarrow \quad r_{k-1} = r_k + d = r + d < 0 + d = d\ .$$

    The right hand side claims that $r_{k-1} < d$. But if the algorithm performed step 2 exactly k times, then it performed step 2 on r_{k-1}, which requires $r_{k-1} \geq d$, a contradiction. So $0 \leq r < d$, as claimed.

- Is $n = qd + r$?
  - If the algorithm does not perform step 2 at all, then
    - $r = n$ and $q = 0$, so $qd + r = 0 \times d + n = n$, satisfying the claim that $qd \leq n$; and
    - since the algorithm did not perform step 2, we must have $r < d$, and $r = n \in \mathbb{N}$ implies that $0 \leq r$.
  - Suppose that the algorithm continues to step 2. Observe that

    $$\begin{aligned} r_1 &= n - d \\ r_2 &= r_1 - d = n - 2d \\ &\vdots \\ r_k &= n - kd\ . \end{aligned}$$

    Meanwhile, q started at 0, and in step 2(a) increased to 1, then to 2, then to 3, … and after the kth step, $q = k$. By substitution,

    $$qd + r = kd + (n - kd) = n.$$

We have shown that $n = qd + r$ and $0 \leq r < d$, so the proof is complete. □

The Division Theorem

Algorithm 1.1 applies to natural numbers only. We can extend this fairly easily to all integers.

Theorem 1.30 (The Division Theorem)**.** *Let $n, d \in \mathbb{Z}$ with $d \neq 0$. There exist $q, r \in \mathbb{Z}$ such that*

- *$n = qd + r$, and*
- *$0 \leq r < |d|$.*

In addition, q and r are uniquely determined by n and d.

Proof. If $n, d \in \mathbb{N}$, then Corollary 1.29 proves existence of $q, r \in \mathbb{Z}$; for uniqueness, see below. Otherwise, at least one of $n, d < 0$. We consider this in three cases. For each case we consider an example.

Case 1. Suppose $n < 0$ but $d > 0$. (Put another way, $n \notin \mathbb{N}$ but $d \in \mathbb{N}^+$.)

Example. Suppose $n = -51$ and $d = 6$. Algorithm 1.1 requires nonnegative numbers, so what happens if we consider 51 and 6? We get $51 = 8 \times 6 + 3$. If we multiply both sides by -1, we have $-51 = -(8 \times 6 + 3) = (-8) \times 6 + (-3)$. The theorem allows $q < 0$, but not $r < 0$. Can we fix this somehow?

We can: add $0 = 6 + (-6)$ on the right hand side. We have $-51 = [(-8) \times 6 + (-3)] + [6 + (-6)]$, which we can rewrite as $-51 = (-8 - 1) \times 6 + (-3 + 6) = -9 \times 6 + 3$. Now we can set $q = -9$ and $r = 3$ and they satisfy the theorem!

This insight allows us to prove the theorem. If $n < 0$, then $-n \in \mathbb{N}$. Divide that by d; Corollary 1.29 tells us that Algorithm 1.1 will give us $\hat{q}, \hat{r} \in \mathbb{N}$ such that $-n = \hat{q}d + \hat{r}$ and $\hat{r} < d$. Multiply both sides by -1 and we have

$$n = -(\hat{q}d + \hat{r}) = (-\hat{q})\,d + (-\hat{r}) \ .$$

If we set $q = -\hat{q}$ and $r = \hat{r}$, then we have $n = qd + r$, but $r \leq 0$. This will work in the theorem only if $\hat{r} = 0$. Otherwise, we try the workaround of the example: let $q = -\hat{q} - 1$ and $r = d - \hat{r}$. By substitution,

$$\begin{aligned} qd + r &= (-\hat{q} - 1) \times d + (d - \hat{r}) \\ &= (-\hat{q}d - d) + (d - \hat{r}) \\ &= -\hat{q}d - \hat{r} \\ &= -(\hat{q}d + \hat{r}) \\ &= -(-n) \\ &= n \ . \end{aligned}$$

So $n = qd + r$, satisfying the first requirement.

As for the second, if $\hat{r} \neq 0$, then $\hat{r} \in \mathbb{N}^+$, so $0 < \hat{r} < d$. Multiply through by -1 to obtain $0 > -\hat{r} > -d$. Add d to every item to obtain $0 + d > -\hat{r} + d > d + (-d)$, or $d > d - \hat{r} > 0$. Recall that $r = d - \hat{r}$, so $0 < r < d$, satisfying the second requirement.

Case 2. Suppose $n > 0$ but $d < 0$. (Put another way, $n \in \mathbb{N}^+$ but $d \notin \mathbb{N}$.)

Example. Suppose $n = 51$ and $d = -6$. Algorithm 1.1 requires nonnegative numbers, so what happens if we consider 51 and 6? We get $d = 8$ and $r = 3$. We have $51 = 8 \times 6 + 3$, but we need an expression for -6 rather than 6. Instead of multiplying both sides by -1, however, we notice that $51 = (-8) \times (-6) + 3$. We satisfy the theorem with $q = -8$ and $r = 3$!

We leave the generalization of this example to a proof as an exercise for the reader.

Case 3. Suppose both $n, d < 0$. (Put another way, $n, d \notin \mathbb{N}$.)

Example. We leave the creation of an example as an exercise for the reader.

We leave the generalization of this example to a proof as an exercise for the reader.

The three cases listed cover all possibilities; in each case we can find $q, r \in \mathbb{Z}$ such that $n = qd + r$ and $0 \leq r < d$, proving the existence of q and r.

We still have to show that q and r are unique. To that end, suppose there exist $q, \hat{q}, r, \hat{r} \in \mathbb{Z}$ such that $n = qd + r$ and $n = \hat{q}d + \hat{r}$ and $0 \leq r, \hat{r} < |d|$. By substitution, $qd + r = \hat{q}d + \hat{r}$. Rewrite this equation as $(q - \hat{q})\,d = \hat{r} - r$. Since d divides the left hand side, it also divides the right. On the other hand, we can rewrite $0 \leq r, \hat{r} < |d|$ as $0 - |d| < r - \hat{r} < |d| - 0$, or $-|d| < r - \hat{r} < |d|$. Recall that d divides $r - \hat{r}$; the only multiple of d that lies between $-|d|$ and $|d|$ is 0, so $r - \hat{r} = 0$, or $r = \hat{r}$. Substitute into $(q - \hat{q})\,d = \hat{r} - r$ to see that $(q - \hat{q})\,d = 0$. This is possible only if $q - \hat{q} = 0$ or $d = 0$. By the theorem's hypothesis, $d \neq 0$, so we must have $q - \hat{q} = 0$, or $q = \hat{q}$, showing that there is only one possible choice for q and r to satisfy the theorem. □

Exercises

Exercise 1.31. When it comes to the natural numbers, the integers, and the real numbers, we will accept "on faith" the four properties we listed for each operation at the beginning of this section; that is, we will accept them without explanation in this text. However, it is not too hard to prove these operations for the rational numbers. For instance, addition of rational numbers is closed, since for any two rational numbers a/b and c/d, their sum is

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \,,$$

which is rational because $bd, ad + bc \in \mathbb{Z}$ and $bd \neq 0$. With this example, show that the following properties are also true.

(a) The rational numbers are closed under multiplication.

(b) Multiplication of rational numbers is commutative.

(c) The rational numbers have an additive identity.
Hint: Be sure to state what the identity is, and show not only that it satisfies the identity property, but also that it is a rational number!

(d) The rational numbers have a multiplicative identity.

(e) Every rational number has an inverse under addition.

(f) Every nonzero rational number has an inverse under multiplication.

Exercise 1.32. Show that the set $S = \{-3, -6, -9, \ldots\}$ has no smallest element.
Hint: We showed above that $\mathbb{Z}$ has no smallest element. Adapt that discussion to show that S has no smallest element.

Exercise 1.33. Show that for any natural numbers a and b, if $a \leq b$ and $b \leq a$ then $a = b$.
Hint: Use the fact that $a \leq b$ implies $b - a \in \mathbb{N}$, and $b \leq a$ implies $a - b \in \mathbb{N}$. If both $b - a$ and its opposite are natural numbers, what does that tell you about $b - a$?

Exercise 1.34. Generalize the example for Case 2 of the proof of Theorem 1.30 to a proof for arbitrary $n > 0$ and $d < 0$.

Exercise 1.35. Suppose $n, d < 0$.

(a) Explain how you can rewrite the expression $51 = 8 \times 6 + 3$ to find q, r such that $-51 = q \times (-6) + r$ and $0 \leq r < 6$.

(b) Generalize your work in part (a) to a proof for arbitrary $n < 0$ and $d < 0$. This completes the proof of Theorem 1.30.

Exercise 1.36. Prove the *zero product property* for the rational numbers: that is, show that if $ab = 0$ then $a = 0$ or $b = 0$.
Hint: Assume that $ab = 0$ but $a \neq 0$. Then a has a multiplicative inverse in $\mathbb{Q}$. Use a^{-1} to show that $b = 0$.

Exercise 1.37. The Well-Ordering Property is not true for the rational numbers. One reason is that $\mathbb{Z} \subseteq \mathbb{Q}$, so if every subset of $\mathbb{Q}$ had a least element, then $\mathbb{Z}$ would, too, but it does not.

However, the Well-Ordering Property fails even if we consider only nonnegative rational numbers. To see why, describe a sequence of rational numbers that does not have a least element. Try to choose a sequence whose elements are all decreasing and never stabilizes. Be sure to prove that the elements really are decreasing.

Hint: To show the elements are decreasing, it might help to consider that $a/b < c/d$ if and only if $ad < bc$.

Exercise 1.38. We have only considered the natural ordering of integers, but there are other ways to order them. For instance, define the relation

$$a \lessdot b \quad \text{if and only if} \quad \begin{cases} |a| < |b|, \text{or} \\ |a| = |b| \text{ and } a < b. \end{cases}$$

(Remember that when we write $a < b$ with no dot, we mean the natural ordering.)

(a) Order the integers $-5, -3, -1, 2, 4, 9$ according to $\lessdot$.

(b) Explain why $\mathbb{Z}$ has a smallest element according to the $\lessdot$ ordering. (It will help to name it explicitly.)

(c) Show that every subset of $\mathbb{Z}$ has a smallest element according to the $\lessdot$ ordering.

In other words, $\mathbb{Z}$ satisfies the well-ordering property if you use the $\lessdot$ ordering!

Exercise 1.39. Another way to prove the Division Theorem, albeit less algorithmically, is as follows. Fill in the blanks to complete the proof.

- Let $n, d \in \mathbb{Z}$ and assume $d \neq 0$. Let $S = \{n - qd : q \in \mathbb{Z}\}$. Let $T = S \cap \mathbb{N}$.
- By ____, T has a least element; call it r.
- By ____, $r = n - qd$ for some $q \in \mathbb{Z}$.
- Rewrite to obtain ____, satisfying the theorem's first criterion.

- It remains to show that $0 \leq r < d$.
  - By ____, $r \in \mathbb{N}$.
  - By ____, $0 \leq r$.
  - By way of contradiction, assume $d \leq r$.
    - Rewrite to obtain $0 \leq$ ____.
    - By ____, $r - d \in T$.
    - On the other hand, $r - d < r$ because ____ $<$ ____.
    - This contradicts ____.
    - Hence, $r < d$.

Exercise 1.40. Recall the lattice of natural numbers. Suppose we order its elements in the following way: for any $(a, b), (c, d) \in \mathbb{N}^2$ we have

$$(a, b) \prec (c, d) \qquad \text{if and only if} \qquad \begin{cases} a + b < c + d, \text{ or} \\ a + b = c + d \text{ and } a < c \,. \end{cases}$$

(a) Order the points $(3, 7)$, $(2, 5)$, $(4, 1)$, $(0, 0)$, $(4, 6)$, $(3, 2)$ according to $\prec$.

(b) Explain why $\mathbb{N}^2$ has a smallest element according to the $\prec$ ordering. (It will help to name it explicitly.)

(c) Show that every subset of $\mathbb{N}^2$ has a smallest element according to the $\prec$ ordering.

In other words, $\mathbb{N}^2$ satisfies the well-ordering property if you use the $\prec$ ordering!

Sage supplement

You can divide integers using the / operator in Sage, but that gives you a rational number.

```
sage: 7 / 3
7/3
sage: type( _ )
<type 'sage.rings.rational.Rational'>
```

The _ symbol asks Sage for the result of the last statement. The `type` command gives us the type of an object in Sage; here we get a long string that, for all practical purposes, means that the result of `2 / 3` is something that Sage considers a rational number.

This works if you want rational division, but what if we're interested in integer division, as we were in this section? Sage offers a different command for that, `.quo_rem`. From the dot that precedes the command you would be right to conclude that it is a method, and is used accordingly.

```
sage: 7.quo_rem(3)
(2, 1)
```

This indicates that the quotient is 2 and the remainder is 1. You can assign names to these values if you like.

```
sage: q, r = 7.quo_rem(3)
```

After this, `q` and `r` would have the values 2 and 1, respectively.

This command is perfectly fine, but to illustrate some more aspects of Sage, we define a new command that implements the simplistic division algorithm (Algorithm 1.1).

```
sage: def simplistic_division(n, d):
        r, q = n, 0
        while r >= d:
           q += 1
           r -= d
        return q, r
```

Before trying it out, let's consider what it should do when we execute it. First, you should compare it to Algorithm 1.1 and verify that it looks extremely similar. Next, observe the use of keywords you already know: `def` and `return`. As for the lines themselves:

- The first line defines the function and ends with a colon. Subsequent lines are indented.
- The second line assigns the values `n` and `0` to `r` and `q`, as in the first line of Algorithm 1.1's instructions.
- The third line begins a repetition of statements, called a *loop*. Sage offers several kinds of loops; this one is called a `while` loop, and performs the indented statements only if, and as long as, the stated condition remains true. Here, the condition is that $r \geq d$.
- The fourth line is the first line repeated in the loop. The `+=` symbol is an operator, and it tells Sage to increment the value before it by the value after it. In this case, it increments `q` by `1`.
- The fifth line is the second line repeated in the loop. The `-=` symbol behaves just like the `+=` symbol, except it decrements the value before it. In this case, it decrements `r` by `d`.
- The sixth line is indented but not as much as the ones before. It lines up with the `while` statement. That means that it should be the first command performed after the `while` statement terminates. In this case, it is a `return` statement, so it indicates that the procedure `simplistic_division` should terminate and give the result `q,r`.

One aspect of Sage that it shares with its roots in Python is that you can return multiple values from a procedure. Here; we return `q` and `r`.

Let's go ahead and try this.

```
sage: simplistic_division(7, 3)
(2, 1)
```

We end up with the same result as the `quo_rem` command.

The `simplistic_division` command also reveals the importance of the Well-Ordering Property. If we try to execute it with numbers that are not natural, strange things will result. For instance:

```
sage: simplistic_division(-7, 3)
(0, -7)
```

Here the result is $q = 0$ and $r = -7$. This is incorrect according to our definition of division, because we require $r \geq 0$, but it is true that $-7 = 0 \times 3 + (-7)$.

Things can get worse! If you apply `simplistic_division` with a negative divisor, an "infinite loop" will result. The `while` loop's condition never becomes false, because every time we subtract d (which is negative) from r, the value of r *increases*. You can force the command to stop either by holding `control` and pressing `C` (if you're using Sage via a command line terminal), or by pushing the `Stop` button (if you're using it via a graphical interface). You will then encounter an error message similar to the one shown. Go ahead and try it.

```
sage: simplistic_division(7, -3)
Error in lines 1-1
Traceback (most recent call last):
  File "/cocalc/lib/python2.7/site-packages/
smc_sagews/sage_server.py", line 1188, in execute
flags=compile_flags) in namespace, locals
  File "", line 1, in <module> File "", line 4, in
simplistic_division
  File "src/cysignals/signals.pyx", line 265, in
cysignals.signals.python_check_interrupt
  File "src/cysignals/signals.pyx", line 98, in
cysignals.signals.sig_raise_exception
KeyboardInterrupt
```

The phenomenon of the infinite loop illustrates why we must always prove that an algorithm terminates. So long as the algorithm is described properly, we should only need to worry about "while" statements; all other statements should either be clearly one step, such as an assignment or simple operation, or otherwise depend on an algorithm already proved to terminate.

Recall that the proof of the Division Theorem (Theorem 1.30) explained how to use Algorithm 1.1 for unnatural values.[7] For instance, Case 1 says that if $n < 0$ but $d > 0$, divide $|n|$ by d, obtaining quotient $\hat{q}$ and $\hat{r}$. If $\hat{r} = 0$, use $q = -\hat{q}$ and $r = 0$. Otherwise, use the quotient $q = -\hat{q} - 1$ and the remainder $r = d - \hat{r}$.

Let's write a new command that takes care of this case. Sage has a convenient `if` statement that, like the `while` statement, allows us to execute some lines only if the condition is true. Unlike the `while` statement, an `if` statement does not loop! We'll test if the condition of Case 1 is true; if it is, we'll use the `simplistic_division` command as specified to compute $\hat{q}$ and $\hat{r}$, then adjust them as Case 1 indicates, and return the adjusted values.

```
sage: def unnatural_division(n, d):
          # case 1
          if n < 0 and d > 0:
             q_hat, r_hat = simplistic_division(abs(n), d)
             if r_hat == 0:
                q = -q_hat
                r = 0
             else:
                q = -q_hat - 1
                r = d - r_hat
          # additional cases would go here
          return q, r
```

If you try this with −7 and 3, you obtain the desired result!

```
sage: unnatural_division(-7, 3)
(-3, 2)
sage: unnatural_division(-6, 3)
(-2, 0)
```

In fact, $-7 = (-3) \times 3 + 2$.

Exercises

Exercise 1.41. While Sage allows you to create new commands using the `def` keyword, they will not usually be very efficient. To see why, compare how long it takes perform the following commands:

[7] Technically, it explained this for *some* unnatural values. Others were in the section's exercises. Guess what's coming in the exercises to this supplement?

```
sage: 100000000.quo_rem(2)
(50000000, 0)
sage: simplistic_division(100000000, 2)
(50000000, 0)
```

How long does each command take?

Exercise 1.42. The proof of Theorem 1.30 describes three cases where division by negative numbers can happen. The new `unnatural_division` command implemented only the first case. Add additional lines in place of the comment `# additional cases go here` to implement the other cases.
Hint: First implement only Case 2, then make sure it works properly. Use your answers to Exercises 1.34 and 1.35.

1.3 Common divisors

The previous section described for us a division algorithm, and proved that it had several nice properties.

One of the most useful scientific tools is "divide and conquer." Generally this refers to dividing a task or question into smaller tasks and questions, then dividing again and again until you reach a point where you can answer the questions relatively easily. As it turns out, "divide and conquer" sometimes applies in an analogous way to problems that involve integers: divide them into smaller pieces called "factors", and study those.

Let $a, b, c \in \mathbb{Z}$. We say that c ***divides*** a, and write $c \mid a$, if the remainder of dividing a by c is 0. We say that a is ***divisible*** by c, and that c is a ***divisor*** of a. If c does not divide a, but rather a has a nonzero remainder when divided by c, we write $c \nmid a$.

Example 1.43. $4 \mid 8$ but $4 \nmid 6$.

We further say that c is a ***common divisor*** of a and b if $c \mid a$ and $c \mid b$.

Example 1.44. The numbers $a = 12$ and $b = 16$ have common divisors 1, 2, and 4.

People sometimes make the mistake of saying, "If $d \mid a$, then $d \leq a$." This is not generally true; after all, $2 \mid -2$, but $2 \not\leq -2$. We make this mistake because we think of a few examples, usually natural numbers, then generalize from that small pattern. Sometimes this works, but all too often it doesn't. Our world contains negatives as well as positives, so we have to be more careful than this.[8]

Lemma 1.45. *Let $a, d \in \mathbb{N}^+$. If $d \mid a$, then $d \leq a$.*

[8] It's always worth asking, why does a theorem (or lemma) require all its qualifiers? If I drop one of the qualifiers, does the theorem (or lemma) remain true? In the case of Lemma 1.45, we already know that dropping the requirement $a, d \in \mathbb{N}^+$ makes it false in general: we gave an example before the lemma. All the same, there is a set of cases where the lemma remains true even without that strict qualifier. We're not interested in it here, but see if you can identify it.

Proof. Assume $d \mid a$. By definition, there exists $q \in \mathbb{N}$ such that $qd = a$. By way of contradiction, suppose $d > a$. Then $2d = d + d > a$, $3d = 2d + d > a$, … until $qd > a$. By substitution, $a = qd > a$, so $a > a$, a contradiction. The assumption that $d > a$ must be invalid; we conclude that $d \leq a$. □

If a is natural, then the the only natural numbers smaller than a are $\{0, 1, 2, \ldots, a - 1\}$, a finite set. By Lemma 1.45, all of a's divisors come from that set, so a has finitely many divisors. Given two positive integers a and b, each can have only finitely many divisors, so they can have only finitely many common divisors. We call the largest of these the ***greatest common divisor***, written $\gcd(a, b)$.

Example 1.46. Building on Example 1.44, $\gcd(12, 16) = 4$.

What if one of a or b is not positive? For negatives, we can define $\gcd(a, b) = \gcd(|a|, |b|)$. This is both intuitive and consistent, because both a and $-a$ have the same divisors. For instance, the divisors of -16 are the same as the divisors of 16: ± 1, ± 2, ± 4, ± 8, ± 16. So $\gcd(12, -16) = \gcd(12, 16) = 4$.

What if one of a or b is zero? So long as *only* one of them is zero, we can define $\gcd(a, 0) = |a|$. This is both intuitive and consistent, because $|a| \mid 0$, and $|a|$ is also a's largest divisor. So $\gcd(12, 0) = 12$.

What if both a and b are zero? There's no good answer for $\gcd(0, 0)$: *every* integer c divides 0, and there is no largest integer c. We resolve this conundrum by keeping $\gcd(0, 0)$ undefined, consistent with the resolution of problems with dividing by 0.

Over two thousand years ago, Euclid described a very nice way to use compute the greatest common divisor via division. Surprisingly, it is still one of the most efficient methods to compute a gcd.

Algorithm 1.2 The Euclidean Algorithm

```
Input
  • $a, b \in \mathbb{N}^+$
Output
  • $\gcd(a, b)$
Do
  1. let $m = \max(a, b)$, $n = \min(a, b)$
  2. while $n \neq 0$
     (a) determine $q, r$ that satisfy the Division Theorem
     (b) replace $m$ by $n$, then replace $n$ by $r$
  3. return $m$
```

As with Algorithm 1.1, we'll have to prove that Algorithm 1.2 terminates correctly. First let's look at an example to see how it works.

Example 1.47. We compute $\gcd(142, 64)$. Step 1 of the algorithm assigns $m = 142$, $n = 64$.

Since $n \neq 0$, we proceed to step 2, compute $q = 2$ and $r = 14$, and replace m by 64 and n by 14.

Since $n \neq 0$, we repeat step 2, compute $q = 4$ and $r = 8$, and replace m by 14 and n by 8.

Since $n \neq 0$, we repeat step 2, compute $q = 1$ and $r = 6$, and replace m by 8 and n by 6.

Since $n \neq 0$, we repeat step 2, compute $q = 1$ and $r = 2$, and replace m by 6 and n by 2.

Since $n \neq 0$, we repeat step 2, compute $q = 3$ and $r = 0$, and replace n by 2 and n by 0.

We now have $n = 0$, so the algorithm terminates with $\gcd(142, 64) = 2$. You can confirm this result by listing all the divisors of 142 and 64.

Theorem 1.48. *The Euclidean Algorithm terminates correctly.*

Proof. Enumerate each m and n computed in steps 1 and 2(b) as $m_0, m_1, \ldots$ and $n_0, n_1, n_2, \ldots$. For convenience, write $d = \gcd(a, b)$.

Termination? Consider that for $i > 0$ we know that n_i is the remainder of dividing m_{i-1} by n_{i-1}, so $n_{i-1} \geq n_i$ for each $i = 1, 2, \ldots$. This is a non-increasing sequence of natural numbers; by Theorem 1.28, it must stabilize at a least element, say n_k. If $n_k \neq 0$, then the algorithm would perform step 2 again, and obtain a remainder from dividing m_k by n_k, and assign it to n_{k+1}. That makes $n_k > n_{k+1}$, contradicting the observation that n_k is our smallest n. The assumption that $n_k \neq 0$ must have been wrong; we conclude that $n_k = 0$. Yet once $n_k = 0$ the algorithm terminates.

Correctness? We claim that $\gcd(m_i, n_i) = \gcd(m_{i+1}, n_{i+1})$ for each $i = 0, 1, 2, \ldots$. If the claim is true, then the last pair is $\gcd(r, 0) = r$, and we would have $\gcd(a, b) = \gcd(m_0, n_0) = \gcd(m_{\text{last}}, n_{\text{last}}) = r$.

But is the claim true? Let $d = \gcd(m_i, n_i)$, and choose $x, y \in \mathbb{N}$ such that $m_i = xd$ and $n_i = yd$. The algorithm assigns $m_{i+1} = n_i$ and n_{i+1} to be the remainder of dividing m_i by n_i. Let q be the quotient of that division, so that

$$m_i = qn_i + n_{i+1} .$$

By substitution,

$$xd = q(yd) + n_{i+1} .$$

Rewrite this as

$$d(x - qy) = n_{i+1} .$$

By definition, $d \mid n_{i+1}$, so d is a common divisor of n_{i+1} and $n_i = m_{i+1}$. So $d \leq \gcd(m_{i+1}, n_{i+1})$.

Now let $d' = \gcd(m_{i+1}, n_{i+1})$. Recall that $m_{i+1} = n_i$, so $d' = \gcd(n_i, n_{i+1})$. Choose $u, v \in \mathbb{N}$ such that $n_i = ud'$ and $n_{i+1} = vd'$. By substitution,

$$m_i = qn_i + n_{i+1} \quad \Longrightarrow \quad m_i = q(ud') + vd' \quad \Longrightarrow \quad m_i = d'(qu + v) .$$

By definition, $d' \mid m_i$, so d' is a common divisor of m_i and n_i. So $d' \leq \gcd(m_i, n_i)$.

Putting together the previous two paragraphs, we have

$$d \leq \gcd(m_{i+1}, n_{i+1}) = d' \leq \gcd(m_i, n_i) = d .$$

In short,

$$d \leq d' \text{ and } d' \leq d ;$$

by Exercise 1.33, $d = d'$. By substitution,

$$\gcd(m_i, n_i) = \gcd(m_{i+1}, n_{i+1}) \ .$$

□

The various quotients and remainders turn out to be even more useful.

Theorem 1.49 (Extended Euclidean Algorithm)**.** *For any $a, b \in \mathbb{Z}$, there exist $x, y \in \mathbb{Z}$ such that $ax + by = \gcd(a, b)$.*

Example 1.50. For $a = 142$ and $b = 64$, we have $142 \times (-9) + 64 \times 20 = 2$.

We call the equation

$$ax + by = \gcd(a, b)$$

the ***Bézout identity***, and we call x and y ***Bézout coefficients*** of $\gcd(a, b)$. There are infinitely many Bézout coefficients, but it suffices to find only one pair. We can find these coefficients by back-substituting through the various divisions of the Euclidean Algorithm, in reverse order.

Example 1.51. We found $\gcd(142, 64) = 2$ via the divisions

$$\begin{aligned} 142 &= 2 \times 64 + 14 && (1.1)\\ 64 &= 4 \times 14 + 8 && (1.2)\\ 14 &= 1 \times 8 + 6 && (1.3)\\ 8 &= 1 \times 6 + 2 \ . \end{aligned}$$

First we isolate $\gcd(142, 64)$ in the last equation,

$$2 = 8 + (-1) \times 6 \ . \qquad (1.4)$$

Now isolate the remainder of equation (1.3),

$$6 = 14 + (-1) \times 8 \ .$$

Substitute that into equation (1.4) and we have

$$2 = 8 + (-1) \times [14 + (-1 \times 8)] \ ,$$

or

$$2 = 2 \times 8 + (-1) \times 14 \ . \qquad (1.5)$$

Isolate the remainder of equation (1.2),

$$8 = 64 + (-4) \times 14 \ .$$

Substitute that into equation (1.5) and we have

$$2 = 2 \times [64 + (-4) \times 14] + (-1) \times 14 \ ,$$

or

$$2 = 2 \times 64 + (-9) \times 14 . \tag{1.6}$$

Equation (1.1) tells us that

$$14 = 142 + (-2) \times 64 .$$

Substitute that into equation (1.6) and we have

$$2 = 2 \times 64 + (-9) \times [142 + (-2) \times 64] ,$$

or

$$2 = 20 \times 64 + (-9) \times 142 ,$$

as in Example 1.50.

We can use this technique to prove the Extended Euclidean Algorithm.

Proof of Theorem 1.49. We only prove the algorithm for the case where $a, b \in \mathbb{N}^+$. For the case where a or b is not positive, we have the following easy cases:

- When $a \notin \mathbb{N}$ and $b \in \mathbb{N}^+$, we first find $s, t \in \mathbb{Z}$ such that $|a|\, s + bt = \gcd(|a|, b)$. Recall that $\gcd(a, b) = \gcd(|a|, b)$ and $a = -|a|$, so by substitution, $a(-s) + bt = \gcd(a, b)$.
- We leave to the reader the cases where (i) $a \in \mathbb{N}^+$ and $b \notin \mathbb{N}$, and (ii) exactly one of a or b is 0.

Assume therefore that $a, b \in \mathbb{N}^+$. Enumerate the various divisions performed during the Euclidean Algorithm as

$$m_0 = q_0 n_0 + r_0 \quad , \quad m_1 = q_1 n_1 + r_1, \quad \ldots, \quad m_k = q_k n_k + r_k ,$$

where $r_k = \gcd(a, b)$ is the last non-zero remainder. Rewrite the last division as

$$\gcd(a, b) = m_k - q_k n_k . \tag{1.7}$$

Recall that $m_k = n_{k-1}$ and $n_k = r_{k-1}$, so rewrite this equation as

$$\gcd(a, b) = n_{k-1} - q_k r_{k-1} .$$

Rewrite the previous division as

$$r_{k-1} = m_{k-1} - q_{k-1} n_{k-1} .$$

Substitute into equation (1.7) to obtain

$$\gcd(a, b) = n_{k-1} - q_k (m_{k-1} - q_{k-1} n_{k-1}) = (1 + q_k q_{k-1}) n_{k-1} + (-q_k) m_{k-1} .$$

By repeating this process, we eventually obtain an expression

$$\gcd(a, b) = Q_0 m_0 + Q_1 n_0 ,$$

which by substitution becomes

$$\gcd(a, b) = Q_0 a + Q_1 b .$$

The values $x = Q_0$ and $y = Q_1$ satisfy the theorem. □

Algorithm 1.3 describes a step-by-step method for the Extended Euclidean Algorithm.

Algorithm 1.3 Extended Euclidean Algorithm

```
Inputs
  • $a, b \in \mathbb{Z}$, not both zero
Outputs
  • $s, t \in \mathbb{Z}$ such that $as + bt = \gcd(a, b)$
Do
  1. apply the Euclidean Algorithm on $|a|$ and $|b|$, enumerating the divisions as $m_i = q_i n_i + r_i$
  2. let $k$ be the number of the last division with a nonzero remainder
  3. solve $m_k = q_k n_k + r_k$ for $r_k$, obtaining an expression
         $\gcd(a, b) = m_k s_k + n_k t_k$        (1.8)
     (in the first case we have $s_k = 1$ and $t_k = -q_k$)
  4. let $i = k - 1$
  5. while $i \in \mathbb{N}$
       (a) substitute $r_i = m_i - q_i n_i$ in place of $n_{i+1}$ in (1.8)
       (b) decrement $i$ by 1
  6. return $s = s_0 \cdot |a|/a$, $t = t_0 \cdot |b|/b$
```

Theorem 1.52. *Algorithm 1.3 terminates correctly.*

Proof. As with Theorem 1.49, we only prove the case where

Termination? Step 1 terminates by Theorem 1.48. Steps 2, 3, 4, 5(a,b), and 6 are simple statements, so they will not inhibit termination. That brings us to step 5, a while statement. The statement continues as long as i is a natural number; it starts at $i = k - 1$ and is changed only by step 5(b), which decreases it by 1. In other words, step 5 will consider the values $i = k - 1, k - 2, \ldots, 1, 0, -1$, at which point the condition $i \in \mathbb{N}$ is no longer true, and the algorithm proceeds to step 6. Hence the algorithm terminates.

Correctness? The computations of steps 3 and 5(a) replicate the proof of Theorem 1.49 for the case where $a, b \in \mathbb{N}$. We leave the remaining cases as an exercise to the reader. □

We point out one more fact about the Bézout coefficients.

Theorem 1.53. *For any $a, b \in \mathbb{Z}$, $\gcd(a, b)$ is the smallest $z \in \mathbb{N}^+$ that can be written in the form $z = as + by$, where $s, t \in \mathbb{Z}$.*

Proof. Write $d = \gcd(a, b)$, and choose $x, y \in \mathbb{Z}$ such that $a = xd$ and $b = yd$. Choose any $s, t \in \mathbb{Z}$. Then

$$as + bt = (dx)s + (dy)t = d(sx + ty) \ .$$

Since s and y were arbitrary, d divides every integer that can be written in the form $as + bt$. Since d is the smallest positive integer that divides d, the claim stands. □

Exercises

Exercise 1.54. For each pair $a, b \in \mathbb{N}$, use the Euclidean Algorithm to compute $\gcd(a, b)$. Then use the Extended Euclidean Algorithm to compute the Bézout coefficients of a and b.

(a) $a = 4, b = 9$

(b) $a = 100, b = 112$

(c) $a = 255, b = 51$

Exercise 1.55. Show that the Extended Euclidean Algorithm gives a correct result for negative inputs, as well.
Hint: You know that it gives the correct result for nonnegative inputs, so try the algorithm on two inputs where at least one is negative, and determine where, and how, it automatically corrects the result. Generalize your insight to a proof.

Exercise 1.56. Let $n \geq 2$.

(a) Show that $\gcd(n + 1, n) = 1$.

(b) What is $\gcd(n + 2, n)$? Explain why.
Hint: It depends on the value of n. Try a few examples before deciding and explaining.

Exercise 1.57. In this section we have discussed common divisors only of positive integers, and the greatest common divisor of two positive integers.

(a) Explain why it makes sense to speak of common divisors of negative numbers, as well.

(b) How would you compute the greatest common divisor of two integers if they are negative?

(c) How would you define $\gcd(a, 0)$ where a is nonzero?

(d) Why does $\gcd(0, 0)$ not make sense?

Exercise 1.58. Suppose $\gcd(a, n) = 1$.

(a) Show that if $\gcd(b, n) = 1$, then $\gcd(ab, n) = 1$.
Hint: Find Bézout identities for $\gcd(a, n)$ and $\gcd(b, n)$, then — since you want to look at $\gcd(ab, n)$ — multiply the identities and see what happens.

(b) Show that $\gcd\left(a^k, n\right) = 1$ for any $k \in \mathbb{N}$.
Hint: It can be easy if you use part (a) and induction.

Algorithm 1.4 Alternate Extended Euclidean Algorithm

Inputs

- $a, b \in \mathbb{N}^+$

Outputs

- $s, t \in \mathbb{Z}$ such that $as + bt = \gcd(a, b)$

Do

```
1. set up the following table, which will probably extend by more rows
```

i	s_i	t_i	m_i	n_i	q_i	r_i
-1	1	0				
0	0	1	$\max(a, b)$	$\min(a, b)$		

```
   ($m_{-1}, n_{-1}, q_{-1}, r_{-1}$ are not needed and remain undefined)
2. let $i = 0$
3. repeat...
   (a) compute $q_i, r_i$ to satisfy the Division Theorem for dividing $m_i$ by $n_i$
   (b) let $s_{i+1} = s_{i-1} - q_i s_i$, $t_{i+1} = t_{i-1} - q_i t_i$, $m_{i+1} = n_i$, $n_{i+1} = r_i$
   (c) increment $i$ by 1
   ...until $r_{i-1} = 0$
4. return $s = s_i$, $t = t_i$
```

Exercise 1.59. Let $a, b, c \in \mathbb{N}$. Suppose $\gcd(a, b) = 1$ and both $a \mid c$ and $b \mid c$. Show that $(ab) \mid c$. *Hint:* Multiply the Bézout identity of $\gcd(a, b)$ by c.

Exercise 1.60. Algorithm 1.4 gives another way to compute the Bézout coefficients of $a, b \in \mathbb{Z}$.

(a) Compute the Bézout coefficients of $\gcd(255, 51)$ using this algorithm, and compare your result to Exercise 1.54.

(b) Prove that Algorithm 1.4 terminates. (Don't worry about correctness. It is correct, but we won't consider those details here.)

Sage supplement

Sage will compute the greatest common divisor of two integers via the `gcd` command. It will compute the Bézout coefficients at the same time as the gcd via the `xgcd` command.

```
sage: gcd(132, 72)
12
sage: xgcd(132, 72)
(12, -1, 2)
```

The result of `xgcd(a,b)` is (d, x, y) where $\gcd(a, b) = d = ax + by$.

In the last section we illustrated basic programming in Sage by implementing a simplistic division algorithm, even though Sage already has a division operator (which is much faster, anyway). In this section we illustrate how to implement the Extended Euclidean Algorithm, even though Sage already has the `xgcd` command. In this case we implement Algorithm 1.4, the Alternate Extended Euclidean Algorithm. One reason is to show how one can keep track of the results from an iteration.

Line 3(b) of the Alternate Extended Euclidean Algorithm requires us to track several values of s and t. We can do this using a list. We can create a list in Sage using the `list` command or brackets `[]`; we add elements to the end of a list using the `.append` command.

```
sage: L = []
sage: L.append(3)
sage: L.append(5)
sage: L.append(-2)
sage: L.append(5)
[3, 5, -2, 5]
```

Unlike a set, a list can contain multiple copies of an element, so we see 5 twice in `L`.

We access elements using the `[]` operator. Proper usage of the `[]` operator might be a little counterintuitive if you aren't accustomed to languages like C and Python where the first element is element 0, not element 1.

```
sage: L
sage: L[0]
3
sage: L[1]
5
sage: L[2]
-2
sage: L[-1]
-2
```

As indicated above, `L[0]` gives you 3, the list's first element. Another curiosity is that negative indices refer to elements from the back of the list. Just as `L[-1]` gives us the last element, `L[-2]` gives us the element before that, and so on.

In a manner similar to sets, we can build lists using Sage's analogy to set-builder notation.

```
sage: L2 = [ i^2 for i in range(20) if is_even(i) ]
sage: L2
[0, 4, 16, 36, 64, 100, 144, 196, 256, 324]
```

There are a number of useful operations you can perform on a list, but discussing them lies beyond the scope of our current motivation, which is to implement the Alternative Extended Euclidean Algorithm. Let's remind ourselves why we need a list anyway: the instructions in Algorithm 1.4 require us to keep track of previously computed s and t values. To do this, we will maintain two lists, `s` and `t`. We will initialize them with the values indicated at the beginning of Algorithm 1.4. The last values in the lists, `s[-1]` and `t[-1]`, will correspond to s_i and t_i. We can thus compute the next values — the ones we have to add, or `.append` to the lists — by translating

$$s_{i+1} = s_{i-1} - q_i s_i \quad\longrightarrow\quad \texttt{s.append(s[-2] - q*s[-1])}$$

and similarly for t_{i+1}. Thus, translating Algorithm 1.4 into Sage code yields the following.

```
sage: def alternate_euclidean(a, b):
         s = [ 1, 0 ]
         t = [ 0, 1 ]
         m, n = max(a, b), min(a, b)
         r = n
         while r != 0:
            q, r = m.quo_rem(n)
            s.append( s[-2] - q*s[-1] )
            t.append( t[-2] - q*t[-1] )
            m, n = n, r
         return s[-2], t[-2]
```

Before testing it, let's point out an important difference. First, Algorithm 1.4 uses a "repeat…until" construction in step 3. This tells a computer to perform steps 3(a)–3(c) *at least once*, test the subsequent condition ($r_{i-1} = 0$), and if it is false repeat the loop. Sage does not have a command that does this, so when implementing it we had to convert it to a "while" construction. This typically involves setting up the variable(s) in the condition so that the while condition will be true at least once (in this case, `r = n` should do the trick).

Another thing to notice is the use of the `!=` operator. This is how we tell Sage to test whether two values are *not* equal to each other. The statement `r != 0` will be true so long as $r \neq 0$, and becomes false only if $r = 0$.

Now let's try the algorithm.

```
sage: alternate_euclidean(132, 72)
(-1, 2)
```

These are precisely the Bézout coefficients that `xgcd` gave us. While `alternate_euclidean` does not return the gcd itself, we can obtain it as follows.

```
sage: _[0] * 132 + _[1] * 72
12
```

(Recall that the `_` symbol asks Sage for the result of the last statement.) In our case, the last statement's result was the pair `(-1, 2)`, so `_[0]` gives us the number -1, and `_[1]` gives us the number 2.

Exercises

Exercise 1.61. Use Sage's `gcd` and `xgcd` commands to verify your answers to Exercises 1.54 and 1.56.

Exercise 1.62. Modify the `alternate_euclidean` command's `return` statement so that it returns $\gcd(a,b)$ in addition to the Bézout coefficients.
Hint: Recall that in both Algorithms 1.2 and 1.4, the gcd appears as the last nonzero remainder. Where does that appear in the program? If necessary, work a simple example by hand to see.

Exercise 1.63. Define a procedure that implements Algorithm 1.2 in Sage. Don't call your procedure `gcd`, as that would "overwrite" Sage's `gcd` command. Rather, call it `euclidean`, then test it with several examples, such as `euclidean(132,72)`.

1.4 Prime and composite numbers

An important example of divisibility occurs when 2 divides a number; we call it ***even***. A number that is not even is ***odd***.

We return to a question we raised in Example 1.5: Is every real number rational? We focus on $\sqrt{2}$. We know from Exercise 1.15 that $\sqrt{2}$ is real; suppose that it is also rational. By definition, we could write $\sqrt{2} = a/b$ such that $a, b \in \mathbb{Z}$ and $b \neq 0$. To make things simple, we'll say that $\gcd(a,b) = 1$; this is a valid assumption, since if $\gcd(a,b) \neq 1$ we can divide both a and b by $\gcd(a,b)$ and write it in simpler terms.

Square both sides of $\sqrt{2} = a/b$, to obtain $2 = a^2/b^2$, or $2b^2 = a^2$. This says that a^2 is an even number. Is a also even?

Lemma 1.64. *An integer is even if and only if its square is even.*

Proof. Let $a \in \mathbb{Z}$, and suppose $2 \mid a$. By definition of divisibility, choose $q \in \mathbb{Z}$ such that $2q = a$. Square both sides to see that $4q^2 = a^2$, or $2\left(2q^2\right) = a^2$, so $2 \mid a^2$.

We prove the converse via its contrapositive: suppose a is *not* even. By definition, the remainder of dividing a by 2 is 1, so we can write $a = 2q + 1$ for some $q \in \mathbb{Z}$. That gives $a^2 = 4k^2 + 4k + 1 = 2\left(2k^2 + 2k\right) + 1$, which is also not even. If a is not even, then a^2 is not even, either. Hence if a^2 is even, then a is. □

We return to the equation $2b^2 = a^2$; Lemma 1.64 tells us that a is even. Choose $q \in \mathbb{Z}$ such that $a = 2q$ and substitute into $2b^2 = a^2 = 4q^2$; now we have $b^2 = 2q^2$. This says that b^2 is an even number. As with a, we conclude that b itself must be even.

Hold on — we assumed above that $\gcd(a, b) = 1$. Now we've found that 2 is a common divisor of a and b, contradicting $\gcd(a, b) = 1$. *Something* isn't right here; which is it? As we pointed out above, it's perfectly reasonable to suppose that $\gcd(a, b) = 1$; we can always reduce a fraction to lowest terms. We had also assumed that $\sqrt{2}$ is rational — that's not so clear. *That* assumption must be wrong: $\sqrt{2}$ is irrational.

We conclude that $\mathbb{Q} \neq \mathbb{R}$.

Prime and composite numbers

That's a neat trick we pulled with Lemma 1.64: $2 \mid a^2$ if and only if $2 \mid a$. Is that true for any divisor besides 2? No, as it turns out: $4 \mid 2^2$, but $4 \nmid 2$. So this property of 2 is rather special: it's not easy to divide 2.

Here's another way to generalize Lemma 1.64: instead of looking at squares, look at products. Let n be any even number, and suppose $n = ab$. By definition of even, $2 \mid n$, so $2 \mid (ab)$. Must a or b be even? Yes! To see why, suppose neither is even. By definition, there exist $q_a, q_b \in \mathbb{N}$ such that $a = 2q_a + 1$ and $b = 2q_b + 1$. By substitution,

$$ab = (2q_a + 1)(2q_b + 1) = 4q_aq_b + 2q_a + 2q_b + 1 = 2(2q_aq_b + q_a + q_b) + 1 .$$

The product is odd! If *both a and b* are odd, then ab is odd. By the contrapositive, if ab is even, then *one of a or b* is even.

This does not apply to all numbers! Consider that $6 \mid 12$ and $4 \times 3 = 12$. By substitution, $6 \mid (4 \times 3)$, but $6 \nmid 4$ and $6 \nmid 3$.

Here's a similar situation. If you have 6 chocolates, then you can divide them evenly among your friends only if you have 1 friend, 2 friends, 3 friends, or 6 friends. But if you have 7 chocolates, you can only divide them without remainder if you have 1 friend or 7 friends; any other number of friends leaves either broken chocolate or broken friendships.

This second property is so important that we give it a name. A natural number is ***prime*** if it has exactly two natural divisors: itself and 1. We can also call a prime number ***irreducible***, because it does not "reduce" by factorization. If, however, a number is greater than 1 and not prime, we call it ***composite***.

Example 1.65. The numbers 2 and 5 are prime. The numbers 4 and 6 are composite. The numbers 0 and 1 are neither prime nor composite,[9] because they don't have exactly two natural divisors (0 has infinitely many while 1 has only itself), and they are smaller than 2.

Prime numbers enjoy a special property that composite numbers do not: they generalize Lemma 1.64 to arbitrary products.

[9] Some grade school textbooks teach children that 1 is prime. It's hard to say that they are "wrong", because an author can define a word however he or she likes, so long as the use is consistent. However, this is arguably an inelegant choice, as theorems become much, much messier to write when 1 is prime. We will point out at least one example of this.

Of course, every grade school textbook the author has seen contains a maximum of zero proofs. This probably explains a great deal about why they say that 1 is prime.

Theorem 1.66 (Euclid's Lemma)**.** *Let $a, b \in \mathbb{Z}$, and $d \in \mathbb{N}^+$. Then d is prime if and only if any time d divides a product ab, we also have $d \mid a$ or $d \mid b$.*

Proof. Assume d is prime and $d \mid ab$. If $d \mid a$, then the statement "$d \mid a$ or $d \mid b$" is true, and we're done. Otherwise, $d \nmid a$. The definition of prime tells us that d's only divisors are 1 and itself, so a and d have only 1 as a common divisor. Thus $\gcd(a, d) = 1$. By the Euclidean Algorithm, we can find $x, y \in \mathbb{Z}$ such that $ax + dy = 1$. Multiply both sides by b, and we have $(ab)x + d(by) = b$. By hypothesis, $d \mid ab$, so we can choose $z \in \mathbb{Z}$ such that $ab = dz$. By substitution, $(dz)x + d(by) = b$, or $d(xz + by) = b$. By definition, $d \mid b$. We have shown that if d is prime, then $d \mid a$ or $d \mid b$.

Conversely, assume that any time $d \mid ab$, we also have $d \mid a$ or $d \mid b$. Choose any $x, y \in \mathbb{N}$ such that $d = xy$. By Lemma 1.45, $x, y \leq d$. Rewrite $d = xy$ as $d \cdot 1 = xy$. By definition, $d \mid xy$. By hypothesis, $d \mid x$ or $d \mid y$; let's say $d \mid x$. By Lemma 1.45, $d \leq x$. We now have $x \leq d \leq x$; by Exercise 1.33, $x = d$, and thus $y = 1$. Since xy was an arbitrary factorization of d, the only factors of d are 1 and itself. Hence d is prime. □

Euclid's Lemma is actually a special case of a more general fact.

Theorem 1.67. *Let $a, b, d \in \mathbb{Z}$ with $d \neq 0$. If $d \mid ab$ and $\gcd(a, d) = 1$, then $d \mid b$.*

We leave the proof to Exercise 1.73, but the following example bears comment.

Example 1.68. Suppose we know that $4 \mid (5x)$. Now, 4 is not prime, but Theorem 1.67 tells us that $4 \mid x$, so in this case 4 behaves like a prime number. We already saw that $4 \mid (2 \times 2)$ but $4 \nmid 2$, so there must be something special about the fact that $\gcd(4, 5) = 1$.

More generally, Theorem 1.67 shows that two numbers that are not prime can act with each other much the same way that a prime number interacts with other numbers. This phenomenon is important enough that we make the following definition: If $\gcd(a, b) = 1$, then we call a and b ***relatively prime***.

How do we find prime numbers? Algorithm 1.5 gives one way.

Algorithm 1.5 The Sieve of Eratosthenes

```
Inputs
  • $n \in \mathbb{N}^+$ with $n > 2$
Outputs
  • every prime $p \leq n$
Do
  1. write the numbers from 2 to $n$
  2. let $i = 2$
  3. while $i \leq \sqrt{n}$
     (a) if $i$ is not itself crossed out, cross out all multiples of $i$ except for $i$ itself
     (b) increment $i$ by 1
  4. return the numbers that are not crossed out
```

Theorem 1.69. *Algorithm 1.5 terminates correctly.*

Proof. Termination? Each of steps 1, 2, 3(a), 3(b), and 4 can be done in finite time. That leaves the while loop of step 3, which repeats as long as $i < \sqrt{n}$. But i starts at 1, and increments by 1 at step 3(b) each time, so eventually it rises above $\sqrt{n}$, which is fixed. Hence the algorithm terminates.

Correctness? As the algorithm crosses out numbers that are obviously composite, we need merely show that the remaining numbers are all prime. Suppose $2 < a \leq n$ and a is not prime. By Exercise 1.78, it has a prime divisor, say p. By Lemma 1.45, $p < a$. If $p \leq \sqrt{n}$, then we would have encountered p in step 3(a) of the algorithm, and crossed out a as a result.

Otherwise, $p > \sqrt{n}$. Choose $q \in \mathbb{N}$ such that $a = pq$. By Lemma 1.45, we also have $q < a$. If $q > \sqrt{n}$ also, then

$$a = pq > \sqrt{n}^2 = n \geq a\,,$$

which says that $a > a$, a contradiction. So we must have $q < \sqrt{n}$. If q is prime, then we would have encountered q in step 3(a) of the algorithm, and crossed out a as a result. Otherwise, q is composite, and as before it must have a prime divisor, say q'; again, $q' < q$. We now have $a = (pq)\,q'$, and since $q' < q$ and $q < \sqrt{n}$ we have $q' < \sqrt{n}$, so we would have encountered q' in step 3(a) of the algorithm, and crossed out a as a result.

No matter how we go about it, we crossed out the composite number a, so the algorithm could not return it in step 4. Hence the algorithm returns only prime numbers, and is correct. □

Factorization

In Exercise 1.78 you will show that every composite number has at least one prime divisor. We can actually say something much stronger.

Theorem 1.70 (The Fundamental Theorem of Arithmetic)**.** *Let $n \geq 2$ be an integer. There exist prime numbers $p_1, \ldots, p_k$ such that $n = p_1 \cdots p_k$. (The p's might not be distinct.) Moreover, if $p_1 \leq \cdots \leq p_k$, then this expression is unique.*

Example 1.71. We can factor $40 = 2 \times 2 \times 2 \times 5$. Here $p_1 = p_2 = p_3 = 2$ and $p_4 = 5$.

Proof. If n is prime, then put $p_1 = n$ and we are done.

Otherwise, let $m_1 = n$ and $i = 1$. While m_i is composite, Exercise 1.78 tells us that m_i has a prime divisor; call it q_i, and choose m_{i+1} such that $m_i = m_{i+1} q_i$; then increment i by 1. By Lemma 1.45, $m_{i+1} \leq m_i$ for each $i = 1, \ldots$; since q_i is prime and m_i is not, we see that $m_{i+1} < m_i$. By Theorem 1.28, the sequence of m's stabilizes at a least value, m_k. Were m_k composite, we could prolong the sequence, but the sequence has now stabilized, so m_k must be prime; let $q_k = m_k$. By substitution, we have

$$n = m_1 = q_1 m_2 = q_1 (q_2 m_3) = \cdots = q_1 (q_2 (\cdots q_k)),$$

a product of primes.

The list of q's now consists of prime numbers. Let $p_1 = \min(q_1, \ldots, q_k)$ and for $i = 2, \ldots, k$ let $p_i = \min(\{q_1, \ldots, q_k\} \setminus \{p_1, \ldots, p_{i-1}\})$. By construction, $p_1 \leq \ldots \leq p_k$. Moreover, each of these corresponds to a unique q_i, so $n = p_1 \cdots p_k$.

It remains to show uniqueness. Suppose we factor n twice and obtain $p_1 \cdots p_k$ and $q_1 \cdots q_\ell$, with the p's and q's prime, but not necessarily distinct. By substitution

$$p_1 \cdots p_k = q_1 \cdots q_\ell . \tag{1.9}$$

By Euclid's Lemma, $p_1 \mid q_i$ for some $i = 1, \ldots, \ell$; similarly, $q_1 \mid p_j$ for some $j = 1, \ldots, k$. The p's and q's are both irreducible, so $p_1 = q_i$, and $q_1 = p_j$. Now,

$$p_1 \leq p_j = q_1 \leq q_i = p_1 ,$$

or, more succinctly,

$$p_1 \leq q_1 \leq p_1 ,$$

so $p_1 = q_1$. Divide both sides of (1.9) by $p_1 = q_1$ and we have $p_2 \cdots p_k = q_2 \cdots q_\ell$. Continuing in this fashion, we find that $p_2 = q_2, \ldots, p_k = q_k$, and $k = \ell$. The factorization is unique. □

Exercises

Exercise 1.72. Use the Sieve of Eratosthenes to compute all the prime numbers smaller than 200.

Exercise 1.73. Prove Theorem 1.67. The proof should be similar to that of the first part of Euclid's Lemma.

Exercise 1.74. Show that $\sqrt{p} \notin \mathbb{Q}$ for any prime number p.

Exercise 1.75. Let $n > 1$. Show that $\sqrt[n]{p} \notin \mathbb{Q}$ for any prime number p.

Exercise 1.76. Show that if a is not a perfect square, then $\sqrt{a}, \sqrt[n]{a} \notin \mathbb{Q}$.

Exercise 1.77. It turns out that if $6 \mid a^2$, then $6 \mid a$.

(a) Explain why. Why does the same argument not apply when $4 \mid a^2$?
Hint: Think about the prime factors of 6 and the prime factors of 4.

(b) We call a composite number ***squarefree*** when it factors as $n = p_1 \cdots p_k$, *and the p's are all distinct.* Show that if n is squarefree and $n \mid a^2$, then $n \mid a$.

Exercise 1.78. Show that every composite number has at least one prime divisor. *Do not* use the Fundamental Theorem of Arithmetic, or anything after that.
Hint: Let $n \in \mathbb{N}^+$, and let $m_1 = n$. While m_i is composite, you can find a factorization $m_i = a_i m_{i+1}$ such that $a_i, m_{i+1} \neq 1$. By Lemma 1.45, $a_i, m_{i+1} \leq m_i$, but neither is 1, so in fact $a_i, m_{i+1} < m_i$. You have a nonincreasing sequence of integers. Can it go on decreasing indefinitely? If it not, what is the only way it can stabilize? What sort of number do you end up with?

Sage supplement

To compute the prime factorization of an integer, use the `factor()` command. The `.divides()` and `.is_prime()` dot commands return `True` or `False` to indicate what their names imply: whether one number divides another, and whether a number is prime.

```
sage: factor(100)
2^2 * 5^2
sage: 2.divides(5)
False
sage: 2.divides(4)
True
sage: 2.is_prime()
True
sage: 4.is_prime()
False
```

Sage will produce a list of primes up to n using a command fittingly called `eratosthenes()`.

```
sage: eratosthenes(100)
[2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53,
59, 61, 67, 71, 73, 79, 83, 89, 97]
```

That said, we will do here the same thing we've done in previous sections: implement the Sieve of Eratosthenes as a procedure. This will give us some more practice manipulating lists.

The Sieve requires only one input, an integer n. Algorithm 1.5 first instructs us to "write the numbers from 2 to n." We could ask Sage to `print` the numbers from to to n, but that won't actually help us in our task; the algorithm tells us to write the numbers so that we can manipulate them. To do that in Sage, we want a list or a set. In step 3(a) we see that we need to test if a number *i* is in our list or set of numbers, and cross out its multiples. For both of these types, the `in` command tests for membership and the method `.remove` removes an element.[10] However, sets have the convenient *difference* operation, implemented in Sage as `.difference` or `.difference_update`; using that makes the code easier to follow, so we'll go with that.

```
sage: def sieve(n):
          S = set( i for i in range(2,n+1) )
          i = 2
          while i <= sqrt(n):
             if i in S:
                S.difference_update(
                  i*j for j in range(2,n/i+1)
                )
             i += 1
          return S
```

Let's look at what this procedure attempts to do.

- The first line defines a procedure named `sieve`, which takes one input, `n`.
- The second line creates a set `S`, which contains the numbers from 2 to *n*. (Recall that the `range` command does *not* include the last number!) This corresponds to step 1 of Algorithm 1.5.
- The third line corresponds to step 2 of Algorithm 1.5.
- The fourth line begins the same loop that we see in step 3 Algorithm 1.5.
- Lines 5–8 implement Step 3(a) of Algorithm 1.5:
  - The fifth line tests if `i` is a member of `S`. This corresponds to the beginning of step 3(a) of Algorithm 1.5!
  - Lines 6–8 creates a new set that consists of all multiples of `i` (computed using `i*j`) starting from `2*i` until `(n+i)/i`, then uses the `.difference_update` method to remove its elements from `S`. We have "split up" the invocation of `.difference_update` over three lines in part because we didn't have much space in the text here, but also to show that Sage suspends its rules on indentation between parentheses.
- The ninth line corresponds to step 3(b) of Algorithm 1.5.
- The tenth line corresponds to step 4 of Algorithm 1.5.

Let's try the command.

[10] It's also more efficient to test for membership in a set. Depending on the set's structure, it may also be more efficient to remove items from a set than from a list.

```
sage: sieve(100)
set([2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47,
53, 59, 61, 67, 71, 73, 79, 83, 89, 97])
```

The only difference between this result and the one we obtained from `eratosthenes` is that this one comes back as a set rather than a list. It's always good to get the same result!

Exercises

Exercise 1.79. Does Sage factor negative numbers? Try it and see if the answer makes sense.

Exercise 1.80. Sage has an `.is_prime` method, but we can implement our own function to test if an integer n is prime:

Algorithm 1.6 Naïve primality test

```
Inputs
    • $n$, an integer
Outputs
    • True if $n$ is prime; False otherwise
Do
    1. let $P$ be the set of all primes no larger than $\sqrt{n}$
    2. for each $p \in P$
        (a) if $p \mid n$, return False
    3. return True
```

Implement Algorithm 1.6 as a Sage procedure, and test it on several "large" integers, both prime and not-so-prime.
Hint: The Sieve of Eratosthenes will do step 1 for you, so just `sieve(sqrt(n))` for that step.

1.5 Congruence

For the rest of this chapter, m is an integer, and $m > 2$.

It's 10:00pm and I tell you we should meet in 4 hours. At what time will we meet? Not at "14:00pm", but at 2:00am. A clock's hour dial has only twelve settings; once we move past 12:59, time resets to 1:00. In algebra, we generalize this idea to dials with different settings and later to polynomials.

Congruence

Throughout this section, $m \in \mathbb{Z} \setminus \{0\}$.

Recall that division is not an operation on $\mathbb{Z}$, but a function on $\mathbb{Z}^2$: we start with a dividend and divisor, and end with a quotient and remainder. While division is not an operation, we can adapt it to one by keeping only part of the result. Let $\overline{a}$ be the remainder after dividing $a \in \mathbb{Z}$ by m.

Example 1.81. If $m = 5$, we would have $\overline{18} = 3$.

Here we have

$$(a, m) \mapsto \overline{a}, \quad \text{an element of} \quad \mathbb{Z}^2 \times \mathbb{Z} .$$

In other words, $\overline{a}$ is an operation on $\mathbb{Z}$. Given our assumption that $n > 2$, it is also *closed*: after all, the Division Theorem guarantees us a remainder.

In our example above, 14:00 changes to 2:00 because 2 is the remainder of 14 after division by 12. We say that a is ***congruent*** *to* b, ***modulo*** m, *written* $a \equiv b \pmod{m}$, if a and b have the same remainder after division by m. The following characterization based on divisibility is often more convenient.

Theorem 1.82. *Let $a, b \in \mathbb{Z}$. Then $a \equiv b \pmod{m}$ if and only if $m \mid (a - b)$.*

Proof. Recall that the phrase "if and only if" signals that the two phrases are equivalent, and thus we have to prove two directions.

Suppose that $a \equiv b \pmod{m}$. By definition, there exist $q_a, q_b \in \mathbb{Z}$ and $r \in \{0, 1, \ldots, , m-1\}$ such that $a = q_a m + r$ and $b = q_b m + r$. By substitution, $a - b = (q_a m + r) - (q_b m + r) = (q_a - q_b) m$. By definition, $m \mid (a - b)$. We have shown that if $a \equiv b \pmod{m}$, then $m \mid (a - b)$.

For the converse, we show its contrapositive. Assume that $a \not\equiv b \pmod{m}$. By definition, there exist $q_a, q_b \in \mathbb{Z}$ and $r_a, r_b \in \{0, 1, \ldots, m-1\}$ such that $a = q_b m + r_a$ and $b = q_b m + r_b$ and $r_a \neq r_b$. By substitution, $a - b = (q_a - q_b) m + (r_a - r_b)$, so $m \mid (a - b)$ if and only if $m \mid (r_a - r_b)$. However,

$$-|m| \quad < \quad r_a - r_b \quad < \quad |m| \ ,$$

so $m \mid (r_a - r_b)$ if and only if $r_a - r_b = 0$. This is true if and only if $r_a = r_b$, which contradicts a hypothesis. Hence $r_a - r_b \neq 0$, and $m \nmid (r_a - r_b)$, and $m \nmid (a - b)$. □

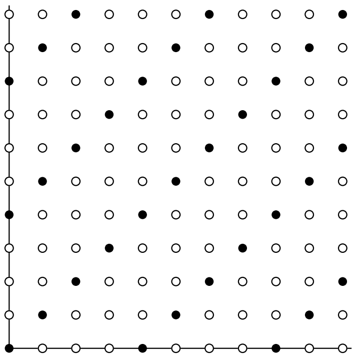
Example 1.83. By definition, $14 \equiv 2 \pmod{12}$. As the theorem indicates, $12 \mid (14 - 2)$.

Congruence is a relation! After all, we can write $a \equiv b \pmod{m}$ as the ordered pair (a, b), and define the set

$$R_n = \{(a, b) : a \equiv b \pmod{m}\} \ .$$

Example 1.84. Some elements of R_{12} would be $(3, 15)$, $(-24, 144)$, $(38, 2)$.

Example 1.85. Some elements of R_4 would be $(3, 3)$, $(7, 3)$, $(3, 7)$. The dots in the lattice diagram below indicate elements of R_4; the circles indicate pairs (a, b) that are *not* congruent modulo 4.



Henceforth, we keep with convention and write $a \equiv b \pmod{m}$ instead of $(a, b) \in R_m$.

The congruence symbol ≡ resembles the equality symbol = for a reason: congruence shares many interesting properties with equality.

Theorem 1.86. *For any $m \geq 2$, congruence modulo m is an equivalence relation.*

Proof. Let $m \geq 2$. We have to show three properties: reflexive, symmetric and transitive.

Reflexive: Let $a \in \mathbb{Z}$. We want to show that $a \equiv a \pmod{m}$. By Theorem 1.82, this is true if and only if $m \mid (a - a)$, or $m \mid 0$. The last statement is definitely true, since $m \times 0 = 0$. Hence $a \equiv a \pmod{m}$; or, congruence modulo n is reflexive.

Symmetric: Let $a, b \in \mathbb{Z}$. We want to show that if $a \equiv b \pmod{m}$, then $b \equiv a \pmod{m}$. We leave this as Exercise 1.104 to the reader.

Transitive: Let $a, b, c \in \mathbb{Z}$. We want to show that if $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$. Assume that $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$. By Theorem 1.82, $m \mid (b - a)$ and $m \mid (c - b)$. By definition, there exist $x, y \in \mathbb{Z}$ such that $mx = b - a$ and $my = c - b$.

We need to show that $a \equiv c \pmod{m}$. By Theorem 1.82, this is true if and only if $m \mid (a - c)$. By definition, this is true if and only if there exists $z \in \mathbb{Z}$ such that $mz = a - c$.

Can we somehow combine $mx = b - a$ and $my = c - b$ to find the desired z? We can:

$$mx - my = (b - a) + (c - b) = c - a \ .$$

If we set $z = x - y$, then

$$mz = m(x - y) = c - a \ .$$

Hence $a \equiv c \pmod{m}$; or, congruence is transitive.

□

Another similarity between congruence and equality is that congruence also allows basic algebra.

Theorem 1.87. *Let $a, b, c \in \mathbb{Z}$, and suppose $a \equiv b \pmod{m}$.*

(A) $a \pm c \equiv b \pm c \pmod{m}$.

(B) $ac \equiv bc \pmod{m}$.

Proof. By Theorem 1.82, $m \mid (a - b)$. We leave a proof of (A) to Exercise 1.105, proving only (B) here.

We want to show that $ac \equiv bc \pmod{m}$. By Theorem 1.82, this is true if and only if $m \mid (ac - bc)$, or $m \mid [(a - b) c]$. We already know that $m \mid (a - b)$; by definition, there exists $q \in \mathbb{Z}$ such that

$$mq = a - b . \tag{1.10}$$

Can we find another integer x such that $mx = (a - b) c$? Certainly: just multiply both sides of (1.10) to obtain

$$m (qc) = (a - b) c .$$

Hence $m \mid [(a - b) c]$, and $ac \equiv bc \pmod{m}$. □

Example 1.88. Since $14 \equiv 2 \pmod{12}$, we also have

$$14 - 2 \equiv 2 - 2 \pmod{12} \quad \text{and} \quad 14 \times 2 \equiv 2 \times 2 \pmod{12} .$$

You can verify this by computing their remainders.

These properties may seem obvious, and proving them may seem silly. Still, consider the following "obvious" property:

$$\text{if } ab = 0 \text{ , then } a = 0 \text{ or } b = 0 \text{ .}$$

This property is always true for integers, rationals, real numbers, and complex numbers; we know it as the ***zero product property***.

Strangely enough, *the zero product property is not guaranteed by congruence!* To wit,

$$2 \times 3 \equiv 0 \pmod{6} \qquad \text{but} \qquad 2, 3 \not\equiv 0 \pmod{6} .$$

The following section elaborates on a similar phenomenon.

Can we divide modulo a number?

Theorem 1.87 tells us that we can add, subtract, and multiply modulo a number. Did you notice that we left something out?

Example 1.89. $14 \equiv 2 \pmod{12}$, but if you divide both sides by 2, you end up with $7 \equiv 1 \pmod{12}$, which is false.

Once again, division stands apart from the arithmetic operations. Yet sometimes it does preserve congruence.

Example 1.90. Suppose $ac \equiv bc \pmod 5$. By definition,

$$5 \mid (ac - bc), \quad \text{or,} \quad 5 \mid [(a-b)c] \ .$$

Now, 5 is a prime number, so by Euclid's Lemma $5 \mid (a-b)$ or $5 \mid c$.

If $5 \mid c$, then $c \equiv 0 \pmod 5$, so by Theorem 1.87, $ac \equiv bc \equiv 0 \pmod 5$.

If $5 \nmid c$, then we must have $5 \mid (a-b)$, and by Theorem 1.82 $a \equiv b \pmod 5$.

In summary, if $5 \nmid c$ then we can go from $ac \equiv bc$ to $a \equiv b$.

Does this apply to other moduli? The key to our example is that when $5 \nmid d$ we could apply Euclid's Lemma. Euclid's Lemma requires only a prime number. In other words, if p is prime, then we should be able to divide modulo p.

Theorem 1.91. *The number p is prime if and only if $ac \equiv bc \pmod p$ implies $a \equiv b \pmod p$ for every $a, b, c \in \mathbb{Z}$ such that $p \nmid c$.*

Proof. Recall that the phrase "if and only if" signals that the two phrases are equivalent, and thus we have to prove two directions.

Suppose p is prime, and let $a, b, c \in \mathbb{Z}$ such that $ac \equiv bc \pmod p$ and $p \nmid c$. By Theorem 1.82, $p \mid (ac - bc)$, or $p \mid [(a-b)c]$. By hypothesis, $p \nmid c$, so by Euclid's Lemma we must have $p \mid (a-b)$. By Theorem 1.82, $a \equiv b \pmod p$.

Conversely, suppose p is not prime. By definition, p is composite, and we can find $a, c \in \mathbb{N}$ such that $ac = p$ and $1 < a, c < p$. Since $ac = p \cdot 1$, we know that $p \mid ac$. By definition, $p \mid ap$. Hence $p \mid (ap - ac)$, and by Theorem 1.82, $ap \equiv ac \pmod b$, but it is *not* the case that $p \equiv c \pmod p$, because $1 < c < p$ implies $c \not\equiv 0 \equiv p$. □

Just as Euclid's Lemma generalized to a theorem for relatively prime numbers, so does Theorem 1.91 generalize to a theorem for relatively prime numbers.

Theorem 1.92. *The numbers c and m are relatively prime if and only if $ac \equiv bc \pmod m$ implies $a \equiv b \pmod m$ for every $a, b \in \mathbb{Z}$.*

Proof. Suppose that c and m are relatively prime. Let $a, b \in \mathbb{Z}$ such that $ab \equiv ac \pmod m$. By Theorem 1.82, $m \mid (ab - ac)$, or $m \mid [(a-b)c]$. By Theorem 1.67, $m \mid (a-b)$. By Theorem 1.82, $a \equiv b \pmod m$.

Conversely, suppose that c and m are not relatively prime. Let $d = \gcd(c, m)$; by hypothesis, $d \neq 1$, and by definition, there exist $x, y \in \mathbb{Z}$ such that $dx = c$ and $dy = m$. By Lemma 1.45, $1 \leq y < m$. By substitution, $cy = (dx)y = (dy)x = mx$, and so $m \mid cy$ by definition. Also by definition, $m \mid cm$. Hence $m \mid (cy - cm)$, so by Theorem 1.82, $cy \equiv cm \pmod m$. However, it is *not* the case that $y \equiv m \pmod m$, because $1 \leq y < m$ implies $y \not\equiv 0 \equiv m$. □

The set $\mathbb{Z}_m$ and its arithmetic

We use our observations in this section to define a new kind of arithmetic. Let

$$\mathbb{Z}_m = \{0, 1, 2, \ldots, n-1\} \ .$$

Recall the remainder operation $\overline{a}$ from page 45. Define addition, subtraction, and multiplication on $\mathbb{Z}_m$ in the following way: for any $a, b \in \mathbb{Z}_n$, we say that

$$\begin{aligned} a \oplus b &= \overline{a+b} \\ a \ominus b &= \overline{a-b} \\ a \otimes b &= \overline{ab} \ . \end{aligned}$$

(We use the symbols $\oplus$, $\ominus$, and $\otimes$ to distinguish this addition, subtraction, and multiplication from the one you're used to.) In other words, whenever we add or multiply two elements of $\mathbb{Z}_m$ we also divide and take the remainder, and this remainder is the sum or product. By the definition of a remainder on page 16, all three of $\overline{a+b}$, $\overline{a-b}$, and $\overline{ab}$ are natural numbers, and also smaller than n. Hence, these operations are *closed;* that is, their result is always in $\mathbb{Z}_m$.

While addition, subtraction, and division on $\mathbb{Z}_m$ are different from addition, subtraction, and division on $\mathbb{Z}$, so that we should probably continue using the symbols $\oplus$, $\ominus$, and $\otimes$, we will go ahead and use ordinary symbols for addition of elements of $\mathbb{Z}_m$, so that in $\mathbb{Z}_5$ we write $3 + 3 = 1$ instead of $3 \oplus 3 = 1$. As long as we keep in mind the set we're working in, there will be no confusion.

Why study $\mathbb{Z}_m$? Let a be any integer. By the Division Theorem, there exists a unique remainder r when we divide a by m, and moreover $r \in \mathbb{Z}_m$. Hence, every integer is congruent to some element of $\mathbb{Z}_m$, so by Theorem 1.87, $\mathbb{Z}_m$ gives us a very powerful tool to study integers. We will see some of this in the next section.

Example 1.93. In $\mathbb{Z}_{12}$, $10 + 4 = 2$, $2 - 4 = 10$, and $10 \times 4 = 4$.

By which elements can we "divide"? A better way to phrase this might be, which elements have multiplicative inverses? By Theorem 1.92, only those numbers relatively prime to the modulus.

Example 1.94. In $\mathbb{Z}_{12}$, only the numbers 1, 5, 7, and 11 have multiplicative inverses: $1^{-1} = 1$, $5^{-1} = 5$, $7^{-1} = 7$, and $11^{-1} = 11$. The other values do not, because they are not relatively prime to 12. If that leaves you skeptical, consider that in $\mathbb{Z}_{12}$ we have

$$\begin{array}{lll} 2 \times 0 = 0, & 2 \times 1 = 2, & 2 \times 2 = 4, \\ 2 \times 3 = 6, & 2 \times 4 = 8, & 2 \times 5 = 10, \\ 2 \times 6 = 0, & 2 \times 7 = 2, & \quad \ldots \end{array}$$

so that all products cycle among 0, 2, 4, 6, 8, 10.

Example 1.95. In $\mathbb{Z}_{14}$, the numbers 1, 3, 5, 9, 11, 13 have multiplicative inverses: $1^{-1} = 1$, $3^{-1} = 5$, $5^{-1} = 3$, $9^{-1} = 11$, $11^{-1} = 9$, and $13^{-1} = 13$.

In small moduli like 12 and 14, it isn't too hard to discover the inverses via brute force. Often, however, the modulus is very large (for example, 32003). In a case like this, how can we find the multiplicative inverse of a number in $\mathbb{Z}_m$?

Amazingly, the same method tells us both *whether* a has a multiplicative inverse modulo m, as well as *what* the inverse is. By Theorem 1.92 tells us that a has multiplicative inverse if and only if $\gcd(a, m) = 1$. To compute the gcd, perform the Euclidean Algorithm. If in fact

$\gcd(a, m) = 1$, then by the Extended Euclidean Algorithm we can use its divisions to compute the Bézout coefficients x and y such that

$$ax + my = 1 \ .$$

Rewrite this as

$$my = 1 - ax \ .$$

By definition,

$$m \mid (1 - ax), \quad \text{or} \quad 1 \equiv ax \pmod{m} \ .$$

In other words, x is the multiplicative inverse of a, modulo m. We have just proved the following theorem.

Theorem 1.96. *Let $m \geq 2$. A nonzero $a \in \mathbb{Z}_m$ has a multiplicative inverse in $\mathbb{Z}_m$ if and only if* $\gcd(a, m) = 1$.

Example 1.97. In $\mathbb{Z}_{14}$, performing the Euclidean algorithm on 14 and 9 gives us the divisions

$$\begin{aligned} 14 &= 1 \times 9 + 5 \\ 9 &= 1 \times 5 + 4 \\ 5 &= 1 \times 4 + 1 \\ 4 &= 4 \times 1 + 0 \ . \end{aligned}$$

Reversing these according to the Extended Euclidean Algorithm, we obtain the equations

$$\begin{aligned} 1 &= 5 + (-1) \times 4 \\ 1 &= 5 + (-1) \times [9 + (-1) \times 5] \\ &= 2 \times 5 + (-1) \times 9 \\ 1 &= 2 \times [14 + (-1 \times 9)] + (-1) \times 9 \\ &= 2 \times 14 + (-3) \times 9 \ . \end{aligned}$$

So the Bézout coefficient of 9 is -3, and $-3 \equiv 11 \pmod{14}$. Hence $11 = 9^{-1}$ in $\mathbb{Z}_{14}$.

Example 1.98. In $\mathbb{Z}_{14}$, performing the Euclidean algorithm on 14 and 12 gives us the divisions

$$\begin{aligned} 14 &= 1 \times 12 + 2 \\ 12 &= 6 \times 2 + 0 \ . \end{aligned}$$

This tells us that $\gcd(14, 12) \neq 0$, so 12 has no multiplicative inverse modulo 14.

We conclude this section by observing that division is related to the zero product property. If we could always divide equal numbers from both sides of a congruence, then

$$2 \times 3 \equiv 0 \quad \text{would imply that} \quad 2 \times 3 \equiv 2 \times 0 \quad \text{and thus} \quad 3 \equiv 0 \pmod{6} \ .$$

Put another way,

Corollary 1.99. *The zero product property holds in $\mathbb{Z}_m$ if and only if m is prime.*

Proof. If m is prime, then it is relatively prime to each of the numbers $1, 2, \ldots, m-1$. Hence if $a, b \in \mathbb{Z}_m$ and $ab = 0$, then by our notation $ab \equiv 0 \pmod{m}$, so $ab \equiv a \times 0 \pmod{m}$, so by Theorem 1.92 $b \equiv 0 \pmod{m}$, or $b = 0$.

On the other hand, if m is not prime, then there exist $a, b \in \mathbb{Z}$ such that $ab = m$ and $1 < a, b < m$. Thus $a, b \in \mathbb{Z}_m$ and $ab = m \equiv 0 \pmod{m}$. By our notation, this means $ab = 0$ in $\mathbb{Z}_m$. □

We will explore this relationship between division (or, more properly, "cancellation") and the zero product property as the material unfolds.

Exercises

Exercise 1.100. In which of the following congruences is the indicated division reliable?

(a) $12 \times a \equiv 12 \times b \pmod{7}$ implies $a \equiv b \pmod{7}$

(b) $12 \times a \equiv 12 \times b \pmod{6}$ implies $a \equiv b \pmod{6}$

(c) $12 \times a \equiv 12 \times b \pmod{25}$ implies $a \equiv b \pmod{25}$

Exercise 1.101. Make a lattice diagram of congruence modulo 3. The x- and y-axes should each go to at least 10. Compare and contrast your answer to Example 1.85.

Exercise 1.102. Compute multiplicative inverses of the following numbers, according to the given moduli. If an inverse does not exist, explain why not.

(a) 72^{-1} in $\mathbb{Z}_{101}$

(b) 105^{-1} in $\mathbb{Z}_{539}$

Exercise 1.103. Show that every nonzero element of $\mathbb{Z}_m$ has a multiplicative inverse if and only if m is prime.

Exercise 1.104. Complete the proof of Theorem 1.86 by showing that congruence modulo n is symmetric.

Exercise 1.105. Complete the proof of Theorem 1.87 by showing that if $a \equiv b \pmod{n}$, then $a \pm c \equiv b \pm c \pmod{n}$.

Exercise 1.106. Recall Euclid's Lemma (Theorem 1.66).

(a) Use Euclid's Lemma to show that if p is prime and $ab \equiv 0 \pmod{p}$, then $a \equiv 0 \pmod{p}$ or $b \equiv 0 \pmod{p}$.

(b) Find $a, b, m \in \mathbb{N}$ such that $ab \equiv 0 \pmod{m}$ but neither $a \equiv 0 \pmod{m}$ nor $b \equiv 0 \pmod{m}$.

(c) For any fixed $m \in \mathbb{N}$, the numbers a and b of part (b) above are called ***zero divisors***. Show that if $m > 2$ is not prime, then you can always find zero divisors $a, b \in \{1, 2, \ldots, m-1\}$. *Hint:* Try factoring m.

Exercise 1.107. Recall the generalization of Euclid's Lemma (Theorem 1.67).

(a) Use Theorem 1.67 to show that if $ab \equiv 0 \pmod{m}$ and $\gcd(a, m) = 1$, then $b \equiv 0 \pmod{m}$.

(b) Find $a, b, m \in \mathbb{N}^+$ that satisfy part (a). (Notice $b \in \mathbb{N}^+$ means $b \neq 0$.)

Sage supplement

You already know that we can compute the remainder of a after division by n using `a.quo_rem(n)` and taking the second entry of the result. Another, more convenient way to do this is by using either the `%` operator[11] or the `.mod` method.

```
sage: (-10).quo_rem(3)
(-4, 2)
sage: -10 % 3
2
sage: (-10).mod(3)
2
```

Sage doesn't have a command named `is_congruent`, but we can easily create one.

```
sage: def is_congruent(a, b, n):
          return (a % n) == (b % n)
```

The first line defines our procedure with three arguments, `a`, `b`, and `c`. The second computes the remainders of `a` and `b` after division by `n`, and returns whether they are the same.

```
sage: is_congruent(7, 5, 4)
false
sage: is_congruent(9, 5, 4)
true
```

There's also another way: Sage makes it easy to compute in $\mathbb{Z}_n$! To do this, we must set up every number as an element of $\mathbb{Z}_n$. There are two steps for this: one you have to do only once, but the other you have to do every time you work with a new number.

1. Define `Zn` as the quotient of $\mathbb{Z}$ and the modulus.

2. Initialize a number a as elements of `Zn` by typing `Zn(`a`)`. This is called ***coercion***.

```
sage: Z3 = ZZ.quo(3)
sage: a = Z3(-10)
sage: a
2
```

One nice consequence of this is that any arithmetic between `a` and other numbers also occurs in $\mathbb{Z}_n$:

[11] You may recognize the `%` operator if you are familiar with any of the programming languages C, Java, and Python.

```
sage: a + 7
0
sage: a * 7
2
```

Let's review what happened in these computations.

- We defined `a` as an element of $\mathbb{Z}_3$: originally we set it to -10, but the remainder of dividing -10 by 3 is 2, so Sage sets `a` to 2.
- If `a` has the value 2, how did Sage figure `a + 7` to be 0? The sum is 9, but again Sage automatically reduces it modulo 3 to 0.
- In the same way, for `a * 7` Sage computes the product 14, then reduces it modulo 3 to 2.

Sage similarly coerces numbers involved in a comparison, giving us an easier way to test congruence than the `is_congruent` command we defined earlier. (Make sure you type *two* equality signs in the example below, not one. Otherwise you'll reassign `a`.)

```
sage: a == 5
True
```

Sage will also compute the multiplicative inverse of `a` in a very natural manner: use the exponent -1.

```
sage: a^(-1)
2
```

Recall that not every number in every modulus has a multiplicative inverse. Sage has a way to tell you this, too. We'll look at the same example we conidered in the text, the number 12 in $\mathbb{Z}_{14}$.

```
sage: Z14 = ZZ.quo(14)
sage: a = Z14(12)
sage: a
12
sage: a^(-1)
Error in lines 1-1
Traceback (most recent call last):
  File "/cocalc/lib/python2.7/site-packages/
smc_sagews/sage_server.py", line 1188, in execute
    flags=compile_flags) in namespace, locals
  File "", line 1, in <module>
  File "sage/rings/finite_rings/integer_mod.pyx",
line 2704, in sage.rings.finite_rings.integer_mod.
IntegerMod_int.__pow__ (build/cythonized/sage/
rings/finite_rings/integer_mod.c:30197)
    return ~self._new_c(res)
  File "sage/rings/finite_rings/integer_mod.pyx",
line 2722, in sage.rings.finite_rings.integer_mod.
IntegerMod_int.__invert__ (build/cythonized/sage/
rings/finite_rings/integer_mod.c:30372)
    raise ZeroDivisionError(f"inverse of Mod({self},
{self.__modulus.sageInteger}) does not exist")
ZeroDivisionError:  inverse of Mod(12, 14) does not exist
```

Whenever you encounter an error in Sage, the first line to examine is the *last* line. It tells you the precise error, gives a brief explanation, and sometimes even suggests how to fix it. In this case, it gives us a *ZeroDivisionError*, which seems like a strange thing to get when looking for a multiplicative inverse; this will make sense later when we talk about zero divisors.

Exercises

Exercise 1.108. Use Sage to find the multiplicative inverse of every invertible element of $\mathbb{Z}_{100}$. *Hint:* Doing this one number at a time would be tedious. Use a `for` loop to make Sage do them all for you in one go! To avoid a *ZeroDivisionError*, use an `if` statement to check whether the number is relatively prime to 100.

1.6 Linear equations in $\mathbb{Z}_m$

In this section we introduce the reader to linear algebra in $\mathbb{Z}_m$; that is, we consider the question of solving a congruence of the form

$$ax \equiv b \pmod{m}$$

or, more generally, a system of congruences of the form

$$\begin{cases} ax \equiv b & \pmod{m} \\ cx \equiv d & \pmod{n} \end{cases}.$$

These are called *linear congruences* because the variable, x, is only to the first power.

If we can find one solution to a linear congruence, then we can find *infinitely many.* Here's why:

- Suppose y is a solution to $ax \equiv b$. Then

  $$ay \equiv b \pmod{m},$$

  but also

  $$a(y+m) = ay + am \equiv b + am \equiv b \pmod{m},$$

  because $(b+am) - b = am \equiv 0 \pmod{m}$.

- In a similar fashion, suppose y is a solution to both $ax \equiv b \pmod{m}$ and $cx \equiv d \pmod{n}$. This time, consider the fact that

  $$a(y+mn) = ay + a(mn) \equiv b + (an)m \equiv b \pmod{m},$$

  and similarly

  $$c(y+mn) = cy + c(mn) \equiv d + (cm)n \equiv d \pmod{mn}.$$

So *if* there is a solution, there are infinitely many.

That said, the other solutions given can be viewed as not especially interesting; after all:

- For a solution y to $ax \equiv b \pmod{m}$, we know that $y + m \equiv y \pmod{m}$. If we consider y as a solution of $\mathbb{Z}_m$, it might well be unique; we're only guaranteed another solution when we add or subtract a multiple of m. What would be interesting is if we *don't* have a solution that is congruent to y modulo m.

- For a solution y to $ax \equiv b \pmod{m}$ and $cx \equiv d \pmod{n}$, we know that $y + mn \equiv y \pmod{mn}$. If we consider y as a solution of $\mathbb{Z}_{mn}$, it might well be unique; we're only guaranteed another solution when we add or subtract a multiple of mn. What would be interesting is if we *don't* have a solution that is congruent to y modulo mn.

Because of this, this section considers two questions:

- Does a solution exist?

- When it does, is it unique *up to congruence?*

"Up to congruence" means that the solution is the only one in $\mathbb{Z}_m$, where the value of n depends on the problem.

One linear congruence

We begin by looking at $ax \equiv b \pmod{m}$. From the previous section we know that if $\gcd(a, m) = 1$, then a has a multiplicative inverse modulo m; that is, we can find $s \in \mathbb{Z}_m$ such that $as \equiv 1 \pmod{m}$. This leads to a very simple solution to the congruence.

Algorithm 1.7 Solving linear congruences

```
Inputs
  • $ax \equiv b \pmod{m}$, where
      – $a, b, n \in \mathbb{Z}$,
      – $n \geq 2$, and
      – $\gcd(a, m) = 1$
Outputs
  • $x \in \mathbb{Z}_m$ satisfying $ax \equiv b$
Do
  1. let $s$ be the multiplicative inverse of $a$ modulo $m$
  2. return the remainder of $bs$ after division by $m$
```

Theorem 1.109. *Algorithm 1.7 terminates correctly. The solution is unique up to congruence.*

Proof. Termination? We can perform step 1 via the Extended Euclidean Algorithm (1.3), and that terminates by Theorem 1.52. Step 2 is a return statement, so the algorithm terminates.

Correctness? If we substitute $x = bs$ into the left hand side of $ax \equiv b \pmod{m}$ we have

$$a(bs) = a(sb) = (as)b \equiv 1 \cdot b = b \pmod{m}.$$

So $x = bs$ would be a correct solution. The algorithm actually returns the remainder after division by m; that is, it returns r where $x = qm + r$. Rewrite that equation to $qm = x - r$ and we see that $x \equiv r \pmod{m}$. In other words, m is also a solution.

Is the solution unique up to congruence? Let y, z be solutions to $ax \equiv b \pmod{m}$. Then $ay \equiv b \pmod{m}$ and $az \equiv b \pmod{m}$. By the transitive property of congruence (Theorem 1.86), $ay \equiv az \pmod{m}$. By hypothesis, $\gcd(a, m) = 1$, so a has a multiplicative inverse modulo m; call it s. Multiply both sides of the congruence by s, and we have

$$(sa)y \equiv (sa)z \quad \Longrightarrow \quad y \equiv z \qquad \pmod{m}.$$

By definition, $m \mid (y - z)$, so $y \equiv z \pmod{m}$, and the solution is unique up to congruence. □

Example 1.110. Consider the linear congruence $9x \equiv 3 \pmod{14}$. In Example 1.95 we found that $9^{-1} \equiv 11 \pmod{14}$. Multiply both sides by 11 and we have $99x \equiv 33 \pmod{14}$. However, $99 \equiv 1 \pmod{14}$, and $33 \equiv 5 \pmod{14}$, so

$$99x \equiv 33 \quad \Longrightarrow \quad x \equiv 5 \qquad \pmod{14}.$$

If you substitute 5 in for x in $9x \equiv 3 \pmod{14}$, you will see that the equation checks out as true. Moreover, Theorem 1.109 tells us the solution is unique up to congruence; that is, none of $\{0, 1, 2, 3, 4\} \cup \{6, 7, \ldots, 13\}$ will work for x.

You may have noticed a little wrinkle; Algorithm 1.7 requires $\gcd(a, n) = 1$. What if this is not true? We consider two examples.

Example 1.111. Algorithm 1.7 cannot solve $6x \equiv 2 \pmod{14}$ directly, because $\gcd(6, 14) = 2 \neq 1$. However, x is a solution to the congruence if and only if $14 \mid (6x - 2)$, which is true if and only if we can find an integer q such that $14q = 6x - 2$. Both sides of the equation are divisible by 2, so we divide and obtain $7q = 3x - 1$. This is equivalent to $7 \mid (3x - 1)$, which is true if and only if x is a solution to $3x \equiv 1 \pmod{7}$. Algorithm 1.7 *can* solve this congruence; it returns $x = 5$. You can easily check that this is also a solution to $6x \equiv 2 \pmod{14}$, precisely what all the equivalences imply.

This solution is not, however, unique! We can add 7 to our solution $x = 3$ to obtain $x = 10$ as another solution to $3x \equiv 1 \pmod{7}$. This is not a new solution modulo 7, but it *is* another solution modulo 14, and it is certainly different from 5 in that modulus!

Example 1.112. Algorithm 1.7 cannot solve $6x \equiv 1 \pmod{14}$ directly, because $\gcd(6, 14) = 2 \neq 1$. In fact, Algorithm 1.7 cannot solve $6x \equiv 1 \pmod{14}$ *at all*, because it is equivalent to the equation $6x = 14q + 1$, or $6x - 14q = 1$, or $2(3x - 7q) = 1$. This latter equation is true if and only if $2 \mid 1$, which it does not!

These results allow us to answer completely the question of solving $ax \equiv b \pmod{m}$.

Theorem 1.113. *Let $a, b, m \in \mathbb{Z}$, with $m \geq 2$. Let $d = \gcd(a, m)$.*

- *If $d = 1$, then the linear congruence $ax \equiv b \pmod{m}$ has one solution, and the solution is unique up to congruence.*
- *If $d \neq 1$, then:*
  - *If $d \nmid b$, then there the linear congruence $ax \equiv b \pmod{n}$ has no solution.*
  - *If $d \mid b$, then the linear congruence $ax \equiv b \pmod{n}$ has d incongruent solutions modulo n, and we can find them by first solving $(a/d)\,x \equiv b/d \pmod{n/d}$, then enumerating the solutions $x, x + n/d, x + 2n/d \ldots, x + (d-1)n/d$.*

Proof. If $d = 1$, then Theorem 1.109 applies.

If $d \neq 1$ and $d \nmid b$, then we can rewrite $ax \equiv b \pmod{m}$ as $ax + mq = b$ for some integer q. Then d divides the left hand side, but it does not divide the right, so there can be no solution.

If $d \neq 1$ and $d \mid b$, then choose $\hat{a}, \hat{b}, \hat{m} \in \mathbb{Z}$ such that $\hat{a}d = a$, $\hat{b}d = b$, and $\hat{m}d = m$. The congruence $ax \equiv b \pmod{m}$ is equivalent to the equation $ax + mq = b$ for some $q \in \mathbb{Z}$. Divide both sides by d to see that this is equivalent to $\hat{a}x + \hat{m}q = \hat{b}$. This latter equation is equivalent to $\hat{a}x \equiv \hat{b} \pmod{\hat{m}}$. We obtained $\hat{a}$ and $\hat{m}$ by dividing a and m by their greatest common divisor, so $\hat{a}$ and $\hat{m}$ can have no common divisor; otherwise, a and m would have a larger one. Hence $\gcd(\hat{a}, \hat{m}) = 1$ and Theorem 1.109 applies; a solution x to $\hat{a}x \equiv \hat{b} \pmod{\hat{m}}$ exists. As explained above, this congruence is equivalent to $ax \equiv b \pmod{m}$, so x is a solution to that one, as well.

Moreover, $y = x + {}^{km}/_{d} = x + k\hat{m}$ is likewise a solution to $\hat{a}x \equiv \hat{b} \pmod{\hat{m}}$ for every $k \in \mathbb{N}$, so as with x, y is also a solution to $ax \equiv b \pmod{m}$.

To show that there are only d incongruent solutions modulo m, let $x, y \in \mathbb{Z}_m$ be any solutions to $ax \equiv b \pmod{m}$. By substitution,

$$ax \equiv ay \pmod{m} .$$

$$ai\left(\frac{m}{d}\right) \equiv aj\left(\frac{m}{d}\right) \pmod{m} .$$

By definition,

$$m \mid a(x-y) .$$

Choose $q \in \mathbb{Z}$ such that

$$mq = a(x-y) ;$$

rewrite this as

$$\hat{m}q = \hat{a}(x-y) .$$

Recall that $\gcd(\hat{a}, \hat{m}) = 1$. By Theorem 1.67, $x - y$ is a multiple of $\hat{m}$. Since $m = \hat{m}d$, there are only d multiples of $\hat{m}$ in $\mathbb{Z}_m$, so there can be be only d distinct values of $x - y$ in $\mathbb{Z}_m$. Thus, there are only d solutions to $ax \equiv b \pmod{m}$. □

Several simultaneous congruences

This section's theorems have been tough, but they lay the groundwork we need to make the remainder easy (no pun intended).

Consider a *system* of congruences

$$\begin{cases} x \equiv a & \pmod{m} \\ x \equiv b & \pmod{n} \end{cases} .$$

Theorem 1.109 tells us that $1 \cdot x \equiv a \pmod{m}$ has a solution $x = a$. Now, a might not be a solution to $x \equiv b \pmod{n}$, but we have pointed out several times that there are infinitely many solutions to $x \equiv a \pmod{m}$, all of them having the form $x = a + km$, where k is any integer. We'd like to find a value of k that makes the second congruence true; in other words, we can treat k as an unknown.

Substitute this expression for x into the second congruence, obtaining

$$a + km \equiv b \pmod{n} .$$

Rewrite this as

$$mk \equiv b - a \pmod{n} .$$

By Theorem 1.109, we can solve this congruence so long as $\gcd(m, n) \mid (b - a)$. In fact, it is *very* easy to solve if $\gcd(m, n) = 1$, and this is such a useful fact that the result is very ancient.

Theorem 1.114 (The Chinese Remainder Theorem)**.** *Let $m, n \in \mathbb{N}$ such that $m, n \geq 2$ and* $\gcd(m, n) = 1$*. The system of linear congruences*

$$\begin{cases} x \equiv a & (\text{mod } m) \\ x \equiv b & (\text{mod } n) \end{cases}$$

has a solution, and the solution is unique up to congruence modulo mn.

Proof. As we explained before the theorem, this system has a solution if $\gcd(m, n) = 1$. It remains to show that the solution is unique up to congruence modulo mn. Let $y, z \in \mathbb{Z}$ be solutions to the system, so that

$$y \equiv a \equiv z \pmod m \quad \text{and} \quad y \equiv b \equiv z \pmod n .$$

By the transitive property, $y \equiv z$ under both moduli. By definition, $m \mid (y - z)$ and $n \mid (y - z)$. By Exercise 1.59, $mn \mid (y - z)$. By definition, $y \equiv z \pmod{mn}$. □

What about the more complicated case

$$\begin{cases} ax \equiv b & (\text{mod } m) \\ cx \equiv d & (\text{mod } n) \end{cases},$$

which we mentioned at the beginning of the section? As with the Chinese Remainder Theorem, we can find solutions to this system using a similar principle. See Algorithm 1.8.

Example 1.115. Solve

$$\begin{cases} 5x \equiv 72 & (\text{mod } 99) \\ 3x \equiv 4 & (\text{mod } 101) \end{cases}.$$

It is easy to verify that $\gcd(5, 99) = \gcd(3, 101) = 1$, so we pass through steps 1 and 2 without difficulty. For step 3, we find the multiplicative inverses of 5 modulo 99 and of 3 modulo 101 to rewrite the system as

$$\begin{cases} x \equiv 54 & (\text{mod } 99) \\ x \equiv 35 & (\text{mod } 101) \end{cases}.$$

Again, it is easy to see that $\gcd(99, 101) = 1$ (the result of Exercise 1.56 makes it *very* easy) so we pass through step 4 without difficulty.

We finally come to something new in step 5. We have to follow step 5(a), which tells us to solve according to the manner outlined in the proof of the Chinese Remainder Theorem. We rewrite the first equation as

$$x = 99q + 54$$

and substitute this into the second equation,

$$99q + 54 \equiv 35 \pmod{101} .$$

Solve this in the usual manner for linear congruences,

$$\begin{aligned} 99q &\equiv 82 \pmod{101} \\ 50 \times 99q &\equiv 50 \times 82 \\ q &\equiv 60 . \end{aligned}$$

Algorithm 1.8 Solving two linear congruences

Inputs

- $a, b, c, d, m, n \in \mathbb{N}^+$ such that $m, n \geq 2$

Outputs

- a solution to
$$\begin{cases} ax \equiv b & \pmod{m} \\ cx \equiv d & \pmod{n} \end{cases},$$
if it exists; otherwise, $\emptyset$

Do

```
1. if $\gcd(a, m) \nmid b$ or $\gcd(c, n) \nmid d$, return $\emptyset$
2. rewrite the system as
       $\begin{cases} \hat{a}x \equiv \hat{b} & \pmod{\hat{m}} \\ \hat{c}x \equiv \hat{d} & \pmod{\hat{n}} \end{cases},$
   where $\gcd(\hat{a}, \hat{m}) = \gcd(\hat{c}, \hat{n}) = 1$
3. multiply both sides of the first congruence by $\hat{a}^{-1}$, and both sides of the second by $\hat{c}^{-1}$, to obtain the equivalent system
       $\begin{cases} x \equiv b' & \pmod{\hat{m}} \\ x \equiv d' & \pmod{\hat{n}} \end{cases},$
4. if $\gcd(\hat{m}, \hat{n}) \nmid (b' - d')$, return $\emptyset$
5. else
   (a) if $\gcd(\hat{m}, \hat{n}) = 1$,
       i. compute the unique solution $x$ found by the Chinese Remainder Theorem
       ii. return the resulting solution
   (b) else
       i. substitute $x = b' + mk$ into $x \equiv d' \pmod{\hat{n}}$
       ii. return the resulting solution
```

Back-substitute to obtain

$$x = 99 \times 60 + 54 \equiv 5994 \pmod{9999} .$$

Verifying this answer is straightforward.

Theorem 1.116. *Algorithm 1.8 terminates correctly.*

Proof. Termination? Steps 5(a)(ii), 5(b)(i), and 5(b)(ii) always terminate. By Theorem 1.48, computing the gcd terminates, so computing the gcd in steps 1, 2, 4, and 5(a) always terminates. Computing a multiplicative inverse in step 3 requires the Extended Euclidean Algorithm, which terminates by Theorem 1.52, so step 3 terminates. By Theorem 1.7, step 5(b)(ii) terminates. Step 5(a)(i) requires us to solve a Chinese Remainder Theorem, and the proof of Theorem 1.114 shows that this is just a matter of twice solving a linear congruence; by Theorem 1.109, that also terminates. This covers all the steps, so the algorithm terminates.

Correctness? If the algorithm returns a result in steps 1, 4, or 5(a)(ii), then the Chinese Remainder Theorem guarantees correctness. The only other way it returns a result is in step 5(b)(ii), where Theorem 1.109 and the subsequent discussion imply that the solution is a result to the system. □

Remark. The result obtained from steps 5(a)(ii) and 5(b)(ii) in Algorithm 1.8 are not necessarily unique. See Exercise .

Exercises

Exercise 1.117. Solve the following linear congruences. If they cannot be solved, explain why not. If there is more than one solution among the canonical residues, list them all.

(a) $4x \equiv 7 \pmod 9$

(b) $100x \equiv 18 \pmod{112}$

(c) $100x \equiv 20 \pmod{112}$

Exercise 1.118. Solve

$$\begin{cases} 5x \equiv 3 & \pmod 7 \\ 4x \equiv 2 & \pmod 9 \end{cases} .$$

List all distinct solutions modulo 63.

Exercise 1.119. Solve

$$\begin{cases} 2x \equiv 1 & \pmod 7 \\ 7x \equiv 5 & \pmod{11} \\ 4x \equiv 7 & \pmod{15} \end{cases} .$$

Indicate the modulus in which this solution is unique.
Hint: Divide and conquer. First solve the first two congruences; by the Chinese Remainder Theorem, this gives a unique solution modulo 77. Let’s call that solution b; you now know that the

solution must satisfy the smaller system

$$\begin{cases} x \equiv b & \pmod{77} \\ 4x \equiv 7 & \pmod{15} \end{cases}.$$

You can solve this the same as before, and find your solution.

Exercise 1.120. Show that if $\gcd(a,b) \neq 1$ but divides m, then $ax \equiv b \pmod{m}$ still has a solution.
Hint: Consider the equation $ax = mq + b$. "Simplify" this to show that a solution exists. Explain why the solution works for $ax \equiv b \pmod{m}$.

Exercise 1.121. Solve

$$\begin{cases} x \equiv 12 & \pmod{15} \\ x \equiv 17 & \pmod{20} \end{cases}.$$

Hint: There is one unique solution modulo 60, and distinct 5 solutions modulo 300.

Exercise 1.122. Consider the system

$$\begin{cases} x \equiv a & \pmod{m} \\ x \equiv b & \pmod{n} \end{cases}$$

where $\gcd(a,m) = 1$ and $\gcd(b,n) = 1$, but $\gcd(m,n) \neq 1$. Suppose that $\gcd(m,n) \mid (b-a)$, as in Exercise 1.121. In this case, step 5(b)(ii) returns one answer to such a system.

(a) Use the algorithm, as well as insight from Exercise 1.121, to write a symbolic formula for one solution.

(b) Explain how to find *all* solutions.

(c) The solutions are incongruent modulo what number?

Sage supplement

Sage's `solve` command will solve regular equations. You need to specify the name of the variable to solve for.

```
sage: solve( 5*x == 2, x)
[x == (2/5)]
```

The result is a list of equations that satisfy the solution. Higher-degree equations will have multiple solutions.

```
sage: solve( 5*x^2 == 2, x)
[x == -1/5*sqrt(5)*sqrt(2), x == 1/5*sqrt(5)*sqrt(2)]
```

Sage will also solve a linear congruence, but it requires a different command. Instead of solve, we'll use solve_mod, which expects an equation as the first argument, and the modulus as the second.

```
sage: solve_mod( 5*x == 7, 12 )
[(11,)]
```

The solution is a list of incongruent values for each variable; when substituted for the variable(s), they make the congruence true. In this case, there is only one solution, and only one variable, so the list [...] gives us one solution (...) which lists only one number, 11.

Oftentimes there can be more than one incongruent solution to a linear congruence. In this case, the list will have more entries.

```
sage: solve_mod( 4*x == 8, 12 )
[(8,), (5,), (2,), (11,)]
```

There are four solutions! One of these solutions is obvious: $x = 2$. What about the others? It isn't hard to verify that in fact

$$4 \times 8 \equiv 4 \times 5 \equiv 4 \times 2 \equiv 4 \times 11 \equiv 8 \pmod{12},$$

but how would you know this in advance? You will explore this in the exercises.

There is also a more traditional way to solve linear congruences. Remember that Theorem 1.82 tells us we can think of a linear congruence $ax \equiv b \pmod{n}$ as $n \mid (ax - b)$. Choose $q \in \mathbb{Z}$ such that $ax - b = nq$ and we're looking at an equation whose solutions must be via integers.[12] Sage gives us a way to specify that via an assume command. This command allows us to specify many kinds of constraints on variables. To specify that the variable x is an integer, we use assume(x, 'integer'). We can then solve an equation with x, and only integer solutions will be allowed.

We illustrate this with the congruence $5x \equiv 7 \pmod{12}$. As per the discussion above, we consider the equation $5x = 12q + 7$. We need to define a symbol in Sage for the variable q, which we can do with the var command.

```
sage: var( 'q' )
sage: assume( x, 'integer' )
sage: assume( q, 'integer' )
sage: solve( 5*x == 12*q + 7 )
12*t_0 + 35
```

This tells us that for any integer t_0, $x = 12t_0 + 35$ will satisfy $5x = 12q + 7$. We know that this means it should satisfy $5x \equiv 7 \pmod{12}$, and in fact

$$5 \times (12t_0 + 35) = 60t_0 + 175 \equiv 0 + 7 \pmod{12},$$

[12] Equations of the form $ax + by = c$ where every constant variable is an integer are called *Diophantine equations*. Solving them is a major topic in Number Theory.

as desired. So $x \equiv 35 \pmod{12}$, or preferably we use its canonical linear residue, $x \equiv 11 \pmod{12}$.

Solving Chinese Remainder Theorem problems is also fairly easy in Sage; we use the `crt` command. It expects two arguments, and each of those arguments is to be a list of two integers:

$$\texttt{crt([a, b] , [m, n])} \text{ corresponds to } \begin{cases} x \equiv a & \pmod{m} \\ x \equiv b & \pmod{n} \end{cases}.$$

The coefficients of x must be 1; if they are not, we must first rewrite the system.

Example 1.123. To solve the system in Example 1.115,

$$\begin{cases} 5x \equiv 72 & \pmod{99} \\ 3x \equiv 4 & \pmod{101} \end{cases},$$

we first have to rewrite it using the multiplicative inverses. We took care of that already in Example 1.115, obtaining

$$\begin{cases} x \equiv 54 & \pmod{99} \\ x \equiv 35 & \pmod{101} \end{cases}.$$

We can finally apply `crt` to this form.

```
sage: crt( [ 54, 35 ], [ 99, 101 ] )
5994
```

This is the solution we expect.

Exercises

Exercise 1.124. Use Sage to solve the following linear congruences.

(a) $5123x \equiv 1001 \pmod{32003}$

(b) $7719x \equiv 10017 \pmod{35}$

(c) $1024x \equiv 256 \pmod{65536}$
Hint: In this last case there are a *lot* of solutions. Don't write down all of them: find a pattern $x = ai + b$, where a and b are fixed integers and i ranges from a smallest to a largest value. Be sure to indicate the smallest and largest values of i that describe a solution.

Exercise 1.125. Try using Sage to solve $6x \equiv 7 \pmod{12}$. What happens? Why?

Exercise 1.126. Use Sage to solve the following systems of linear congruences. Be sure to list all incongruent solutions if there are more than one. *Be careful with the latter:* Sage will not automatically tell you all incongruent solutions when there are more than one. You'll have to think about it a bit!

(a) $$\begin{cases} 38x & \equiv 52 \pmod{101} \\ 82x & \equiv \ \ 7 \pmod{103} \end{cases}$$

(b) $$\begin{cases} x & \equiv 17 \pmod{20} \\ x & \equiv 13 \pmod{32} \end{cases}$$

Exercise 1.127. Sage will not automatically solve systems of linear congruences of the form

$$\begin{cases} ax & \equiv b \quad \pmod{m} \\ cx & \equiv d \quad \pmod{n} \end{cases}.$$

We can however outline an algorithm to solve them, based on this section's discussions and the procedure of Example 1.115:

Algorithm 1.9 CRT with coefficients

Inputs

- $a, b, c, d \in \mathbb{N}$
- $m, n \in \mathbb{N}^+$

Outputs

- a solution to the system of linear congruences

$$\begin{cases} ax \equiv b & \pmod{m} \\ cx \equiv d & \pmod{n} \end{cases} \tag{1.11}$$

Do

1. if $\gcd(a, m) \nmid b$ or $\gcd(c, n) \nmid d$ then return $\emptyset$
2. find $\hat{a}, \hat{b}$ such that the system (1.11) is equivalent to

$$\begin{cases} x \equiv \hat{a} & \pmod{m} \\ x \equiv \hat{b} & \pmod{n} \end{cases}$$

3. use Sage to solve the system obtained in step 2 and return that solution

Implement this algorithm as a Sage procedure.
Hint: For step 2, you don't have to create any equations; you just have to find $\hat{a}$ and $\hat{b}$ so that you can use them in step 3. You can do this with Sage by computing $\hat{a} = a^{-1}b$ and $\hat{b} = c^{-1}d$.

1.7 Public-key encryption

We end this chapter with a famous application of the ideas we have studied: *secret communication.* Suppose that Person A and Person B want to exchange messages, but are afraid that Person E might overhear.[13] They need a function f that transforms a readable message m into an unreadable cipher c, typically using an encryption key e. That is,

$$c = f(m, e) \ .$$

They also need a way to *undo* the encryption.

In the modern age we use computers to do this. Computers work with numbers, so we need some way to turn letters into numbers. We will adopt a very simple method where

$$A \mapsto 0, \quad B \mapsto 1, \quad \ldots \quad Z \mapsto 25, \quad ; \mapsto 26,$$
$$, \mapsto 27, \quad . \mapsto 28, \quad ␣ \mapsto 29, \quad \text{-} \mapsto 30 \ .$$

(The symbol ␣ indicates a space.) This would encode the message

`STOP␣-␣DANGER␣AHEAD`

as

$$18\ 19\ 14\ 15\ 29\ 30\ 29\ 3\ 0\ 13\ 6\ 4\ 17\ 29\ 0\ 7\ 4\ 0\ 3\ .$$

The numbers are elements of $\mathbb{Z}_{31}$. We don't *have* to use the modulus 31; we can use any modulus that is sufficiently large to encode the alphabet. The benefit is that we can leverage modular arithmetic to find a way to communicate secretly.

Classical versus public-key encryption

The "classical" approach to encryption requires A and B to know both the encryption method f and the private key e. Only the ciphertext c is public knowledge, so E's challenge is to determine both f and e. Once E determines this information, decryption is a snap, since typically

$$m = f^{-1}(c, e) \ .$$

and it is "easy" to compute f^{-1} from f. For example;

- The *Cæsar cipher* consists of choosing some $k \in \mathbb{Z}$ and using

  $$f(m, e) = \overline{m + e} \ ,$$

  where the line over $m + e$ means to divide and take the remainder modulo 31 (or whatever modulus one chooses, only both A and B must know it). Decryption consists of computing

  $$f^{-1}(c, e) = \overline{c - e} \ .$$

  This cipher takes its name from Julius Cæsar; according to several ancient Romans, he used a version of this method.

[13] Authors often use "Alice," "Bob," and "Eve" instead of A, B, and E. In our internet economy one could well use "Amazon," "Buyer," and "Eavesdropper."

- The *Vigenère* cipher consists of choosing a short sequence $e_1, \ldots, e_\ell$ (sometimes corresponding to an easy-to-remember word) and enciphering several characters at a time

  $$f\left((m_1, \ldots, m_\ell), (e_1, \ldots, e_\ell)\right) = \left(\overline{m_1 + e_1}, \ldots, \overline{m_\ell + e_\ell}\right) .$$

  Decryption consists of computing

  $$f^{-1}\left((c_1, \ldots, c_\ell), (e_1, \ldots, e_\ell)\right) = \left(\overline{c_1 - e_1}, \ldots, \overline{c_\ell - e_\ell}\right) .$$

  It takes its name from Blaise de Vigenère, though Giovan Battista Bellasio discovered it. For a long time people considered the Vigenère cipher indecipherable if E does not know the key.

- A *one-time pad* uses for its key a sequence of random numbers $e_1, \ldots, e_\ell$ in exactly the same fashion as the Vigenère cipher, except that the sequence must be *long*, at least as long as the message. It takes its name from the fact that you use a one-time pad exactly once, then never again. Because of this, the cipher has been proved indecipherable if E does not know the key. Unfortunately, generating and storing sequences of random numbers is burdensome.

- A *stream cipher* uses for its key a sequence of *pseudo*-random numbers $e_1, e_2, \ldots$ in exactly the same fashion as the one-time pad. Here, "pseudo-random" means that the numbers are generated according to a formula designed to produce numbers that look random, even though they are not — the reasoning being that if you can generate the numbers according to a formula, then they aren't truly random.

- The *Navajo code talkers* were Navajo men who translated English messages to Navajo, which was then radioed between airplanes in the Pacific theater during World War II. The Japanese Navy had never heard anything like it before, and was completely unable to make sense of it.

Again, these techniques require both A and B to know *both* the method *and* the key, and indecipherability depends on keeping at least one of the two secret. This makes classical encryption practically difficult, as both A and B must not only keep a record of the keys — in a codebook, for instance — they must also keep the record hidden from E. Failure either to use a secure method or to keep the method secret forfeits the security.

- The American Department of State at one time used an encryption method so poor that every half-competent intelligence agency was reading our "secret" communications. One history of encryption called us the "laughing stock of the world."

- The German military in World War II used an encryption device called Enigma. The United States and Britain invested heavily in early computer technology precisely to decrypt Nazi communications. These efforts received an enormous boost when the Allies captured a codebook that a captured submarine's commander was unable to destroy before being boarded.

By contrast, public-key encryption works as follows.

- A chooses a method f and a public "encryption key e."
- A broadcasts in public that anyone who wishes to communicate secretly with A should use the method f and the key e.
- B computes and broadcasts $c = f(m, e)$, so that everyone now knows f, c, and e.
- A also has a second, private "decryption key" d. To decipher the method, E needs to find d.

This is much easier to deal with on a large scale than private encryption: A and B do not need to keep secret, hidden codebooks. Whenever they want to communicate, they broadcast clearly each other's encryption key. What's more, *anyone* can send messages securely to A using this method, not just B.

RSA encryption

RSA encryption takes its name from "Rivest, Shamir, Adelman," the mathematicians who first described the technique publicly. One convenient aspect of RSA is that encryption and decryption use the same mathematical operations; the only difference is in the key. Another convenient aspect is that a computer can perform the operations relatively quickly.

A does the following in preparation to receive messages.

- Choose two prime numbers, p and q.
- Compute $N = pq$.
- Let e be a number that is relatively prime to $\phi = (p-1)(q-1)$.
- Let d be the multiplicative inverse of e, modulo ϕ.
- Invite everyone to send messages using RSA, encryption key e, modulo N.

To send a message, B does the following.

- Compute, then broadcast, $c = m^e \pmod{N}$.

To decrypt the message, A does the following.

- Compute $x = c^d$.

Theorem 1.128. *If A and B perform the steps above, then $x \equiv m \pmod{m}$.*

We postpone the proof until we build up some background theory. To begin with, you are probably wondering why $\phi = (p-1)(q-1)$ is special. To explain this we need a new set: let $\mathbb{Z}_m^*$ be the subset of $\mathbb{Z}_m$ whose elements are all relatively prime to n.

Example 1.129. $\mathbb{Z}_{15}^* = \{1, 2, 4, 7, 8, 11, 13, 14\}$ and $\mathbb{Z}_{31}^* = \{1, 2, \ldots, 31\}$.

Lemma 1.130. *Let p and q be prime, and $N = pq$. The number $\phi = (p-1)(q-1)$ counts the elements of $\mathbb{Z}_N^*$. That is, $\phi = \left|\mathbb{Z}_{pq}^*\right|$.*

To help understand the proof, we illustrate it with an example.

Example 1.131. Let $p = 3$ and $q = 5$. We have $N = 15$ and $\phi = 2 \times 4 = 8$. To see why ϕ really does count the number of integers in $\{1, \ldots, 15\}$ that *are* relatively prime to 15, count the number of integers that are *not.*

Since $N = 3 \times 5$, and both 3 and 5 are prime, a number has a common divisor with N only if it is a multiple of 3 or 5. These are

$$\begin{array}{ll} 3, 6, 9, 12, 15 & \text{(multiples of 3)} \\ 5, 10, 15 & \text{(multiples of 5)}\ . \end{array}$$

The first sequence has 5 multiples of 3; the second has 3 multiples of 5. They have no common elements except the last one, 15. In fact, they *should* have nothing in common, since one consists of multiples of 3, the other consists of multiples of 5, and the least common multiple of 3 and 5 is 15. So the number of integers in $\{1, \ldots, 15\}$ that share a common divisor with 15 is

$$\underbrace{5}_{\text{multiples of 3}} + \underbrace{3}_{\text{multiples of 5}} - \underbrace{1}_{\text{extra 15}} .$$

Hence, the number of integers in $\{1, \ldots, 15\}$ that are relatively prime to 15 is

$$15 - (5 + 3 - 1) = 8\ ,$$

which is precisely the value of ϕ we computed above.

Proof of Lemma 1.130. Let $a \in \{0, 1, \ldots, N - 1\}$, and suppose that $\gcd(a, N) \neq 1$. By the Fundamental Theorem of Arithmetic, $\gcd(a, N)$ has a unique prime factorization, say

$$\gcd(a, N) = r_1 \cdots r_k\ .$$

By definition of divisibility, we can find $s \in \mathbb{N}$ such that $N = s \gcd(a, N)$, so by substitution

$$pq = sr_1 \cdots r_k\ .$$

By Euclid's Lemma, $p \mid s$ or $p \mid r_i$ for some $i = 1, \ldots, k$. Suppose $p \nmid r_i$ for any i; this forces $p \mid s$. Choose $t \in \mathbb{N}$ such that $pt = s$. We divide both sides by p and obtain the equation

$$q = tr_1 \cdots r_k\ .$$

By Euclid's Lemma, $q \mid t$ or $q \mid r_i$ for some $i = 1, \ldots, k$. Suppose $q \nmid r_i$ for any i; this forces $q \mid t$. Choose $u \in \mathbb{N}$ such that $qu = t$. We divide both sides by q and obtain the equation

$$1 = ur_1 \cdots r_k\ .$$

This happens only if $u = r_1 = \cdots = r_k = 1$. Recall that $r_1 \cdots r_k = \gcd(a, N)$; by substitution, $1 = \gcd(a, N)$. This contradicts the hypothesis that $\gcd(a, N) \neq 1$, so $p \mid r_i$ or $q \mid r_i$ for some i. Either way, p or q is a divisor of $\gcd(a, N)$, so p or q is a divisor of a.

We have shown that if $\gcd(a, N) \neq 1$, then p or q divides a. How many such a's are there in $\{0, 1, \ldots, N-1\}$? The multiples of p are

$$p,\ 2p,\ \ldots\ pq\ ;$$

the multiples of q are

$$q,\ 2q,\ \ldots pq\ .$$

The number pq appears in both sequences; are there others? Suppose i, j satisfy $ip = jq$. By Euclid's Lemma, $p \mid j$ or $p \mid q$. Both p and q are prime, and $p \neq q$, so $p \mid j$. Similarly, $q \mid i$. So the smallest number that is a multiple of both p and q is pq itself. Hence our sequences above are completely distinct except for pq itself, and there are

$$\underbrace{q}_{\text{multiples of } p} + \underbrace{p}_{\text{multiples of } q} - \underbrace{1}_{\text{extra } pq}$$

common multiples of p and q from $\{0, 1, \ldots, N-1\}$. These are the only numbers in $\mathbb{Z}_N^*$ that share a common divisor with N, so the number of integers in $\{0, 1, \ldots N-1\}$ that are *not* multiples of p or q are

$$\begin{aligned} N - (p + q - 1) &= pq - p - q - 1 \\ &= p(q-1) - (q-1) \\ &= (p-1)(q-1) \\ &= \phi\ , \end{aligned}$$

as claimed. □

More generally, suppose $\phi(m) = \left|\mathbb{Z}_m^*\right|$ for any integer n. This number has a useful property.

Theorem 1.132 (Euler's Theorem). *For any $a \in \mathbb{Z}_m^*$, $a^{\phi(m)} \equiv 1 \pmod m$.*

Example 1.133. In $\mathbb{Z}_{15}$, $2^8 \equiv 1 \pmod{15}$. One way to compute this is by evaluating

$$\underbrace{2 \times 2 \times \cdots \times 2}_{8 \text{ times}},$$

but a more clever way to do it is to realize that $8 = 2^3$ and compute

$$\left(\left(2^2\right)^2\right)^2\ .$$

If we reduce modulo 15 every chance we get, we see that in fact

$$\begin{aligned} 2^2 &= 4 \\ \left(2^2\right)^2 &= 4^2 = 16 \equiv 1 \\ \left(\left(2^2\right)^2\right)^2 &\equiv 1^2 = 1\ . \end{aligned}$$

Here we encountered 1 at 2^4, illustrating that we might meet it sooner than the $\phi(n)$ power. Nevertheless, we will still always reach it at $\phi(n)$.

Proof of Euler's Theorem. Let $a \in \mathbb{Z}_n$, and suppose a is relatively prime to n. The set $\mathbb{Z}_n$ is finite, so there can be only finitely many distinct powers of a, modulo n. Let $T = \left\{a, a^2, \ldots, a^k\right\}$ be a complete list of the powers of a. We now make two observations.

Claim. The elements of T are all distinct.

Subproof. Suppose $a^i = a^j$ in $\mathbb{Z}_n$ for some $i < j$. The exponents are natural, so we can choose i so that it is the smallest power of a for which this occurs. By hypothesis, a and n are relatively prime; this means that a has a multiplicative inverse, call it s. Multiply both sides of $a^i = a^j$ by s^{i-1} to obtain

$$a = a^{j-i+1} .$$

Remember that i was supposed to be the smallest positive power where repetition occurred, so we have just proved that $i = 1$. By substitution,

$$a = a^j , \quad a^2 = a^{j+1} , \quad \ldots ;$$

that is, all powers from j on simply repeat powers that already appear in T. Repetition cannot occur until after we have reached the jth power, which means $k = j - 1$ and the elements of T are indeed distinct.

Claim. $a^k = 1$ in $\mathbb{Z}_n$.

Subproof. We showed above that $a^{k+1} = a$. Multiply both sides by the inverse of a to see that $a^k = 1$.

We now perform the following iteration.

```
1. let $U_1 = T$
2. let $i = 2$
3. while $U_1 \cup \cdots \cup U_{i-1} \neq \mathbb{Z}_n^*$
    (a) let $b_i \in \mathbb{Z}_n^* \backslash (U_1 \cup \cdots \cup U_{i-1})$
    (b) let $U_i = \left\{ab_i, a^2b_i, \ldots, a_kb_i\right\}$
    (c) increment $i$ by 1
```

We claim this iteration terminates with $U_1 \cup \cdots \cup U_{\text{last}} = \mathbb{Z}_n^*$, no pair of distinct U's has even one element in common, and the U's all have the same size. We prove each claim individually.

Example 1.134. Before examining the claims, we illustrate the iteration on a concrete example. Let $n = 15$ and $a = 4$. We have $\mathbb{Z}_{15}^* = \{1, 2, 4, 7, 8, 11, 13, 14\}$. We start with

$$U_1 = T = \left\{4, 4^2, 4^3, \ldots\right\} = \{4, 1\} .$$

Notice that $k = 2$. Since $U_1 \neq \mathbb{Z}_{15}^*$, let $b = 2$ and we have

$$U_2 = \{2 \times 4, 2 \times 1\} = \{8, 2\} .$$

Since $U_1 \cup U_2 \neq \mathbb{Z}_{15}^*$, let $b = 7$ and we have

$$U_3 = \{7 \times 4, 7 \times 1\} = \{13, 7\} \ .$$

Since $U_1 \cup U_2 \cup U_3 \neq \mathbb{Z}_{15}^*$, let $b = 11$ and we have

$$U_4 = \{11 \times 4, 11 \times 1\} = \{14, 11\} \ .$$

The iteration has now terminated with $U_1 \cup U_2 \cup U_3 \cup U_4 = \mathbb{Z}_{15}^*$. Distinct U's have no elements in common, and they are all the same size.

Claim. The iteration terminates.

Subproof. Steps 1, 2, 3(a), 3(b), and 3(c) are simple assignments, so by themselves they do not inhibit termination. Only the repetition of step 3 might lead to a never-ending task, but that requires $U_1 \cup \cdots \cup U_{i-1} \neq \mathbb{Z}_n^*$. There are only finitely many elements of $\mathbb{Z}_n^*$, and each time we perform steps 3(a) and 3(b) we move at least one element of $\mathbb{Z}_n^*$ that is not in some U into a new U. Eventually, we run out of elements of $\mathbb{Z}_n^*$, so the iteration must terminate.

Claim. $U_1 \cup \cdots U_{\text{last}} = \mathbb{Z}_n^*$.

Subproof. We showed in the previous claim that the iteration must terminate, but by step 3 the iteration terminates only when $U_1 \cup \cdots \cup U_{\text{last}} = \mathbb{Z}_n$.

Claim. $U_i \cap U_j = \emptyset$ only if $i = j$.

Subproof. Suppose $U_i \cap U_j \neq \emptyset$ and let c be a common element. Without loss of generality, $i \leq j$. By construction, $c \in U_i$ implies that $c = a^\ell b_i$ for some $j = 1, \ldots, k$. Similarly, $c \in U_j$ implies that $c = a^m b_j$ for some $m = 1, \ldots, k$. By substitution,

$$a^\ell b_i = a^m b_j \ .$$

If $m \leq \ell$, then

$$a^{\ell-m} b_i = b_j \ ,$$

which means that $b_j \in T$. This contradicts the choice of b_j as *not* being an element of $U_1 \cup \cdots \cup U_{j-1}$. On the other hand, if $m > \ell$, recall that $a^k = 1$ and $m \leq k$, so

$$a^\ell b_i \times a^{k-m} = a^m b_j \times a^{k-m} = a^k b_j = 1 \times b_j = b_j \ ;$$

in other words,

$$a^{\ell+k-m} b_i = b_j \ .$$

Recall that $\ell < m$, so $\ell + k - m < m + k - m = k$, so $a^{\ell+k-m} b_i \in U_i$, the same contradiction as before.

Claim. The U's all have the same size.

Subproof. It suffices to show that each U has k distinct elements. We have already shown that $U_0 = T$ has k distinct elements. For any other i, suppose there exist $\ell, m \in \{1, \ldots, k\}$ such that $a^\ell b_i = a^m b_i$. Without loss of generality, $\ell \leq m$. Let s be the multiplicative inverse of a in $\mathbb{Z}_n$ and multiply both sides by a^ℓ; we have $b_i = a^{m-\ell} b_i$. Recall that $b_i \in \mathbb{Z}_n^*$; that is, b_i is also relatively prime to n, so it has a multiplicative inverse in $\mathbb{Z}_n$. Let t be the multiplicative inverse of b in $\mathbb{Z}_n$ and multiply both sides by t; we have $1 = a^{m-\ell}$. Since $0 < m, \ell \leq k$ and k is the smallest positive power where $a^k = 1$, we must have $m - \ell = 0$; in other words, $\ell = m$. The elements of U_i are thus all distinct, and it has k elements.

Our three claims show that the U's "divide" $\mathbb{Z}_n^*$ into equally-sized sets. Recall that $\phi = \left|\mathbb{Z}_n^*\right|$ If we put $\ell =$ last, then

$$\phi = k \times \ell \ .$$

Hence $a^\phi = a^{k\times\ell} = \left(a^k\right)^\ell = 1^\ell = 1$, as claimed. □

We can now prove RSA's correctness.

Proof of Theorem 1.128. Observe that

$$x = c^d \equiv (m^e)^d \equiv m^{ed} \pmod N \ .$$

If we can show that $m^{ed} \equiv m \pmod N$, then we will have proved the theorem.

By construction, $ed \equiv 1 \pmod \phi$. By definition, there exists $q \in \mathbb{N}$ such that $ed = 1 + q\phi$. Rewrite

$$x \equiv m^{ed} = m^{1+q\phi} = m \times \left(m^\phi\right)^q \pmod N \ .$$

Recall that $\phi = (p-1)(q-1)$ is the number of integers $\{0, 1, \ldots, N-1\}$ that are relatively prime to N. By Euler's Theorem, $m^\phi \equiv 1 \pmod N$. By substitution,

$$x \equiv m \times 1^q = m \pmod N \ .$$

The result of A's decryption is B's original message. □

Is RSA secure?

Before we consider this question, let's review what E knows about B's message. E knows that:

- B used RSA with the parameters N and e;
- in RSA, $N = pq$ where p and q are both prime;
- in RSA, the decryption exponent d is e's multiplicative inverse modulo $\phi(N)$;
- by Lemma 1.130, $\phi(N) = (p-1)(q-1)$;
- applying the Extended Euclidean Algorithm to e and $\phi(N)$ will reveal d.

The one thing E needs is the factorization $N = pq$. Once E knows p and q, E can compute $\phi(N)$ and d, then apply them to decrypt the message.

How would E determine p and q? Again, E *already knows* that N is the product of p and q, which are both prime. So E's task is as "easy" as this:

$$6 = 2 \times 3\,,$$

or as "easy" as this:

$$15 = 3 \times 5\,,$$

or as "easy" as this:

$$33 = 3 \times 11\,.$$

This is a grade-school problem! How is RSA secure?

Factoring is much easier for small numbers than large ones. In real-world RSA encryption, the primes used are quite large. One of the strange quirks of mathematics is that many grade-school problems are easy with small numbers, but unfeasible with large ones. Factoring a number into primes is one of those! In fact, RSA was first described in the late 1970s, and 40 years later there is still no practical way to defeat it. If you could find a practical way to factor two large primes, you would became very famous, possibly very rich, and also possibly very dead, depending on whom you informed first!

Other methods of public-key encryption exist, such as Elgamal or elliptic curve encryption. The long-term security of many of these methods is also not clear in general. Some schemes have been proposed, only to be cracked very quickly: problems that seem difficult to do can sometimes be cracked open quickly once someone finds the right approach — it's just that no one was motivated to find that approach before. Modern cryptography is, therefore, an exciting and active field of research that grows from number theory and algebra — two fields that were once considered as abstract and useless and mathematics could possibly be!

Exercises

Exercise 1.135. Encode the message `LEAVE␣ME␣ALONE` according to the technique described at the beginning of this section.
Hint: The problem asks you to en*code*, not to en*crypt*. Make sure you understand the difference.

Exercise 1.136. Another way to encode a message as numbers is to pair consecutive letters together, adding a random letter at the end if needed to get a pair. For example, the message

`STOP␣-␣DANGER␣AHEAD`

pairs up as

`ST, OP, ␣-, ␣D, AN, GE, R␣, AH, EA, DX.`

We then encode each pair `XY` as

$$x \times 31 + y\,,$$

where x is the value we'd use for `X` in the encoding described at the beginning of this section, and y is the value we'd use for `Y`. Complete the encoding of the message.

Exercise 1.137. Use a Cæsar cipher with $k = 3$ to encode `LEAVE␣ME␣ALONE`.

Exercise 1.138. The message `JUPNKAJADUI` has been encoded using a Cæsar cipher with $k = 9$. Decrypt the message.

Exercise 1.139. A stream cipher needs a function that generates pseudo-random numbers. One example generator is the following:

$$x_i = \begin{cases} 27, & i = 1 \\ 3x_{i-1} \in \mathbb{Z}_{31}, & i > 1 \end{cases}.$$

Compute the first 31 numbers generated by this sequence. Do you think the sequence looks random? Why or why not?

Exercise 1.140. In Example 1.133, we showed a shortcut for computing 2^8. Adapt this method to compute the following exponents relatively quickly. If you want to be *really* clever, use Euler's Theorem to make it even faster.

(a) 5^{36} in $\mathbb{Z}_{31}$

(b) 3^{78} in $\mathbb{Z}_{38}$

Exercise 1.141. Example 1.134 uses $\mathbb{Z}_{15}^*$ to illustrate the iterative generation of the U's in the proof of Euler's Theorem. Repeat the example with $\mathbb{Z}_{31}^*$ and $a = 2$. Observe how the U's "cover" $\mathbb{Z}_{31}^*$ completely, how they have no elements in common, and how they are all the same size.

Exercise 1.142. Consider the message (without the period)

`MEET␣AT␣DAWN`.

(a) Encode the message using the encoding described at the beginning of this section. (Don't forget the two spaces!)

(b) Use the RSA algorithm to encrypt the message, using parameters $N = 33$ and $e = 3$.

(c) What value of d would decrypt the message?

Sage supplement

Sage already incorporates fast exponentiation modulo n, *as long as you ask for it.* As it happens, there is a right way and a wrong way.

The *right way* is to define an integer in $\mathbb{Z}_n$. For instance:

```
sage: Z35 = ZZ.quo(35)
sage: Z35(2)^1000000000000
16
```

You'll notice that this computation resolves very quickly. By specifying that $2 \in \mathbb{Z}_{35}$, you have told Sage that you want to compute the power modulo 35. With this information, Sage takes advantage of all the mathematics we have described. On the other hand, suppose we write that second line *only slightly differently:*

```
sage: Z35(2^1000000000000)
```

This takes a lot longer, and might not even work on some machines. The author actually gave up after about a minute passed, so he never saw it produce 16.

What makes the second version take so much longer? The order of operations.

- The first version explicitly tells Sage that we want $2 \in \mathbb{Z}_{35}$, and only then do we raise it to the enormous exponent. Sage can first divide that exponent by $\phi = 24$, obtaining a remainder of 16. It then computes 2^{16}, dividing by 35 to keep the numbers small.

- The second version tells Sage that we want to compute $2^{1000000000000}$ first, and *only afterwards* should it move the result into $\mathbb{Z}_{35}$. Sage thus tries to compute $2^{1000000000000}$ as a regular integer, which takes a really long time[14] and requires a lot of memory,[15] either of which your machine may lack!

The upshot is that when implementing modular arithmetic, we have to take care to specify that our numbers are in $\mathbb{Z}_n$.

With that in mind, we can illustrate RSA encryption and decryption. We'll use $p = 5$ and $q = 7$, so that $N = 35$ and $\phi = 24$. For an encryption exponent we'll choose $e = 7$; then the decryption exponent is $d = 7$,[16] as Sage itself informs us via Bézout coefficients.

```
sage: euler_phi(35)
24
sage: e = 7
sage: xgcd(e, 24)
(1, 7, -2)
```

Encryption and decryption is then a simple matter of encoding the messages and raising them to powers. To encode, we use the Sage command `ord`, which converts a letter to a number. Under the default encoding, the letter `a` has the value 97, so we will subtract its value from `ord(`m`)` in order to obtain numbers between 0 and 35.

```
sage: def encode(m):
        return ord(m) - ord('a’')
```

[14] Try it by hand if you doubt this.

[15] If you do try it by hand, you will probably run out of paper.

[16] This is a terrible choice of parameters for the RSA algorithm; in no way should you have $e = d$. Choosing the right parameters is an art form in itself.

A few examples:

```
sage: encode('a')
0
sage: encode('m')
12
sage: encode('z')
25
```

Decoding requires us to perform the reverse operation. For this, Sage offers `chr`, which converts a number to a character. As before, we deal with numbers between 0 and 25, inclusive, but the default encoding gives "a" the value 97, so we need to add its value to whatever number comes in.

```
sage: def decode(n):
         return chr(n + ord('a'))
sage: decode(12)
'm'
sage: decode(25)
'z'
```

Once we have defined these procedures, encryption and decryption is a fairly straightforward matter using a `for` loop inside a list.

- To encrypt, we tell Sage to encode the letters as numbers, put the numbers in $\mathbb{Z}_{35}$, and raise them to the 7th power (e).
- To decrypt, we tell Sage to raise the numbers to the 7th power again (d), then decode the numbers as letters.

We have to take a little care in the second step, because the numbers resulting from operations in $\mathbb{Z}_{35}$ are not integers in Sage's opinion: they're elements of $\mathbb{Z}_{35}$, which are not quite the same thing. Fortunately, we can convert them back into integers using a simple command called `int`.

```
sage: [ Z35( encode(m) )^7 for m in 'secret']
[32, 4, 23, 3, 4, 19]
sage: [ decode( int(n^7) ) for n in _ ]
['s', 'e', 'c', 'r', 'e', 't']
```

If you examine this result carefully, you may wonder whether it is in fact secure. After all, `e` always turns into the same number (in this case, 4). It is well-known that some letters appear more often than others in English text, and "e" typically shows up the most. Hence, a simple *frequency analysis* would tell us which letter corresponded to which number, making it a snap to decrypt.

This skepticism is well warranted; real-life use of the RSA algorithm is not done in quite this fashion. A course on cybersecurity is well beyond the scope of these notes, but one thing we can do to make the algorithm somewhat more secure is to combine several letters at a time. We have to be careful here, as this simultaneously increases the minimum size of the modulus. For instance:

- If we combine two letters at a time, we need $N > 26^2$.
- If we combine three letters at a time, we need $N > 26^3$.
  ...
- If we combine ℓ letters at a time, we need $N > 26^\ell$.

This requires us to modify the encoding and decoding algorithms. Instead of encoding or decoding one letter at a time, we'll take ℓ at a time, and multiply each by a power of 26 to move it to the right place.

```
sage: def encode(M):
         result = 0
         for m in M:
            result *= 26
            result += ord('m') - ord('a')
         return result
```

The message `’secret’` now encodes in pairs as:

```
sage: encode('se'), encode('cr'), encode('et')
(472, 69, 123)
```

To encrypt it, we need to choose larger values of p and q, since $N = 35$ is too small to capture numbers like 472. How large *does* N need to be? We are encoding two letters at a time, which means we need

$$N > 26^2 = 676 .$$

If we choose $p = 29$ and $q = 31$, then $N = 899$ is sufficiently large. We have $\phi = 28 \times 30 = 840$. For an encryption exponent we choose $e = 11$; the decryption exponent will be 611.

```
sage: euler_phi(29*31)
840
sage: xgcd(11, 840)
(1, -229, 3)
sage: 840 - 229
611
```

(We cannot use -229 as an exponent, so we subtract 229 from 840 in order to find a positive multiplicative inverse of 11.)

We can now encrypt as before:

```
sage: Z899 = ZZ.quo(899)
sage: [ Z899(m)^11 for m in [ 'se', 'cr', 'et'] ]
[206, 764, 371]
```

When we encoded `secret` before, the `e`'s repeated. Here there is no repetition, which makes a frequency analysis impossible. With a long enough message, we would encounter some repetition, and some two-letter pairs, such as "an" or "th," appear more frequently than others.

Decryption remains a simple matter of applying the decryption exponent to the result. Decoding, however, requires us to separate the letters which encoding joined; since that involved multiplication, we can decode using the `%` and `/` operators.

```
sage: def decode(N):
         result = ''
         for n in N:
            m = N % 26
            result = chr(m + ord('a')) + result
            N -= m
            N /= 26
         return result
sage: [ decode(int(Z899(n)^611))
         for n in [206, 764, 371] ]
['se', 'cr', 'et']
```

We have successfully decrypted the message!

Exercises

Exercise 1.143. Reword the encryption of "secret" so that you encrypt and decrypt three letters at a time. This will require you to rewrite the `encode` and `decode` procedures.

Exercise 1.144. Implement in Sage the following algorithm to encrypt a message using the Cæsar cipher. Test it against the message `leavemealone`. There are no spaces, and the letters are all lower-case.

Exercise 1.145. In Exercise 1.139 you experimented with a pseudo-random number generator which gives us the numbers in the key. The following program will give us the first n numbers in a stream cipher's key.

Algorithm 1.10 Cæsar cipher

```
Inputs
  • $M$, a list of numbers corresponding to a message
  • $k \in \mathbb{N}$
Outputs
  • $C$, an encryption (or decryption) of $M$ using a Cæsar cipher with an offset of $k$
Do
  1. for each $i = 1, 2, \ldots, |M|$
     (a) let $c_i$ be the canonical residue of computing $m_i + k$ modulo 26
  2. Return $C = (c_1, c_2, \ldots, c_{|M|})$
```

```
sage: def stream(n):
         ZZ31 = ZZ.quo(31)
         result = [ ZZ31( 27 ) ]
         for each in range(2, n+1):
            result.append( 3*result[-1] )
         return result
```

The command `stream(10)` now gives us the following values:

```
sage: stream(10)
[27, 19, 26, 16, 17, 20, 29, 25, 13]
```

Imagine that you have just exchanged the keys for a stream cipher with a friend, and you want to encrypt the message `secret` using this cipher. Based on the discussion in the text, the following algorithm would do the trick.

Use this algorithm with the pseudo-random generator given above to encrypt the message, `secret`. For extra credit, implement the algorithm in Sage to verify your work.

Algorithm 1.11 Encryption via stream cipher

```
Inputs
  • $A$, a sequence of $n$ letters
  • $N \in \mathbb{N}^+$
Outputs
  • $C$, the text $M$ encrypted by a stream cipher
Do
  1. let $k_1, k_2, \ldots\ k_n$ be the first $n$ numbers of the stream cipher's key
  2. let $b_i$ be the encoding of $a_i$ (the $i$th letter of $a$)
  3. for $i \in \{1, \ldots, n\}$
     (a) let $c_i = \overline{b_i + k_1}$, where the modulus is $N$
  4. return $C = (c_1, \ldots, c_n)$
```

Chapter 2

Solving polynomial equations

This chapter aims to show that we can generalize the previous chapter's concepts of modularity from *integers* to *polynomials*. We can then use this powerful tool to solve polynomial equations. First, however, we have to review the behavior of polynomials in the "ordinary" sense that you are used to.

2.1 Polynomial arithmetic over $\mathbb{Z}$ and $\mathbb{Q}$

A ***polynomial*** f *in the* ***indeterminate*** x *over the integers* has the form[1]

$$f = a_n x^n + \cdots + a_1 x + a_0$$

where $n, a_0, a_1, \ldots, a_n \in \mathbb{N}$. We call a_0, a_1, ..., a_n the ***coefficients*** of f. If $a_n \neq 0$, then we say that f has ***degree*** n, and write $\deg(f) = n$. The degree is always a natural number, and is defined only if the polynomial is nonzero. We say that $\deg(0)$ is undefined.

We call $a_n x^n$ the ***leading term*** of f, and write $\operatorname{lt}(f)$ for short. We call the coefficient of $\operatorname{lt}(f)$ the ***leading coefficient***, and write $\operatorname{lc}(f)$ for short.

We write $\mathbb{Z}[x]$ for the set of all polynomials in x over the integers. If instead the a's are rational numbers, then we say that f is a *polynomial in x over the rationals*. We write $\mathbb{Q}[x]$ for the set of all polynomials in y over the rationals. Naturally, $\mathbb{Z}[x] \subseteq \mathbb{Q}[x]$. Throughout this section, when we speak of "a polynomial" without reference to whether it is over $\mathbb{Z}$ or over $\mathbb{Q}$, the reader should infer that the discussion applies to either case.

Example 2.1. Let $g = x^2 + 1$ and $h = 8x^3 - (1/27)$. Both $g, h \in \mathbb{Q}[x]$, whereas only $g \in \mathbb{Z}[x]$. As for the degrees and leading terms, $\deg(g) = 2$, $\deg(h) = 3$, $\operatorname{lt}(g) = x^2$, and $\operatorname{lt}(h) = 8x^3$.

While the use of x is fairly common, we can use other symbols for the indeterminate.

Example 2.2. The polynomial $2/3 y^3 - 3y$ is an element of $\mathbb{Q}[y]$.

Two polynomials are ***equal*** if they are both zero, or if their degree is equal and their coefficients are equal.

[1] It is common to write $f(x)$ instead of just f, but we will generally stick with f unless we want to emphasize that it is a polynomial in x.

We trust that you are familiar with operations on polynomials, such as addition, subtraction, and multiplication, though not necessarily division, which we turn to in a moment. However, you may not be very comfortable with the formula for polynomial multiplication:

$$pq = \sum_{i=0}^{m+n} \left(\sum_{j+k=i} a_j b_k \right) x^i \ . \tag{2.1}$$

Don't let the summation scare you; break it down and read it! It says that pq is the sum of all terms $a_j b_k x_i$ where i ranges from 0 to $m+n$ and $j+k=i$. Try this out with a few polynomials to see that it is, in fact, true. Don't let the sums and subscripts intimidate you; *work with them!* Get to know them! This is an essential part of mathematical notation.

Example 2.3. Here's an example to demonstrate. Suppose $p = 6x^2 + 4$ and $q = 4x^3 - 2x - 1$. You know from past experience that $pq = 24x^5 + 4x^3 - 6x^2 - 8x - 4$.

How does this compare to equation (2.1)? We have $m = 2$; $n = 3$; $p = a_2x^2 + a_1x + a_0$ where $a_2 = 6$, $a_1 = 0$; and $a_0 = 4$, while $q = a_3x^3 + a_2x^2 + a_1x + a_0$ where $b_3 = 4$, $b_2 = 0$, $b_1 = -2$, $b_0 = -1$. Substituting into (2.1),

$$\begin{aligned}
pq &= \sum_{i=0}^{m+n} \left(\sum_{j+k=i} a_j b_k \right) x^i \\
&= \left(\sum_{j+k=5} a_j b_k \right) x^5 + \left(\sum_{j+k=4} a_j b_k \right) x^4 + \left(\sum_{j+k=3} a_j b_k \right) x^3 \\
&\quad + \left(\sum_{j+k=2} a_j b_k \right) x^2 + \left(\sum_{j+k=1} a_j b_k \right) x^1 + \left(\sum_{j+k=0} a_j b_k \right) x^0 \\
&= a_2b_3x^5 + (a_2b_2 + a_1b_3)\,x^4 + (a_2b_1 + a_1b_2 + a_0b_3)\,x^3 \\
&\quad + (a_2b_0 + a_1b_1 + a_0b_2)\,x^2 + (a_1b_0 + a_0b_1)\,x + (a_0b_0) \\
&= (6 \times 4)\,x^5 + (6 \times 0 + 0 \times 4)\,x^4 + [6 \times (-2) + 0 \times 0 + 4 \times 4]\,x^3 \\
&\quad + [6 \times (-1) + 0 \times (-2) + 4 \times 0]\,x^2 + [0 \times (-2) + 4 \times (-2)]\,x + [4 \times (-1)] \\
&= 24x^5 + 4x^3 - 6x^2 - 8x - 4 \ .
\end{aligned}$$

Try this with some other polynomials until you get the hang of how it works. It isn't hard; it's just tedious.

The next two theorems should not surprise you. As they are relatively simple to prove, but a little tedious to write, we leave most of the details to the exercises.

Theorem 2.4. *Addition, subtraction, and multiplication of polynomials over $\mathbb{Z}[x]$ is closed. Similarly, addition, subtraction, and multiplication of polynomials over $\mathbb{Q}[x]$ is closed. Moreover, if p and q are polynomials in either $\mathbb{Z}[x]$ or $\mathbb{Q}[x]$, then*

- *if $p + q \neq 0$, then $\deg(p+q) \leq \max(\deg(p), \deg(q))$;*

- *if* $p-q \neq 0$, *then* $\deg(p-q) \leq \max(\deg(p), \deg(q))$; *and finally,*
- *if* $pq \neq 0$, *then* $\deg(pq) = \deg(p) + \deg(q)$ *and* $\operatorname{lt}(pq) = \operatorname{lt}(p) \cdot \operatorname{lt}(q)$.

Proof. For addition and subtraction, see Exercise 2.13.

For multiplication, either $pq = 0$ or it is not. If $pq = 0$, then it is in $\mathbb{Z}[x]$, and thus in $\mathbb{Q}[x]$, by virtue of all its coefficients being $0 \in \mathbb{Z}$.

Suppose, then, that $pq \neq 0$. Choose $m, n \in \mathbb{N}$ such that $\deg(p) = m$ and $\deg(q) = n$. Now choose $a_0, \ldots, a_m \in \mathbb{Z}$ and $b_0, \ldots, b_n \in \mathbb{Z}$ such that

$$p = a_m x^m + \cdots + a_1 x + a_0 \quad \text{and} \quad q = b_n x^n + \cdots + b_1 x + b_0 \ .$$

Recall that

$$pq = \sum_{i=0}^{m+n} \left(\sum_{j+k=i} a_j b_k \right) x^i \ .$$

Multiplication of integers (respectively, rationals) is closed; hence, each $a_j b_k$ is also an integer (resp., rational number). Recall further that the addition of integers (resp., rational numbers) is closed; hence, each $\sum a_j b_k$ is an integer (resp., rational number). In other words, the coefficient of each x^i is an integer (resp., rational number), and $pq \in \mathbb{Z}[x]$. That is, multiplication of polynomials is closed.

As for the degree, $\deg(p) = m$ and $\deg(q) = n$ means that $a_m, b_n \neq 0$, so by the zero product property $a_m b_n \neq 0$. This is the only pair of coefficients a_j, b_k such that $j + k = m + n$; every other pair of terms produces a term of smaller degree. Hence $\operatorname{lt}(pq) = a_m b_n x^{m+n} = \operatorname{lt}(p)\operatorname{lt}(q)$ and $\deg(pq) = m + n$. □

Theorem 2.5.
- *Polynomial addition is commutative, associative, and invertible, and it has an identity: the zero polynomial.*
- *Polynomial multiplication is commutative and associative, and it has an identity: the polynomial* 1. *However, polynomial multiplication is not generally invertible.*

Proof. As before, let p and q be polynomials of degree m and n, respectively, in either $\mathbb{Z}[x]$ or $\mathbb{Q}[x]$. Choose a's and b's such that

$$p = a_m x^m + \cdots + a_1 x + a_0 \quad \text{and} \quad q = b_n x^n + \cdots + b_1 x + b_0 \ .$$

Without loss of generality, we may assume that $m \geq n$. By definition of polynomial addition,

$$p + q = a_m x^m + \cdots + a_{n+1} x^{n+1} + (a_n + b_n) x^n + \cdots (a_1 + b_1) x + (a_0 + b_0) \ .$$

Also by definition of polynomial addition,

$$q + p = a_m x^m + \cdots + a_{n+1} x^{n+1} + (b_n + a_n) x^n + \cdots (b_1 + a_1) x + (b_0 + a_0) \ .$$

A naïve glance at $p+q$ and $q+p$ might make us think that they are different, as the coefficients of x^n look different. However, recall from Section 1.2 that addition of natural numbers and integers is commutative! Apply the commutative property to rewrite

$$q + p = a_m x^m + \cdots + a_{n+1} x^{n+1} + (a_n + b_n) x^n + \cdots (a_1 + b_1) x + (a_0 + b_0) \ ,$$

and it becomes clear that $p + q = q + p$.

The proof of the remaining properties is similar in spirit. □

The proof of Theorem 2.4 uses the zero product property of the rational numbers. We saw that the integers modulo m do not necessarily satisfy this property, so it is a good idea to check whether polynomials satisfy the zero product property. If this is true for $f, g \in \mathbb{Q}[x]$, then it is also true for $f, g \in \mathbb{Z}[x]$.

Theorem 2.6. *Polynomials over the rationals satisfy the zero product rule. That is, if $f, g \in \mathbb{Q}[x]$ and $fg = 0$, then $f = 0$ or $g = 0$.*

Proof. We prove the statement's contrapositive that if $f, g \neq 0$, then $fg \neq 0$.[2] Suppose that neither f nor g is a zero polynomial. Let $m = \deg(f)$ and $n = \deg(g)$. As in the proof of Theorem 2.4, $m + n$ is the largest possible degree of any product of a term of f and a term of g. By the zero product property of rational numbers, $\operatorname{lc}(f) \times \operatorname{lc}(g) \neq 0$, so fg has a nonzero coefficient at that degree. By definition, $fg \neq 0$. □

Roots of polynomials

Recall that if f is a polynomial, then "$f(a)$" means to "replace every x in f by a". We say that f *has a* ***root*** *in a set S* if we can find $s \in S$ such that $f(s) = 0$. We also say that s is a *root* of f.

Example 2.7. Recall from Example 2.1 the polynomials $g = x^2 + 1$ and $h = 8x^3 - (1/27)$. The polynomial g has no root in $\mathbb{R}$, as $r^2 + 1 \geq 1 > 0$ for every $r \in \mathbb{R}$. As $\mathbb{Q} \subseteq \mathbb{R}$, this means g has no root in $\mathbb{Q}$, either.

The polynomial h has a root in $\mathbb{Q}$. The root is $1/6$, since

$$h(1/6) \quad = \quad 8\left(\frac{1}{6}\right)^3 - \frac{1}{27} \quad = \quad 2^3 \times \frac{1}{2^3} \times \frac{1}{3^3} - \frac{1}{3^3} \quad = \quad 0\,.$$

Amazingly, we can decide whether a polynomial over the rationals has roots by turning it into a polynomial over the integers.

Theorem 2.8. *Let $f \in \mathbb{Q}[x]$ with degree n and coefficients $a_0, a_1, \ldots, a_n$. Let d be the least common denominator of the nonzero coefficients of f, and set $b_i = a_i d$ for each $i = 1, \ldots, n$. Let*

$$g = b_n x^n + \cdots + b_1 x + b_0\,.$$

Let S be any set. Then f has a root in S if and only if g does.

(Keep in mind that Theorem 2.8 applies to $f \in \mathbb{Z}[x]$ as well as $f \in \mathbb{Q}[x]$.)

Example 2.9. Recall from Example 2.1 the polynomial $h = 8x^3 - (1/27)$. We saw earlier that h has the root $1/6$. The denominators of h's coefficients are 1 and 27; the least common denominator is $d = 27$. If we multiply each coefficient of h by d, we have the polynomial

$$g = 216x^3 - 1\,.$$

It is easy to verify that $g(1/6) = 216(1/6)^3 - 1 = 216(1/216) - 1 = 0$.

[2] Recall that if a statement has the form "if A, then B," its *contrapositive* has the form "if not B, then not A." A statement and its contrapositive are logically equivalent.

Proof of Theorem 2.8. Suppose f has a root $s \in S$. By definition, $f(s) = 0$, which means that

$$a_n s^n + \cdots + a_1 s + a_0 = 0 \ .$$

Multiply both sides of the equation by d; by distribution, the associative property, and the commutative property,

$$(a_n d) s^n + \cdots + (a_1 d) s + a_0 d = 0 \ .$$

By substitution,

$$b_n s^n + \cdots + b_1 s + b_0 = 0 \ ;$$

that is, $g(s) = 0$. By definition, g has a root s in S.

For the converse, assume g has a root s in S. By definition,

$$b_n s^n + \cdots + b_1 s + b_0 = 0 \ .$$

Multiply both sides by $1/d$ to obtain $f(s) = 0$, which by definition implies f has a root s in S. □

The upshot is that if we want to find roots of polynomials over the integers, we can start with polynomials over the rationals; solve those, then return to polynomials over the integers. The reverse is also true. This tack will prove useful in the future.

Exercises

Exercise 2.10. Let $f = 2x^3 + x + 1$, $g = x^4 - x - 1$, and $h = -2x^3 + x - 1$.

(a) Evaluate $f + g$, $f + h$, and $g + h$. For which expressions does the degree change? Why?

(b) Evaluate fg, fh, and gh.

(c) For fg, indicate the values of m, n, a_j, and b_k that satisfy equation (2.1). Work out the formula to show that you get the same answer as in part (b).

Exercise 2.11. Find all roots of $f = \left(x^4/3\right) - 9x^2 + (14x/3) + 40$.
Hint: Take the advice of Theorem 2.8 and turn this into a polynomial over the integers. Then factor by grouping.

Exercise 2.12. Show that if r is a root of $x^{2n} - a$, then $-r$ is also a root.
Hint: $x^{2n} = \left(x^2\right)^n$, and what is the square of a negative?

Exercise 2.13. Show that if f and g are polynomials, then

$$\deg(f \pm g) \le \max(\deg(f), \deg(g)) \ .$$

Hint: Don't forget to consider all possibilities. Exercise 2.10(a) should help examine the behavior.

Exercise 2.14. A remark right before Theorem 2.6 states that "If this is true for $f, g \in \mathbb{Q}[x]$, then it is also true for $f, g \in \mathbb{Z}[x]$." Why?

Sage supplement

You can define a polynomial in Sage as either an *expression* or a *function*. The difference is subtle but important; a function "knows" its independent variable, while an expression doesn't. This makes it easier to substitute into a function than into an expression.

```
sage: f(x) = x^3 + 3
sage: f
x |--> x^3 + 3
sage: g = x^3 + 3
sage: g
x^3 + 3
sage: f(3)
30
sage: g(3)
__main__:3:  DeprecationWarning:  Substitution using
function-call syntax and unnamed arguments is deprecated
and will be removed from a future release of Sage; you can
use named arguments instead, like EXPR(x=..., y=...)  See
http://trac.sagemath.org/5930 for details.
30
```

Observe that Sage displays `f` differently than `g`: the expression `x |--> x^3+3` indicates that `f` is a function that maps x to $x^2 + 3$, so that `f` "knows" its independent variable. On the other hand, `g` does *not* know its independent variable, which is why the substitution raises a `DeprecationWarning`. Sage guesses that `x` is the independent variable for `g`, but it doesn't actually *know* this. You can get around this by telling Sage explicitly how to substitute the value:

```
sage: g(x=3)
30
```

Giving a full description of Sage's abilities with polynomials is beyond the scope of this text. However, we can show how to perform some basic ideas. Beyond that, it is really up to the reader to *experiment* and try out different things.

Basic information about polynomials is generally available via methods.

method	**description**
`.coefficient(x`^{a}`)`	the polynomial's coefficient of x^a
`.coefficients()`	the polynomial's coefficients, in the form `[[` *exponent* `,` *coefficient* `], … ]`
`.degree(x)`	the polynomial's degree in `x`
`.expand()`	the polynomial, expanded as a product
`.factor()`	the polynomial's factorization
`.leading_coefficient(x)`	the polynomial's leading coefficient with respect to `x`
`.operands()`	the polynomial's terms
`.roots()`	returns the polynomial's roots *and* their multiplicities, in the form `[(root, mult), … ]`
`.simplify()`	simplifies the polynomial, though not as much as you might like; see `.expand`

Let's verify that some of these commands are compatible with the behavior we saw in Example 2.1.

```
sage: g, h = x^2 + 1, 8*x^3 - 1/27
      g.degree(), h.degree()
(2, 3)
      g.leading_coefficient(x), h.leading_coefficient(x)
(2, 3)
```

Let's try entering the polynomial of the next example.

```
sage: p = 2/3*y^2 - 3*y
Traceback (click to the left of this block for traceback)
...
NameError: name 'y' is not defined
```

The error here seems fairly obvious, but how do we fix it? You can define a new indeterminate in Sage using the `var` command. Supply as its argument the indeterminate's name *in quotes:*

```
sage: var('y')
y
sage: p = 2/3*y^2 - 3*y
sage: p.degree(x)
0
sage: p.degree(y)
2
```

Sage acknowledges after the first line the definition of the indeterminate `y`. You can now define a polynomial `p` in `y`, and determine its degree. You must tell Sage which variable, even where there is only one.

Let's continue to Example 2.3.

```
sage: p, q = 6*x^2 + 4, 4*x^3 - 2*x - 1
sage: p * q
2*(4*x^3 - 2*x - 1)*(3*x^2 + 2)
```

While this is obviously correct, you probably expected a different answer! Rather than "multiply it out," Sage merely factored `p`'s common factor. Sage's rules for simplification are not necessarily the rules you expect! Nevertheless, you can force the expansion using the `.expand` method:

```
sage: (p * q).expand()
24*x^5 + 4*x^3 - 6*x^2 - 8*x - 4
```

This time, we have the same result as Example 2.3.

We move to Example 2.7. Remember that we defined `h` as an expression, so if we want to substitute, we have to specify the variable of substitution.

```
sage: h(x=1/6)
0
```

No surprises here — unless you forgot to specify `x=`, in which case you may have encountered the `DeprecationWarning` again. On the other hand, you might not see it even when you forget `x=`. A `DeprecationWarning` will appear only once in any Sage session; to see it again, you have to restart the worksheet. It's important to avoid it — the message says that this facility will one day be removed, so you need to adopt the proper syntax. But it's not a "full" error, so it won't appear every time you make the mistake.

Exercises

Exercise 2.15. Use Sage to verify your answers in Exercise 2.10.

Exercise 2.16. Use the `.roots` method to find all the roots of $f = \left(x^4/3\right) - 9x^2 + \left(14x/3\right) + 40$. Compare this with your answer to Exercise 2.11.
Hint: Dont' forget to put parentheses after the `.roots` method; otherwise you'll just get a strange-looking message.

Exercise 2.17. Use the `.roots` method to find all the roots of $h = x^4 + x^3 - 6x^2 - 4x + 8$. Then use the `.factor` method to factor h.

Recall that the `.roots` method gives us *both* roots *and* multiplicities. Where do the root's multiplicities appear in the factorization?

2.2 Polynomial division

We can divide polynomials into quotient and remainder with a result that is similar to that of dividing integers. Even the process is similar.

Theorem 2.18 (The division theorem for $\mathbb{Q}[x]$)**.** *Let $f, d \in \mathbb{Q}[x]$ with $d \neq 0$. There exist $q, r \in \mathbb{Q}[x]$ such that*

$$f = qd + r \quad \textit{and} \quad \textit{either } r = 0 \textit{ or } \deg(r) < \deg(d)\ .$$

In addition, q and r are uniquely determined by f and d.

Algorithm 2.1 on the following page, the commonly-taught algorithm of "long division of polynomials," produces the result we want. Notice that it closely resembles Algorithm 1.1 on page 16.

Example 2.19. We apply Algorithm 2.1 to $f = x^5 + 2x^2 + 1$ and $d = 2x^3 + x$.

- In step 1 we set $r = x^5 + 2x^2 + 1$ and $q = 0$.
- Since $r \neq 0$ and $\deg(r) = 5 > 3 = \deg(d)$, we perform step 2.
  - We set $t = x^5/2x^3 = x^2/2$.
  - Add t to q, resulting in $q = x^2/2$.
  - Subtract td from r, resulting in

$$r = \left(x^5 + 2x^2 + 1\right) - \frac{x^2}{2}\left(2x^3 + x\right) = -\frac{x^3}{2} + 2x^2 + 1\ .$$

- Since $r \neq 0$ and $\deg(r) = 3 = \deg(d)$, we perform step 2.
  - We set $t = (-x^3/2)/2x^3 = -1/4$.
  - Add t to q, resulting in $q = \left(x^2/2\right) - (1/4)$.
  - Subtract td from r, resulting in

$$r = \left(-\frac{x^3}{2} + 2x^2 + 1\right) - \left(-\frac{1}{4}\right)\left(2x^3 + x\right) = 2x^2 + \frac{x}{4} + 1\ .$$

- At this point $r \neq 0$ but $\deg(r) = 2 < \deg(d)$, so we proceed to step 3 and return q and r.

It is easy to verify that

$$\begin{aligned} qd + r &= \left(\frac{x^2}{2} - \frac{1}{4}\right)\left(2x^3 + x\right) + \left(2x^2 + \frac{x}{4} + 1\right) \\ &= \left(x^5 - \frac{x}{4}\right) + \left(2x^2 + \frac{x}{4} + 1\right) \\ &= x^5 + 2x^2 + 1 \\ &= f\ . \end{aligned}$$

Algorithm 2.1 Polynomial division

```
inputs
  • $f, d \in \mathbb{Q}[x]$
outputs
  • $q, d \in \mathbb{Q}[x]$ such that
      – $f = qd + r$, and
      – either $r = 0$ or $\deg(r) < \deg(d)$
do
  1. let $r = f$, $q = 0$
  2. while $r \neq 0$ and $\deg(r) \geq \deg(d)$
      (a) let $t = {}^{\mathrm{lt}(r)}/_{\mathrm{lt}(d)}$
      (b) add $t$ to $q$
      (c) subtract $td$ from $r$
  3. return $q$ and $r$
```

Proof of Theorem 2.18. If Algorithm 2.1 terminates correctly, the resulting q and r will satisfy Theorem 2.18, so we prove that Algorithm 2.1 terminates correctly.

Termination? If $f = 0$, then step 1 sets $q = 0$ and $r = f = 0$, so nothing happens at step 2, and step 3 returns $q = r = 0$, in which case

$$qd + r = 0 = f \ .$$

Not only has the algorithm terminated, we see that the output is correct.

Otherwise, $f \neq 0$. Step 1 sets $q = 0$ and $r = f$. If $\deg(f) < \deg(d)$, then nothing happens at step 2, and step 3 returns $q = 0$ and $r = f$, in which case

$$qd + r = f \ ,$$

and $\deg(r) = \deg(f) < \deg(d)$. Not only has the algorithm terminated, we see that the output is correct.

That leaves the case $f \neq 0$ and $\deg(f) \geq \deg(d)$. We claim that every time we perform step 2, the degree of r decreases. To see why, notice that we choose t such that, by substitution,

$$t \times \mathrm{lt}(d) = {}^{\mathrm{lt}(r)}/_{\mathrm{lt}(d)} \times \mathrm{lt}(d) = \mathrm{lt}(r) \ .$$

Subtracting td from r thus cancels $\mathrm{lt}(r)$, leaving us with a polynomial of smaller degree.

Recall that the degree of a polynomial is a natural number. If we denote the degrees of r on each pass through the loop of step 2 as $n_0, n_1, \ldots$, then $n_0 > n_1 > \cdots$. This is a nonincreasing sequence of natural numbers. By Theorem 1.28, this sequence must eventually stabilize, so we cannot perform step 2 indefinitely. Eventually we must pass on to step 3, which terminates the algorithm.

Correctness? We have two things to prove: that $f = qd + r$, and that $r = 0$ or $\deg(r) < \deg(d)$. We consider the second one first.

- To show that $r = 0$ or $\deg(r) < \deg(d)$ we have two subcases.
  - If the returned value is $r = 0$, then we are done.
  - Otherwise, the condition on step 2 requires the algorithm to continue as long as $\deg(r) \geq \deg(d)$. We now know the algorithm terminates, so the loop cannot continue indefinitely, so the values returned in step 3 satisfy $\deg(r) < \deg(d)$.
- To show that $f = qd + r$ we again have two subcases.
  - If the algorithm does not perform step 2, then we saw already that $f = qd + r$.
  - Otherwise, enumerate each t computed in step 2(a) of the algorithm as $t_0, t_1, \ldots, t_{\text{last}}$. The algorithm returns

    $$q = t_0 + t_1 + \cdots t_{\text{last}} \quad \text{and} \quad r = f - t_0 d - t_1 d - \cdots - t_{\text{last}} d \ .$$

    By substitution,

    $$\begin{aligned} qd + r &= (t_0 + \cdots + t_{\text{last}})\, d + (f - t_0 d - \cdots - t_{\text{last}} d) \\ &= (t_0 d + \cdots + t_{\text{last}} d) + (f - t_0 d - \cdots - t_{\text{last}} d) \\ &= f \ . \end{aligned}$$

We still have to show that q and r are unique. Suppose that in addition to q and r, we can find $\hat{q}, \hat{r} \in \mathbb{Q}[x]$ that satisfy the theorem. By substitution,

$$qd + r = \hat{q}d + \hat{r} \ .$$

Rewrite as

$$(q - \hat{q})\, d = \hat{r} - r \ .$$

By Theorem 2.4, either $\hat{r} - r = 0$ or $\deg(\hat{r} - r) \leq \max(\deg(\hat{r}), \deg(r)) < \deg d$. Similarly,[3] $q - \hat{q} = 0$ or $\deg((q - \hat{q})\, d) = \deg(q - \hat{q}) + \deg(d) \geq \deg(d)$. The degree of the left hand side cannot be smaller than the degree of the right hand side; they have to be equal. We conclude that $\hat{r} - r = 0$ and $q - \hat{q} = 0$; or, $r = \hat{r}$ and $q = \hat{q}$.

Regardless of the situation, the outputs of Algorithm 2.1 satisfy the stated requirements. The algorithm terminates correctly. As per the discussion at the beginning of the proof, this proves Theorem 2.18. □

[3] The zero product property has an implied role here; see if you can spot it!

Theorem 2.18 makes promises about polynomials with *rational* coefficients. Naturally, this implies a promise about polynomials with *integer* coefficients; but it only promises that if we divide them, we obtain a quotient and remainder with *rational* coefficients, not *integer* coefficients. In fact, in Example 2.19, both f and d have integer coefficients, but both q and r have rational coefficients.

If we tinker slightly with the theorem's conclusion, we can divide *and still* obtain a quotient and remainder that are also integer polynomials.

Example 2.20. Recall the result of Example 2.19: the polynomials $f = x^5 + 2x^2 + 1$, $d = 2x^3 + x$, $q = {}^{x^2}\!/_2 - {}^1\!/_4$, and $r = 2x^2 + ({}^x\!/_4) + 1$ satisfy

$$f = qd + r \; .$$

How can we eliminate all fractions from the right hand side of this equation? Clear the denominators! (This is mathematical jargon for, "Multiply both sides by the least common denominator.") That gives us

$$4 \times f = \hat{q}d + \hat{r}$$

where

$$\hat{q} = 2x^2 - 1 \quad \text{and} \quad \hat{r} = 8x^2 + x + 4 \; .$$

We have found an *integer multiple* of f that divides by d into a quotient and remainder over the integers.

Corollary 2.21 (The division theorem for $\mathbb{Z}[x]$)**.** *Suppose that $f, d \in \mathbb{Z}[x]$ with $d \neq 0$. There exist $q, r \in \mathbb{Z}[x]$ and a nonzero $a \in \mathbb{Z}$ such that*

$$af = qd + r \quad \textit{and} \quad \textit{either } r = 0 \textit{ or } \deg(r) < \deg(d) \; .$$

If the leading coefficient of d is 1, then we can also choose q and r such that $a = 1$.

Proof. We leave the theorem's first claim to the exercises; here we show merely that if $\operatorname{lc}(d) = 1$, then we can choose q and r such that $a = 1$, take a moment to review Algorithm 2.1. With $\operatorname{lc}(d) = 1$, step 2(a) always chooses $t = {}^{\operatorname{lt}(r)}\!/_{\operatorname{lt}(d)}$ where the denominator has a coefficient of 1. If the numerator's coefficient is an integer, then t's coefficient will also be an integer. In fact, r's initial coefficients are all integers because they are f's, and $f \in \mathbb{Z}[x]$, so step 2(a) always results in a term whose coefficient is an integer. The end result is that the coefficients of q and r are always integers, and by the proof of Theorem 2.18, $1 \times f = qd + r$. □

Having a leading coefficient of 1 can be useful! We such polynomials ***monic***.

Unlike the Division Theorems for $\mathbb{Z}$ or for $\mathbb{Q}[x]$, (Theorems 1.30 and 2.18), we don't necessarily have a unique remainder. Once we have $af = qd + r$, we can multiply by any integer we like, and obtain another expression of that sort, as well. For instance, if we revisit the example above, we had

$$4f = \hat{q}d + \hat{r} \; ,$$

but we could also have

$$8f = (2\hat{q})\,d + (2\hat{r}) \quad \text{or} \quad -36f = (-9\hat{q})\,d + (-9\hat{r}) \; .$$

Exercises

Exercise 2.22. Let $f = x^4 + x + 1$ by $d = 2x + 3$.

(a) Divide f by d as elements of $\mathbb{Q}[x]$, obtaining Theorem 2.18's $q, r \in \mathbb{Q}[x]$.

(b) Divide f by d as elements of $\mathbb{Z}[x]$, obtaining Corollary 2.21's $a \in \mathbb{Z}$ and $q, r \in \mathbb{Z}[x]$.
Hint: All you have to do is modify the result of part (a).

(c) Is it possible to divide f by d and obtain $a = 1$ for Corollary 2.21?

Exercise 2.23. Complete the proof of Corollary 2.21; that is, for any $f, d \in \mathbb{Z}[x]$ we can find $a \in \mathbb{Z}$ and $q, r \in \mathbb{Z}[x]$ such that $af = qd + r$ and either $r = 0$ or $\deg(r) \leq \deg(d)$.
Hint: You know that Theorem 2.18 is true, so start with its conclusion. Then apply the same technique that we used in Example 2.20 and Exercise 2.22(b).

Exercise 2.24. Let $f \in \mathbb{Q}[x]$ and let s be any number. Let r be the remainder of dividing f by $x - s$. Explain why $r = 0$ or $\deg(r) = 0$.

Exercise 2.25. The *Factor Theorem* states that if $f \in \mathbb{Q}[x]$ and $s \in \mathbb{Q}$ is a root of f, then $x - s$ is a factor of f. Prove this theorem.
Hint: Use the Division Theorem (2.18) to write $f = qd + r$. If $r = 0$, then you are done, so suppose $r \neq 0$. Keep in mind the divisor $d = x - s$ and the result of Exercise 2.24. Substitute $x = s$ into both sides of the equation. What must r be?

Exercise 2.26. The *Remainder Theorem* states that for any polynomial $f \in \mathbb{Q}[x]$ and any $s \in \mathbb{Q}$, the remainder of dividing f by $x - s$ is $f(s)$. Prove this theorem.
Hint: Use the Division Theorem to write $f = qd + r$. Keep in mind the divisor $d = x - s$ and the result of Exercise 2.24. What must r be?

Sage supplement

Dividing polynomials in Sage works like dividing integers in Sage: use the `.quo_rem` method. However, we cannot use it unless we first tell Sage what kind of coefficients we want.

Consider $f = x^5 + 2x^2 + 1$ and $g = 2x^3 + x$, from Example 2.19. If we try to perform division immediately, we encounter an error.

```
sage: f, d = x^5 + 2*x^2 + 1, 2*x^3 + x
sage: f.quo_rem(d)
Traceback (click to the left of this block for traceback)
...
AttributeError:  'sage.symbolic.expression.Expression'
object has no attribute 'quo_rem'
```

To avoid this error, we need to tell Sage which set contains the coefficients of`f` and `d`. You might look at them and think, "Isn't it obvious that their coefficients are integers, so that the polynomials are elements of $\mathbb{Z}[x]$?" Let's go ahead and tell Sage this, and see what happens.

```
sage: f, d = ZZ[x](f), ZZ[x](d)
sage: f.quo_rem(d)
(0, x^5 + 2*x^2 + 1)
```

What are we telling Sage in the first line? Recall that $\mathbb{Z}[x]$ is the set of polynomials with integer coefficients. Sage considers `ZZ[x]` to be the set of polynomials with integer coefficients, and the first line ***coerces*** `f` and `d` so that they are seen as having integer coefficients: `ZZ[x](f)` produces a new polynomial from `f`, only now it is guaranteed to have integer coefficients.

Then, when we ask Sage to divide `f` by `d`, it first checks whether `f`'s leading term is divisible by `d`'s. *The leading term includes the coefficient*, so that even though $x^3 \mid x^5$, we also have $2 \nmid 1$. On this account, Sage will not divide `f` and `d` as *integer* polynomials. The quotient is 0, and the remainder is $x^5 + 2x^2 + 1$.

This is not the answer we found in Example 2.19, but it should actually make sense. We know from Example 2.19 that the quotient and remainder should have rational coefficients, so if we ask Sage to consider f and d as integer polynomials, *it should not* return a quotient and remainder with rational coefficients.

Let's try telling Sage that `f` and `d` have rational coefficients.

```
sage: f, d = QQ[x](f), QQ[x](d)
sage: f.quo_rem(d)
(1/2*x^2 - 1/4, 2*x^2 + 1/4*x + 1)
```

We ended up with the quotient and remainder from Example 2.19, as desired.

You might wonder if we can "fix" the division in $\mathbb{Z}[x]$, the way we did in Example 2.20. Indeed we can: use the `.pseudo_divrem` method.

```
sage: f, d = QQ[x](f), QQ[x](d)
sage: f.pseudo_divrem(d)
(2*x^2 - 1, 8*x^2 + x + 4, 2)
```

Notice that there are three answers. The first and second are precisely the quotient and remainder we obtained in Example 2.20. Let's call the third answer e (for exponent). If we write c as the leading coefficient of q, then the relationship between a, c, e, f, q, and r is,

$$c^e f = dq + r .$$

In short, c^e here stands for a in Corollary 2.21. We did not prove in the corollary that a would have the form c^e, but it does! We leave the proof of this fact as an exercise to the student.

Exercises

Exercise 2.27. Use Sage to verify your answers to Exercise 2.22.

Exercise 2.28. Consider again the polynomials of Example 90. Use Sage to help you perform the division "by hand." Look closely at when the quotient and remainder acquire rational coefficients, and at each step the denominators change. Use your observations to explain why, as we point out at the end of this section, "a (from Corollary 2.21) has the form c^e."

Exercise 2.29. Sage already has a polynomial division algorithm, but it's a good exercise to write your own. So, implement Algorithm 2.1 as a Sage procedure. You will probably need some of the Sage commands listed on page 91, so be sure to have those handy.

2.3 Polynomial divisors

Now that we can divide polynomials, we naturally wonder if the concepts of *divisibility* and *common divisors* will apply, and there really is no reason to think they won't. We say that the polynomial d *divides* the polynomial f if the result of Theorem 2.18 has $r = 0$, and write $d \mid f$ for short. We say that d is a *common divisor* of polynomials f and g if $d \mid f$ and $d \mid g$.

Common divisors

If $f, g \in \mathbb{Q}[x]$ then we say that d is a ***greatest common divisor*** of f and g when

- d is a common divisor of f and g, and additionally
- for any other common divisor c of f and g, $\deg(c) \leq \deg(d)$.

Example 2.30. The polynomial $x - 1$ is a common divisor of $x^4 - 1$ and $x^6 - 1$, but a greatest common divisor is $x^2 - 1$.

Unlike integers, we do not refer to "the" greatest common divisor, because polynomials can have more than one.

Example 2.31. Another greatest common divisor of $x^4 - 1$ and $x^6 - 1$ is $2x^2 - 2$:

$$x^4 - 1 = \left(2x^2 - 2\right)\left(\frac{x^2}{2} + \frac{1}{2}\right) \quad \text{and} \quad x^6 - 1 = \left(2x^2 - 2\right)\left(\frac{x^4}{2} + \frac{x^2}{2} + \frac{1}{2}\right).$$

After all, we said that $f, g \in \mathbb{Q}[x]$, so we are dealing with polynomials over the rationals, not over the integers.

We prefer greatest common divisors whose leading coefficient is 1; that is, we will always aim for a monic greatest common divisor. Eventually we will prove that only one monic polynomial is a greatest common divisor, but for now we do not even know that.

A common divisor of two polynomials enjoys a special relationship with their roots.

Theorem 2.32. *Suppose that $d, f, g \in \mathbb{Q}[x]$, with d a common divisor of f and g. If d has a root in $\mathbb{Q}$, then both f and g have the same root in $\mathbb{Q}$. Conversely, if both f and g have a common root in $s \in \mathbb{Q}$, then any greatest common divisor of f has s as a root.*

Example 2.33. Recall the polynomials of Example 2.30: $f = x^4 - 1$, $g = x^6 - 1$, a common divisor $c = x - 1$, and a greatest common divisor $d = x^2 - 1$.

It is easy to verify that not only $c(1) = 0$, but also $f(1) = g(1) = 0$, as promised by the theorem.

On the other hand, $s = -1$ is also a common root of f and g, but it is not a root of c, as $c(-1) = -2 \neq 0$. This does not trouble us, since c is not a greatest common divisor. However, d is a greatest common divisor, and in fact $d(-1) = 0$.

Proof of Theorem 2.32. Let S be any set, and let $s \in S$.

First suppose that s is a root of d. By definition, $d(s) = 0$. Choose polynomials p and q such that $dp = f$ and $dq = g$. By substitution, $f(s) = (dp)(s) = d(s)p(s) = 0 \times p(s) = 0$, and similarly $g(s) = 0$. So a root of d is always a root of both f and g.

Conversely, suppose s is a root of both f and g, and d is a greatest common divisor of f and g. Using the same definitions of p and q, we know that

$$f(s) = g(s) = 0 , \quad \text{so that} \quad d(s)p(s) = d(s)q(s) = 0 .$$

By Theorem 2.6, the zero product property holds for polynomials over the rationals, so $d(s) = 0$ or both $p(s) = q(s) = 0$. If $d(s) = 0$, then we have arrived at the desired conclusion, so by way of contradiction suppose instead that $d(s) \neq 0$. We just said that this forces $p(s) = q(s) = 0$. By the Factor Theorem (Exercise 2.25), $x - s$ divides both p and q, which means $(x - s) \times d$ divides both f and g, contradicting the choice of d as a greatest common divisor of f and g. □

Greatest common divisors

We turn our attention to two questions:

- Can we adapt the Euclidean Algorithm (Algorithm 1.2 on page 28) with polynomials to compute a greatest common divisor of two polynomials?
- Can we adapt the Extended Euclidean Algorithm (Algorithm 1.3 on page 32) with polynomials to compute coefficients s and t such that $sf + tg = \gcd(f, g)$?

The answer to both is “yes!” …as long as we make certain necessary, but natural (even “obvious”) adjustments to the algorithms. A modified Euclidean algorithm appears in Algorithm 2.2 on the following page. Observe that it is almost identical to Algorithm 1.3 on page 32.

Algorithm 2.2 The Euclidean Algorithm for polynomials

```
Input
  • $f, g \in \mathbb{Q}[x] \setminus \{0\}$
Output
  • a greatest common divisor of $f$ and $g$
Do
  1. let $h$ be the larger of $f$ and $g$ with respect to degree, and let $k$ be the smaller
  2. while $k \neq 0$
     (a) determine $q, r$ that satisfy the Division Theorem
     (b) replace $h$ by $k$, then replace $k$ by $r$
  3. return $h$
```

Theorem 2.34. *Algorithm 2.2 terminates correctly.*

The proof imitates that of the original Euclidean Algorithm (Theorem 1.48). If you understand the former, and if you understand polynomial arithmetic, you should this proof, as well.

Proof. Enumerate each h and k in steps 1 and 2 as $h_0, h_1, \ldots$ and $k_0, k_1, \ldots$.

Termination? By construction, $\deg(k_0) < \deg(h_0)$. For each $i = 1, 2, \ldots$ k_i is the remainder of dividing h_{i-1} by k_{i-1}, so by the Division Theorem either $k_i = 0$ or $\deg(k_i) < \deg(k_{i-1})$. We have a sequence

$$\deg(k_0) > \deg(k_1) > \cdots$$

of nonincreasing natural numbers. By Theorem 1.28, it must stabilize. It cannot stabilize as long as step 2 continues, so eventually it must end, at which point the algorithm passes to step 3 and returns h.

Correctness? We claim that for each i, if the polynomial p divides h_i and k_i, then it also divides h_{i+1} and k_{i+1}. To see why, assume that p divides h_i and k_i, then recall that in step 2 we compute $h_i = q_i k_i + r_i$ and then set $h_{i+1} = k_i$ and $k_{i+1} = r_{i+1}$. By substitution, p divides h_{i+1}. On the other hand, rewrite the division equation so that

$$r_i = h_i - q_i k_i \ .$$

We readily see that p divides the right hand side; thus it must divide the left. Since p divides r_i, by substitution it divides k_{i+1}.

What does this mean? The last pair considered consists of $h_{\text{last}} = r$ and 0, so any common divisor of $h_0, k_0 \in \{f, g\}$ is a divisor of r. This would include any greatest common divisor, so the degree of r must be at least as large as that of a greatest common divisor. On the other hand, moving backwards through the algorithm would show that $k_{\text{last}-1} = r$, and $h_{\text{last}-1} = q_{\text{last}-1} k_{\text{last}-1} + h_{\text{last}} = (q_{\text{last}-1} + 1)\, r$, so r divides $h_{\text{last}-1} = k_{\text{last}-2}$. Continuing backwards through the divisions, we would find that r divides both h_0 and k_0, so that r is a common divisor of f and g, which means the degree of r must be no larger than the degree of a greatest common divisor. Hence $\deg(r)$ *is* the degree of a greatest common divisor, so r is a greatest common divisor of f and g. □

It is not a long step from a Euclidean Algorithm to an Extended Euclidean Algorithm.

Example 2.35. Let $f = x^6 - 1$ and $g = x^4 - 1$. Perform the Euclidean Algorithm (Algorithm 2.2):

- In step 1, let $h = x^6 - 1$ and $k = x^4 - 1$.
- Since $k \neq 0$, we perform step 2:
  - In step 2(a), $q = x^2$ and $r = x^2 - 1$ satisfy the Division Theorem.
  - In step 2(b), replace h by $x^4 - 1$ and k by $x^2 - 1$.
- Since $k \neq 0$, we perform step 2 again:
  - In step 2(a), $q = x^2 + 1$ and $r = 0$ satisfy the Division Theorem.
  - In step 2(b), replace h by $x^2 - 1$ and k by 0.
- Now $k = 0$, so we proceed to step 3, which returns $h = x^2 - 1$. This is in fact $\gcd\left(x^6 - 1, x^4 - 1\right)$.

As with the integers, we can work our way backwards through the divisions to obtain Bézout coefficients. Start with

$$x^6 - 1 = x^2\left(x^4 - 1\right) + \left(x^2 - 1\right) .$$

Isolate the remainder to obtain

$$1 \times \left(x^6 - 1\right) + \left(-x^2\right) \times \left(x^4 - 1\right) = x^2 - 1 .$$

This shows that if we set $s = 1$ and $t = -x^2$, then $sf + tg = \gcd(f, g)$.

Theorem 2.36 (Extended Euclidean Algorithm in $\mathbb{Q}[x]$)**.** *For any $f, g \in \mathbb{Q}[x]$, there exist $s, t \in \mathbb{Q}[x]$ such that $sf + tg = \gcd(f, g)$.*

Again, the proof imitates that for the integers.

Proof. Enumerate the various divisions performed during the Euclidean Algorithm as

$$h_i = q_i k_i + r_i$$

where $i = 1, 2, \ldots, \ell$ and r_ℓ is the last non-zero remainder. Write $d = h_\ell$; we know from Theorem 2.34 that d is a greatest common divisor of f and g. Rewrite the last division as

$$d = h_\ell - q_\ell k_\ell .$$

Recall that $h_\ell = k_{\ell-1}$ and $k_\ell = r_{\ell-1}$, so rewrite this equation as

$$d = k_{\ell-1} - q_\ell r_{\ell-1} . \tag{2.2}$$

Rewrite the previous division as

$$r_{\ell-1} = h_{\ell-1} - q_{\ell-1}k_{\ell-1} \ .$$

Substitute into equation (2.2) to obtain

$$d = k_{\ell-1} - q_{\ell}\left(h_{\ell-1} - q_{\ell-1}k_{\ell-1}\right) = \left(1 + q_{\ell}q_{\ell-1}\right)k_{\ell-1} + \left(-q_{\ell}\right)h_{\ell-1} \ .$$

By repeating this process, we eventually obtain an expression

$$d = Q_0 h_0 + Q_1 k_0 \ ,$$

which by substitution becomes

$$d = Q_0 f + Q_1 g \ .$$

By Theorem 2.4, the values $s = Q_0$ and $t = Q_1$ are polynomials over the rationals, so they satisfy the theorem. □

Exercises

Exercise 2.37. Let $f = x^4 - 4$ and $g = x^5 + 3x^4 - x^3 - 3x^2 - 2x - 6$.

(a) Find $\gcd(f, g)$ and use it to compute the common roots of f and g.

(b) Find the Bézout coefficients $s, t \in \mathbb{Q}[x]$ such that $sf + tg = \gcd(f, g)$.

Sage supplement

We already mentioned the `.factor` method on page 87. Computing greatest common divisors for polynomials is identical to integers: use the `gcd` command if you simply want the greatest common divisor, and the `xgcd` command if you want the Bézout coefficients. We remarked earlier that two polynomials can have more than one greatest common divisor; of these, Sage returns the monic common divisor when that make sense (e.g., rational coefficients). Let's consider Example 96. We have to coerce `f` and `g`into $\mathbb{Z}[x]$ before trying `xgcd`; otherwise, Sage will report an error.

```
sage: f, g = x^4 - 1, x^6 - 1
sage: gcd(f, g)
x^2 - 1
sage: f, g = ZZ[x](f), ZZ[x](g)
sage: xgcd(f, g)
(x^2 - 1, -x^2, 1)
```

This means that $\gcd\left(x^4 - 1, x^6 - 1\right) = x^2 - 1$, *and* that $-x^2\left(x^4 - 1\right) + 1 \times \left(x^6 - 1\right) = x^2 - 1$. This agrees with the results we computed in Example 2.35.

Let's experiment a little with the coefficients. We'll multiply `f` and `g` by 2 and 4, respectively, then compute their gcd in $\mathbb{Z}[x]$.

```
sage: f, g = ZZ[x](2*x^4 - 2), ZZ[x](4*x^6 - 4)
sage: gcd(f, g)
2*x^2 - 2
```

Now we'll compute their gcd in $\mathbb{Q}[x]$.

```
sage: f, g = QQ[x](f), QQ[x](g)
sage: gcd(f, g)
x^2 - 1
```

The greatest common divisor changed. The reason for this is that 2's multiplicative inverse, $1/2$, is clearly not an integer, so in $\mathbb{Z}[x]$ we consider the factorization by 2 to be important. On the other hand, $1/2$ clearly *is* a rational number, so in $\mathbb{Q}[x]$ we consider the factorization by 2 to be *un*important. We consider this distinction more carefully in Section 2.4 when we talk about *units*.

Exercises

Exercise 2.38. Use Sage to verify your answers to Exercise 2.37.

Exercise 2.39. Sage already has a algorithm to compute greatest common divisors, but it's a good idea to implement your own. Implement Algorithm 2.2 in Sage, and make sure you get the correct results by comparing it to Sage's for several pairs of random polynomials. For extra credit, implement an Extended Euclidean Algorithm, so that you get the Bézout coefficients, as well.

2.4 Factoring polynomials

When we studied integer division, we touched on factorization into primes. In the same way, we can discuss factorization of polynomials. With polynomials, factorization is useful because it helps us find roots.

Example 2.40. Suppose $f = x^4 - 5x^2 + 4 \in \mathbb{Z}[x]$. This factors as $f = (x+2)(x+1)(x-1)(x-2)$. Hence, $f = 0$ when

$$(x+2)(x+1)(x-1)(x-2) = 0\ .$$

By the zero product property, one of $x+2=0$, $x+1=0$, $x-1=0$, or $x-2=0$. Hence the roots are $\{\pm 2, \pm 1\}$.

Just as we excluded 1 from the realm of prime numbers, we need to exclude certain "polynomials" from the building blocks of polynomials.

Example 2.41. Although $x^2+1 = 2\times\left[\left(x^2/2\right) + \left(1/2\right)\right]$, this does not seem like a proper factorization, first of all because it's "too easy;" second, it doesn't reduce the degree. It seems reasonable to disallow this sort of thing.

This "naughty" factorization was possible merely because 2 has a multiplicative inverse in $\mathbb{Q}$. It's a bit of a mouthful always to say "a number that has a multiplicative inverse in a set S;" instead we that that a number is a ***unit*** in S. Whether a number is a unit depends very much on the underlying set; 2 is a unit in $\mathbb{Q}$, but not in $\mathbb{Z}$. Similarly we say that f and g are ***associates*** if $f = ag$ where a is a unit.

Remark. You will show in the exercises that the only units in $\mathbb{Q}[x]$ are the nonzero rationals themselves.

Suppose f is a polynomial, but not a unit. We say that f ***factors*** *over a set* S if there exist polynomials p, q whose coefficients are elements of S such that$f = pq$ and neither p nor q is an associate of f. We call p and q ***factors*** of f. If f does not factor over S, then we say that f is ***irreducible***.

Whether a polynomial factors depends very much on the choice of S.

Example 2.42. Suppose $f = x^2 - 2$. If we consider $f \in \mathbb{Z}[x]$, then it will not factor; there are no polynomials $p, q \in \mathbb{Z}[x]$ such that $pq = f$. This remains true if we consider $f \in \mathbb{Q}[x]$. Hence, $x^2 - 2$ is irreducible over $\mathbb{Z}$ and over $\mathbb{Q}$.

If we consider $f \in \mathbb{R}[x]$, then f factors as $\left(x - \sqrt{2}\right)\left(x + \sqrt{2}\right)$. Hence, f factors over $\mathbb{R}$.

The point of excluding associates is to make sure factorization really reduces a polynomial.

Lemma 2.43. *If f factors over $\mathbb{Q}$ as $f = pq$, then $0 < \deg(p), \deg(q) < \deg(f)$.*

Proof. Suppose by way of contradiction that $f = pq$, f and p are not associates, and $\deg(p) = 0$ or $\deg(p) = \deg(f)$. If $\deg(p) = 0$, then $p \in \mathbb{Q}\backslash\{0\}$ itself, in which case p is a unit, which makes q an associate of f, contradicting the definition of a factorization. On the other hand, if $\deg(p) = \deg(f)$, then by Theorem 2.4, $\deg(q) = \deg(f) - \deg(p) = 0$, so q is a unit, which makes p an associate, again contradicting the hypothesis. Hence $0 < \deg(p) < \deg(f)$. A similar argument shows that this is true of $\deg(q)$, as well. □

Lemma 2.43 makes a statement about factorization over $\mathbb{Q}$ that is *not* generally true over $\mathbb{Z}$. This distinction is *very subtle* but *very important.*

Example 2.44. Let $f = 2x^2+2$. This factors over $\mathbb{Z}$ as $f = 2\left(x^2+1\right)$. In addition, the factorization is *into irreducibles.* Yet $\deg(f) = \deg\left(x^2+1\right)$ and $\deg(2) = 0$.

On the other hand, we would not say that f factors over $\mathbb{Q}$ as $f = 2\left(x^2+1\right)$, because 2 is a unit, so x^2+1 is an associate of f in $\mathbb{Q}[x]$.

It didn't matter in Example 2.42 whether we considered $f \in \mathbb{Z}[x]$ or $f \in \mathbb{Q}[x]$, as f factored over neither $\mathbb{Z}$ nor $\mathbb{Q}$.

Recall from Section 1.4 that we sometimes call prime numbers irreducible. It is natural to ask if irreducible polynomials behave the same way as prime numbers; that is,

- a polynomial factors into irreducibles; or,
- if an irreducible divides a product, then it divides one of the factors ($f \mid gh$ implies $f \mid g$ or $f \mid h$).

Interestingly enough, this is the case.

Theorem 2.45 (Euclid's Lemma for polynomials). *Let $f, g, h \in \mathbb{Q}[x]$, and suppose f is irreducible over $\mathbb{Q}$. If $f \mid gh$, then $f \mid g$ or $f \mid h$.*

Proof. This proof will again resemble the one for integers, Theorem 1.66. Assume that $f \mid gh$. If $f \mid g$, then the theorem is satisfied, so assume that $f \nmid g$. Let d be a greatest common divisor of f and g. By the irreducibility of f, $\deg(d) = 0$ or $\deg(d) = \deg(f)$. If $\deg(d) = \deg(f)$, then $f = kd$ for some $k \in \mathbb{Q}$, which implies that $k^{-1}f \mid g$, and hence that $f \mid g$, a contradiction. So $\deg(d) = 0$; that is, d is constant. By the Extended Euclidean Algorithm, there exist $s, t \in \mathbb{Q}[x]$ such that

$$sf + tg = d\ .$$

Multiply both sides by d^{-1} — a constant! — to see that

$$\left(d^{-1}s\right)f + \left(d^{-1}t\right)g = 1\ .$$

Now multiply both sides by h to obtain

$$\left(d^{-1}sh\right)f + \left(d^{-1}t\right)(gh) = h\ .$$

Recall the hypothesis that $f \mid gh$; choose $q \in \mathbb{Q}[x]$ such that $qf = gh$. Substitute into the equation above to obtain

$$\left(d^{-1}sh\right)f + \left(d^{-1}t\right)(qf) = h\ .$$

We can factor the common f to rewrite this equation as

$$f\left(d^{-1}sh + d^{-1}qt\right) = h\ .$$

In other words, $f \mid h$. □

Now we work our way into an analogue of the Fundamental Theorem of Arithmetic. We start by showing that every polynomial over the rationals has an irreducible factor.

Lemma 2.46. *Every $f \in \mathbb{Q}[x]$ has an irreducible factor.*

Proof of Lemma 2.46. Consider the following algorithm.

```
1. let $g_0 = f$, and $i = 0$
2. while $g_i$ is not irreducible
    (a) choose $g_{i+1}, h_{i+1} \in \mathbb{Q}[x]$ such that $g_i = g_{i+1}h_{i+1}$ and $0 < \deg(g_{i+1}), \deg(h_{i+1}) < \deg(g_i)$
    (b) increment $i$ by 1
3. return $g_i$
```

We claim that this algorithm terminates with an irreducible factor of f. If f is itself irreducible, then the algorithm skips step 2 and returns $g_0 = f$, as desired. Otherwise, it enters step 2 and generates a sequence of polynomials $g_0, g_1, \ldots$ such that $\deg(g_0) > \deg(g_1) > \cdots$. This is a nonincreasing sequence of natural numbers, and by Theorem 1.28 the sequence must stabilize. It can only stabilize if the loop ends, and the loop ends only when g_i is irreducible, as desired. □

Now we show that polynomials over the rationals factor uniquely into irreducibles.

Theorem 2.47. *Let $f \in \mathbb{Q}[x]$. There exist irreducible polynomials $p_1, \ldots, p_k \in \mathbb{Q}[x]$ such that $f = p_1 \cdots p_k$. Moreover, this factorization is unique in the sense that if $f = q_1 \cdots q_\ell$ is another factorization of f into irreducibles, then $k = \ell$ and we can reorder the q's such that p_i is an associate of q_i for each i.*

Proof. First we consider the *existence* of a factorization into irreducibles. Consider the following algorithm, where step 2 is justified by Lemma 2.46.

```
1. let $i = 0$ and $r_0 = f$
2. while $r_i$ is not irreducible
    (a) let $p_i \in \mathbb{Q}[x]$ be an irreducible factor of $r_i$
    (b) choose $r_{i+1} \in \mathbb{Q}[x]$ such that $p_i r_{i+1} = r_i$
    (c) increment $i$ by 1
3. let $p_i = r_i$
4. return $p_0, p_1, \ldots, p_i$
```

We claim that this algorithm terminates with an irreducible factorization of f. How so? If f is itself irreducible, then the algorithm skips step 2, assigns $p_0 = r_0 = f$, and returns p_0, as desired. Otherwise, it enters step 2 and generates a sequence of polynomials $r_1, r_2, \ldots$ and *irreducible* polynomials $p_1, p_2, \ldots$ such that

$$\deg(r_0) = \deg(r_1) + \deg(p_1) > \deg(r_1) = \deg(r_2) + \deg(p_2) > \deg(r_2) = \cdots,$$

or,

$$\deg(r_0) > \deg(r_1) > \cdots.$$

This is a nonincreasing sequence of natural numbers; by Theorem 2.46 it must stabilize eventually, which it can do only if the algorithm breaks out of step 2. At that point it assigns $p_i = r_i$. By substitution,

$$f = r_0 = p_1 r_1 = p_1(p_2 r_2) = (p_1 p_2)(p_3 r_3) = \cdots = p_1 \cdots p_i.$$

We have a factorization of f into irreducible polynomials, as desired.

We still have to show that the factorization is unique, in the sense described in the theorem. To see this, suppose that $f = p_1 \cdots p_k$ and $f = q_1 \cdots q_\ell$, where each p_i and each q_j is irreducible over $\mathbb{Q}$. By substitution and the associative property,

$$p_1(p_2 \cdots p_k) = q_1 \cdots q_\ell. \tag{2.3}$$

By definition of divisibility, $p_1 \mid q_1 \cdots q_\ell$. By Theorem 2.45, $p_1 \mid q_i$ for some $i = 1, \ldots \ell$. Recall that q_i is also irreducible, so $p_1 \mid q_i$ only if they are associates. Reorder the q's so that $p_1 = a_1 q_1$ for some $a_1 \in \mathbb{Q}$. By substitution into equation (2.3),

$$p_1(p_2 \cdots p_k) = (a_1 p_1)(q_2 \cdots q_k), \quad \text{so that} \quad p_2 \cdots p_k = a_1 q_2 \cdots q_k.$$

Continuing in this fashion, we can show that q_2 is an associate of p_2, and so forth. □

Exercises

Exercise 2.48.

(a) Explain why the only units of $\mathbb{Z}$ are ±1.

(b) Unlike $\mathbb{Z}$, only one element of $\mathbb{Q}$ is not a unit. Which is it? How do you know that the other elements are units?

(c) The units of $\mathbb{Z}[x]$ are also only ±, while the units of $\mathbb{Q}[x]$ are, in a similar way, precisely the units of $\mathbb{Q}$. That is, polynomials of degree 1 or higher are not units in these settings. Explain why this is the case.

Exercise 2.49. This problem considers linear polynomials, which have the form $f = ax + b$.

(a) Explain why a linear polynomial is always irreducible over $\mathbb{Q}$.

(b) A linear polynomial is *not* always irreducible over $\mathbb{Z}$. Give an example of a linear polynomial over $\mathbb{Z}$ that factors, and of a linear polynomial over $\mathbb{Z}$ that does not factor.

Exercise 2.50. Suppose that $f \in \mathbb{Q}[x]$, that its degree is m, it has m distinct roots, and all of its roots are rational. Show that f factors into exactly m linear polynomials.
Hint: Use the Factor Theorem (Exercise 2.25).

Sage supplement

The main thing to emphasize in this section is that the way Sage factors a polynomial depends very much on the coefficients' set, just as the way Sage computes a greatest common divisor.

We'll start with an easy example that builds on our observations from the previous section: we'll try factoring $2x^2 + 2$, from Example 2.44. First factor it generically, then over $\mathbb{Z}[x]$, and finally over $\mathbb{Q}[x]$.

```
sage: (2*x^2 + 2).factor()
2*x^2 + 2
sage: ZZ[x](2*x^2 + 2).factor()
2 * (x^2 + 1)
sage: QQ[x](2*x^2 + 2).factor()
(2) * (x^2 + 1)
```

The answer came out slightly different each time! The first attempt didn't factor at all; again, this is due to Sage's being careful; it doesn't quite know what you want, so it does nothing. The second attempt factored the way you might expect, given that 2 is in fact the greatest common factor of $2x^2$ and of 2. The third case also factors this way, but placed the 2 in parentheses. This signals that Sage is aware that 2 is a *unit* in $\mathbb{Q}[x]$.

Now consider $x^2 - 2$, from Example 2.42.

```
sage: (x^2 - 2).factor()
x^2 - 2
sage: ZZ[x](x^2 - 2).factor()
x^2 - 2
sage: QQ[x](x^2 - 2).factor()
x^2 - 2
sage: RR[x](x^2 - 2).factor()
(x - 1.41421356237310) * (x + 1.41421356237310)
```

The polynomial $x^2 - 2$ does not factor when we consider its coefficients to be symbolic (the first attempt), integer (the second), or rational (the third). All this is fine. When we consider its coefficients to be real numbers, it does factor; unfortunately, the factorization is not quite correct, for

$$\sqrt{2} \neq 1.41421356237310 \, .$$

In Sage, `RR[x]` represents the set of polynomials whose coefficients are *approximations* of real numbers.

You might wonder if Sage can factor this polynomial exactly. Indeed it is. One is to use the "algebraic numbers" as coefficients; to do this, coerce the polynomial into `AA[x]`, and factor it that way.

```
sage: AA[x](x^2 - 2).factor()
(x - 1.414213562373095?)  * (x + 1.414213562373095?)
```

At first glance, this factorization may look identical to the one before, but it isn't: notice that each of the two constants ends with a question mark. This is Sage's way of indicating that the numbers in question are algebraic numbers, and that Sage can in fact find them. To do so, compute the roots of $x^2 - 2$ in the algebraic numbers:

```
sage: R = AA[x](x^2 - 2).roots()
sage: R
[(-1.414213562373095?, 1), (1.414213562373095?, 1)]
```

We see two roots, each of multiplicity 1. The second root is positive, so let's extract it. We can ask Sage to convert it to a radical expression using an aptly-named command.

```
sage: rpos = R[1][0]
sage: rpos
1.414213562373095?
sage: rpos.radical_expression()
sqrt(2)
```

Sage is telling us that the positive root is $\sqrt{2}$.

To obtain the exact factorization, we'll extend the rationals by $\sqrt{2}$. We'll talk about this more in the next chapter, but the basic idea is that Sage can add $\sqrt{2}$ to $\mathbb{Q}$, creating a new set called $\mathbb{Q}\left[\sqrt{2}\right]$ whose arithmetic remains valid.

```
sage: QQ2 = QQ[sqrt(2)]
sage: QQ2[x](x^2 - 2).factor()
(x - sqrt2) * (x + sqrt2)
```

This is Sage's way of telling you that it has factored $x^2 - 2$ as $\left(x - \sqrt{2}\right)\left(x + \sqrt{2}\right)$.

Exercises

Exercise 2.51. Use Sage to factor $x^4 - 8x^2 + 15$ exactly. Its irreducible factors are all linear, so you need to identify the roots as precisely as possible.

Exercise 2.52. Use Sage to factor $x^3 - 3$ as much as possible using the techniques we have described in this section. What symbol does Sage use to indicate the root in your extension of $\mathbb{Q}$? Why do you think it isn't possible at the moment to factor $x^3 - 3$ into linear factors?

2.5 Imagining something quite real

This section shows how modular arithmetic can help us construct roots of polynomials.

Congruence modulo a polynomial

Let $f, g, d \in \mathbb{Z}[x]$. Throughout this section, $d \neq 0$. In a manner similar to that of Section 1.5, we say that $f \equiv g \pmod{d}$ if f and g have the same remainder after division by d. A familiar theorem applies in this case, too.

Theorem 2.53. *Let $f, g, d \in \mathbb{Q}[x]$, with $d \neq 0$. Then $f \equiv g \pmod{d}$ if and only if $d \mid (f - g)$.*

Proof. Recall that the phrase "if and only if" signals that the two phrases are equivalent, and thus we have to prove two directions.

Assume that f and g have the same remainder after division by d. By definition, there exist $q_f, q_g, r \in \mathbb{Q}[x]$ such that $f = q_f d + r$, $g = q_g d + r$, and either $r = 0$ or $\deg(r) < \deg(d)$. By substitution, $f - g = \left(q_f - q_g\right) d$. By definition, $d \mid (f - g)$.

Conversely, assume that $d \mid (f - g)$. By definition, there exists $q \in \mathbb{Q}[x]$ such that $dq = f - g$. Use the Division Theorem for $\mathbb{Q}[x]$ (Theorem 2.18) to choose $q_f, q_g, r_f, r_g \in \mathbb{Q}[x]$ such that $f = q_f d + r_f$, $g = q_g d + r_g$, and both r_f and r_g satisfy the requirements of a polynomial remainder after division by d. By substitution,

$$dq = \left(q_f d + r_f\right) - \left(q_g d + r_g\right) \quad\Longrightarrow\quad d\left(q - q_f + q_g\right) = r_f - r_g\,.$$

By definition, $d \mid (r_f - r_g)$. If $r_f \neq r_g$, then $\deg d \leq \deg(r_f - r_g)$. By Theorem 2.4, however, $\deg(r_f - r_g) \leq \max(\deg(r_f), \deg(r_g))$, and by the Division Theorem $\deg(r_f), \deg(r_g) < \deg(d)$. All together, we know that

$$\deg(d) \leq \deg(r_f - r_g) \leq \max(\deg(r_f), \deg(r_g)) < \deg(d) ;$$

in short, $\deg(d) < \deg(d)$, a contradiction. Our only assumption that might be unsound is $r_f \neq r_g$; that assumption must be wrong: $r_f = r_g$. In conclusion, $f \equiv g \pmod d$. □

Example 2.54. Consider congruence modulo $d = x^2 + 1$. Let $f = x + 3$, $g = 5x^4 + 2x^2 + x$, and $h = 5x^4 + 2x^2$. The remainder of dividing g by d is f, so $f \equiv g \pmod d$. On the other hand, $g - h = x$, and $d \nmid x$, so $g \not\equiv h \pmod d$.

Just as congruence modulo an integer was an equivalence relation, so is congruence modulo a polynomial.

Theorem 2.55. *Let $f \in \mathbb{Q}[x] \setminus \{0\}$. Congruence modulo f is an equivalence relation.*

Proof. We have to show three properties: reflexive, symmetric and transitive.

Reflexive: We leave this to Exercise 2.65(a).

Symmetric: We leave this to Exercise 2.65(b).

Transitive: Let $p, q, r \in \mathbb{Q}[x]$. We want to show that if $p \equiv q \pmod f$ and $q \equiv r \pmod f$, then $p \equiv r \pmod f$. Assume that $p \equiv q \pmod f$ and $q \equiv r \pmod f$. By definition, p and q have the same remainder after division by f; call that remainder g_1. Similarly, q and r have the same remainder after division by f; call that remainder g_2. The Division Theorem tells us that remainders after division are unique, so $g_1 = g_2$. Hence, p and r have the same remainder after division by f. By definition, then, $p \equiv r \pmod f$; or, congruence modulo a polynomial is transitive. □

Likewise, congruence cooperates with addition, subtraction, and multiplication.

Theorem 2.56. *Let $p, q, r \in \mathbb{Q}[x]$, and let $f \in \mathbb{Q}[x] \setminus \{0\}$. Suppose that $p \equiv q \pmod f$.*

(A) $p \pm r \equiv q \pm r \pmod f$.

(B) $pr \equiv qr \pmod f$.

Proof. By Theorem 2.53, $f \mid (p - q)$. We prove only (B), and leave a proof of (A) to Exercise 2.66.

We want to show that $pr \equiv qr \pmod f$. By Theorem 2.53, this is true if $f \mid (pr - qr)$, or $f \mid [(p - q) r]$. We already know that $f \mid (p - q)$; by definition, choose $g \in \mathbb{Q}[x]$ such that

$$fg = p - q .$$

By substitution and the associative property,

$$f(gr) = (p - q) r .$$

By definition, $f \mid [(p - q) r]$, so that $pr \equiv qr \pmod f$. □

Does division preserve congruence?

Congruence modulo a polynomial shares another property of congruence modulo an integer: division does not preserve congruence for *all* polynomial moduli, but it does preserve congruence for *certain special* polynomial moduli. Here we see that again how irreducible polynomials stand in for prime numbers (which are themselves irreducible).

Theorem 2.57. *Let* $f \in \mathbb{Q}[x]$, *with* $\deg(f) \geq 1$. *The following are equivalent.*

(A) f *is irreducible over* $\mathbb{Q}[x]$

(B) For every $p, q, r \in \mathbb{Q}[x]$ *such that* $f \nmid r$, $pr \equiv qr \pmod{f}$ *implies* $p \equiv q \pmod{f}$.

Proof. Recall that when we claim two phrases are equivalent, we have to prove two directions.

Assume that f is irreducible over $\mathbb{Q}[x]$. Let $p, q, r \in \mathbb{Q}[x]$ such that $f \nmid r$. Assume that $pr \equiv qr \pmod{f}$. By Theorem 2.53, $f \mid (pr - qr)$, or $f \mid [(p - q)r]$. By hypothesis, $f \nmid r$, so by Theorem 2.45, $f \mid (p - q)$. By Theorem 2.53, $p \equiv q \pmod{f}$.

For the converse, we prove its contrapositive. Assume that f is not irreducible. By definition, there exist nonzero $g, h \in \mathbb{Q}[x]$ such that f factors as $f = gh$ and $0 < \deg(g), \deg(h) < \deg(f)$. Rewrite $f = gh$ as $f \times 1 = gh$. By definition of divisibility, $f \mid gh$. Now, $0 < \deg(g) < \deg(f)$ implies $g \not\equiv 0 \equiv f$. Set $p = g$, $q = f$, $r = h$, and we have $pr \equiv qr \pmod{f}$ but $p \not\equiv q \pmod{f}$. □

Does f have to be irreducible for division modulo a polynomial to work? Yes.

Example 2.58. Suppose

$$\begin{aligned} f &= x^3 - 2x^2 - x + 2 = (x+1)(x-1)(x-2)\ , \\ p &= x^3 + 3x^2 - x - 3 = (x+1)(x-1)(x+3)\ , \\ q &= x^3 + x^2 - 10x + 8 = (x-1)(x-2)(x+4)\ , \\ \text{and } r &= x^2 - x - 2 = (x+1)(x-2)\ . \end{aligned}$$

It is not hard to verify that $pr \equiv qr \pmod{f}$, as the factorization makes it clear that both have remainder 0. In addition, $f \nmid r$. However, $p \not\equiv q \pmod{f}$. The cause is that f is not irreducible.

An imaginary number

We come to the main goal of this section — of this chapter, really — which is to introduce you to an example of what we will later call a "quotient ring." Keeping in line with our pattern of showing how polynomials behave in ways that are very similar to integers, we turn our attention to the last part of Section 1.5, where we defined the set $\mathbb{Z}_n$ and its arithmetic. Here we show that we can likewise obtain a set of polynomials with an arithmetic that has an amazing consequence.

For any polynomial $p \in \mathbb{Q}[y]$, define $\overline{p}$ to be the remainder of dividing p by $y^2 + 1$. (This is similar to the remainder operation defined on page 45.) By the Division Theorem for $\mathbb{Q}[y]$ (Theorem 2.18), $\deg(\overline{p}) < \deg(y^2 + 1) = 2$. By definition of degree, $\overline{p} = ay + b$, where $a, b \in \mathbb{Q}$.

Example 2.59. Let $f = y + 3$ and $g = 5y^4 + 2y^2 + y$, simlarly to Example 2.54. Since f is the remainder of dividing g by $y^2 + 1$, $f = \overline{g}$.

Consider the set

$$\mathbb{S} = \{ay + b : a, b \in \mathbb{Q}\} \ .$$

Notice that $\mathbb{S} \subseteq \mathbb{Q}[y]$. Define addition, subtraction, and multiplication of elements of $\mathbb{S}$ as follows: for any $ay + b, cy + d \in \mathbb{S}$,

$$\begin{aligned}(ay + b) \pm (cy + d) &= (a \pm c)\, y + (b \pm d)\\ (ay + b)(cy + d) &= \overline{(ay + b)(cy + d)} \ .\end{aligned}$$

We can simplify the product as

$$(ay + b)(cy + d) = (ad + bc)\, y + (bd - ac) \ , \tag{2.4}$$

since division by $y^2 + 1$ gives us

$$[(ay + b)(cy + d)] \quad = \quad \underbrace{(ac)}_{\text{quotient}} \cdot \left(y^2 + 1\right) \ + \ \underbrace{[(ad + bc)\, y + (bd - ac)]}_{\text{remainder}} \ .$$

Addition, subtraction, and multiplication in $\mathbb{S}$ are thus closed.

Example 2.60. Suppose $f = 2y + 3, g = 6y + 35$. Then

$$\begin{aligned}f + g &= 8y + 38\\ f - g &= -4y - 32\\ fg &= 88y + 93 \ .\end{aligned}$$

Be sure you understand how we arrived at the last one. Everything else in this section depends on understanding that! Here are the details:

$$fg = \overline{(2y + 3)(6y + 35)} = \overline{12y^2 + 88y + 105} = 88y + 93 \ .$$

We build $\mathbb{S}$ using polynomial congruence, and that leads to some interesting behavior.

Lemma 2.61. $\mathbb{Q} \subsetneq \mathbb{S}$.

Proof. Let $b \in \mathbb{Q}$. By definition, $b = 0y + b \in \mathbb{S}$. As $b \in \mathbb{Q}$ is arbitrary, $\mathbb{Q} \subseteq \mathbb{S}$. On the other hand, $y \notin \mathbb{Q}$, so $\mathbb{Q} \neq \mathbb{S}$, as claimed. □

Lemma 2.62. *In* $\mathbb{S}$, $y^2 = -1$ *and* $(-y)^2 = -1$.

Proof. By Equation (2.4),

$$\begin{aligned}y^2 &= (1 \times 0 + 0 \times 1)\, y + (0 \times 0 - 1 \times 1)\\ &= -1\end{aligned}$$

and

$$\begin{aligned}(-y)^2 &= [-1 \times 0 + 0 \times (-1)]\, y + [0 \times 0 - (-1) \times (-1)]\\ &= -1 \ .\end{aligned}$$

□

Let $\mathbb{S}[x]$ be the set of polynomials in x whose coefficients are elements of $\mathbb{S}$. We can perform addition, subtraction, and multiplication in $\mathbb{S}[x]$ using exactly the same techniques as in $\mathbb{Q}[x]$, with only one exception: when we multiply coefficients, we have to reduce them modulo y^2+1.

Example 2.63. Examples of elements of $\mathbb{S}[x]$ are

$$p = x^2 + 1\,, \quad q = (2y+3)\,x + (6y+35)\;, \quad r = yx^3 - (12y+1)\;.$$

The respective leading term of each polynomial is x^2, $(2y+3)\,x$, and yx^3. In addition,

$$\begin{aligned} qr &= \left(\overline{2y^2+3y}\right)x^4 + \left(\overline{6y^2+35y}\right)x^3 - \left(\overline{24y^2+38y+3}\right)x + \left(\overline{6y+35}\right)\\ &= \left(3y^2-2\right)x^4 + (35y-6)\,x^3 - (38y-21)\,x + (6y+35)\;. \end{aligned}$$

We finally arrive at our goal, which is actually easy to prove.

Corollary 2.64. *Both $y, -y \in \mathbb{S}$ are roots of $x^2+1 \in \mathbb{S}$.*

Proof. Let $f = x^2+1$. By Lemma 2.62, $f(y) = y^2+1 = (-1)+1 = 0$, and by definition y is a root of f. The proof for $-y$ is similar. □

What just happened? We have built a concrete set $\mathbb{S}$ that ***extends*** $\mathbb{Q}$; after all, Lemma 2.61 shows that $\mathbb{Q}$ is a subset of $\mathbb{S}$. In addition, by Theorem 2.56, the basic arithmetic of $\mathbb{S}$ is both identical to that of $\mathbb{Q}$ for elements of $\mathbb{Q}$, and also closed for elements of $\mathbb{S}\backslash\mathbb{Q}$. We could further show (as in Exercise 2.67) that $\mathbb{S}$ satisfies additional arithmetic properties we find desirable, so that it really does extend $\mathbb{Q}$ in every conceivable way. It also contains a root of $f = x^2+1$.

Let us empasize that *we constructed* $\mathbb{S}$ in a very *concrete* way. It is no less "real" than any rational number or polynomial. If you consider those things real — and there's no good reason not to — then so are the elements of $\mathbb{S}$.

But $y \in \mathbb{S}$ behaves just like the "imaginary" number i. In short, **the imaginary number is real**.[4]

Exercises

Exercise 2.65. Complete the proof of Theorem 2.55 by showing that congruence modulo a polynomial in $\mathbb{Q}[x]$ is (a) reflexive and (b) symmetric.

Exercise 2.66. Complete the proof of Theorem 2.56 by showing that if $p \equiv q \pmod f$, then $p \pm r \equiv q \pm r \pmod f$.

Exercise 2.67. Show that the following properties for addition and multiplication of elements of $\mathbb{S}$.

(a) the commutative property: both $(ay+b)+(cy+d) = (cy+d)+(ay+b)$ and $(ay+b)(cy+d) = (cy+d)(ay+b)$

[4] We don't mean that i is a real number, but rather that it is not "imaginary" in the sense of "not based in reality". The author of these notes learned in high school that i is imaginary and not real, and for a long time he found it very perplexing that we have to work with imaginary numbers. As we have seen, however, the name is unfortunate: i is as real as any other "real" number, insofar as we can "construct" it in a very concrete way.

(b) the associative property: both $[(ay+b)+(cy+d)]+(ey+f) = (ay+b)+[(cy+d)+(ey+f)]$ and $[(ay+b)(cy+d)](ey+f) = (ay+b)[(cy+d)(ey+f)]$

(c) $1 \in \mathbb{S}$, and any $s \in \mathbb{S}$ satisfies $1 \times s = s$ and $s \times 1 = s$
Hint: Be careful. You have to work in the form of elements of $\mathbb{S}$ for full correctness; that is, $ay+b$ and so forth.

(d) Any nonzero element of $\mathbb{S}$ has a multiplicative inverse; that is, for any $ay+b \in \mathbb{S}\setminus\{0\}$, we can find $cy+d \in \mathbb{S}$ such that $(ay+b)(cy+d) = 1$.

(e) The distributive property holds for elements of $\mathbb{S}$: that is, for any $s,t,u \in \mathbb{S}$, we have $s(t+u) = st+su$.
Hint: Again, you have to respect the form of elements of $\mathbb{S}$; redefine $s = ay+b$ and similarly for t and u, then work with those forms.

Exercise 2.68. The technique we adopted in this section will construct a root of any polynomial. For instance, we can use it to construct $\sqrt{2}$.

(a) What polynomial f should we use to construct $\sqrt{2}$?

(b) Suppose we build a set $\mathbb{S} = \{ay+b : a,b \in \mathbb{Q}\}$ and define addition, subtraction, and multiplication all modulo $f(y)$. What is the simplified form of the product $(ay+b)(cy+d)$? (By "$f(y)$", I mean, "substitute y for the variable of x.)
Hint: y^2 should not appear in the simplified form; use $f(y)$ to reduce it.

(c) Show that, when you take this route, $y \times y = 2$ and also $(-y) \times (-y) = 2$.

(d) Show that y is a root of the polynomial you listed in (a).

Sage supplement

We can define congruence modulo a polynomial in a manner analogous to the way we defined congruence modulo an integer. Let's start with Example 2.54.

```
sage: d = x^2 + 1
sage: Zd = ZZ[x].quo(d)
sage: Zd
Univariate Quotient Polynomial Ring in xbar over Integer
Ring with modulus x^2 + 1
```

This tells us that `Qd` stands for a set of polynomials in "`xbar`" whose coefficients are integers, and whose modulus is x^2+1.

You are probably wondering what `xbar` means. It is common in mathematics to write elements modulo an element with a bar over them. For instance, $\overline{12} = \overline{2} \in \mathbb{Z}_5$. In the same way, Sage is telling us that elements of `Zd` looks are polynomials in the variable $\overline{x}$, and operations are modulo $\overline{x}^2+1$. We can verify this by "injecting" $\overline{x}$ into our work and performing some calculations.

```
sage: Zd.inject_variables()
Defining xbar
sage: xbar^2 + 1
0
```

We see that Sage interprets $\overline{x}^2 + 1 = 0$. Notice that these modular computations *do not* apply to x, so that $x^2 + 1$ simplifies as before.

```
sage: x^2 + 1
x^2 + 1
```

We return to Example 2.54. The example claims that if $f = x + 3$ and $g = 5x^4 + 2x^2 + x$,then $f \equiv g \pmod{d}$. So long as we coerce the polynomials to reside in Zd, we get the expected results. There are two ways to do that: use the same coercion techniques we have used until now, or define the polynomial using xbar instead of x.

All three computations below are correct, but only two obtain the result we want. Make sure you understand why.

```
sage: ( x + 3 ) == ( 5*x^4 + 2*x^2 + x )
x + 3 == 5*x^4 + 2*x^2 + x
sage: Zd( x + 3 ) == Zd( 5*x^4 + 2*x^2 + x )
True
sage: ( xbar + 3 ) == ( 5*xbar^4 + 2*xbar^2 + xbar )
False
```

We can also verify that $h = 5x^4 + 2x^2$ is congruent to neither f nor g. We leave that to you as an exercise.

If you have completed the main text up to this point, you may have realized that xbar corresponds to the object we called y in the set $\mathbb{S}$. We proved in Lemma 2.62 that y behaved just like the imaginary number i. In particular, $y^2 + 1$, or, y is a root of $x^2 + 1$. Let's see if that pans out here.

```
sage: f = x^2 + 1
sage: f(x = xbar)
Traceback (click to the left of this block for traceback)
...
TypeError:  no canonical coercion from Univariate Quotient
Polynomial Ring in xbar over Integer Ring with modulus x^2 +
1 to Symbolic Ring
```

Here, Sage is telling us that it cannot combine $\overline{x}$, which lives in `Zd`, with $x^2 + 1$, which lives in the "symbolic ring." The cause is not with our principles, but that we neglected to tell Sage to view `f` as having integer coefficients. We'll coerce it and try again.

```
sage: f = ZZ[x](x^2 + 1)
sage: f(x = xbar)
0
```

It worked out! Substituting $\overline{x}$ in place of x gave us 0! In other words, we have constructed in Sage the symbol `xbar`, which works exactly the same way that the imaginary number i works: just as $i^2 + 1 = 0$, we get `xbar^2 + 1 == 0`.

You should verify that this works even if we situation `f`'s coefficients in $\mathbb{Q}[x]$.

Exercises

Exercise 2.69. Use Sage to replicate and verify your work in Exercise 2.68. Be sure to check that some other properties hold, such as `xbar^2 + 1 ==` …?

Chapter 3

Rings and fields

The previous chapters reinforced a great deal of what you should have learned already about integers and polynomials, and perhaps you learned something new when it came to modular arithmetic. You may also have noticed that both the integers and the polynomials have a great deal in common. For instance:

- Both have operations called addition, subtraction, and multiplication, where:
  - Addition and multiplication are both commutative and associative, and have identity elements.
  - Multiplication distributes over addition.
  - For addition, we can always find an "inverse" element, and adding a number to its inverse yields the identity.

  It's important to point out that rational and real numbers also enjoy these properties.

- For rational and real numbers, division is an operation, but for integers and polynomials, division is not an operation, as it results in two elements instead of one: quotient and remainder. This distinction is worth keeping in mind, as it tells us that, at least in this sense, integers are more akin to polynomials than to rational or real numbers!

- For instance, this curious property of integer and polynomial division results in modular operations for both integers and polynomials. The modular operation is closed, commutative, and associative, and the operation also has an identity, though it may not be invertible, even for non-zero elements. It also turns out to be very useful for certain applications:
  - With modular integer arithmetic, we derive modern cryptography.
  - With modular polynomial arithmetic, we can construct the so-called "imaginary" number, i.

  By contrast, division of rational and real numbers does *not* lead to a new, modular arithmetic. It has important applications, but those applicatons are very different.

Other similarities exist, and you might want to think about what they are.

Noticing that integers and polynomials share so many similarities, mathematicians of the 18th and 19th centuries began to describe a structure that underlies both systems. Investigating this common structure can be a bit tough at first, because we have to work with abstract symbols rather than concrete numbers or polynomials (or other objects!), but if you keep at it, it will reward you! Since the discovery of these structures, mathematics has becme a much more powerful tool for solving problems.

This chapter introduces you to those structures.

3.1 Rings

The main structure we consider is a ***ring***.

Informally, we can think of a ring as "a set of objects where addition is *very* well-behaved, but multiplication is *merely* well-behaved, and distributes over addition."

More precisely, let R be a set where addition and multiplication are defined in some way. We say that R is a ring if:

- addition on R is closed, commutative, associative, invertible, and satisfies the identity;
- multiplication is closed, commutative, associative, and satisfies the identity; and
- multiplication distributes over addition; that is, any three elements $r, s, t \in R$ satisfy $r(s+t) = rs + rt$.

We normally write a ring's additive identity as 0, and its multiplicative identity as 1. Given a ring element r, we normally write its additive inverse as $-r$.

Example 3.1. The set $\mathbb{N}$ is *not* a ring, because addition is not invertible: $2 + (-5) = -3$, and $-3 \notin \mathbb{N}$.

Example 3.2. Although we didn't call it a ring, we established in Section 1.2 that the set $\mathbb{Z}$ is a ring under addition and multiplication. We established in Section 2.1 that $\mathbb{Z}[x]$, and $\mathbb{Q}[x]$ are rings.

We also showed earlier that $\mathbb{Q}$ and $\mathbb{R}$ are also rings, but they differ from $\mathbb{Z}$ in that every nonzero element has a *multiplicative* inverse. It is worth remembering when a ring has this property, so we give it a special name: a ring with a multiplicative inverse is a ***field***.

Exercise 2.67 shows that the set $\mathbb{S}$ that we defined there is also a ring. What we have not yet considered is whether $\mathbb{Z}_n$ is a ring.

Is $\mathbb{Z}_m$ a ring?

Theorem 3.3. *For any $m \geq 2$, the set $\mathbb{Z}_m$ is a ring under addition and multiplication modulo m.*

Proof. We split this into three cases: one to show that addition is very well-behaved; one to show that multiplication is merely well-behaved; and one to show that multiplication distributes over addition.

Case 1. Is addition very well-behaved?

First we consider addition. For the sake of clarity in this proof, we briefly revive the use of $\oplus$ from page 49 to indicate "addition modulo m."

closure? We explained on page 49 that addition modulo m is closed.

commutative? Let $a, b \in \mathbb{Z}_m$. Let $r \in \mathbb{Z}_m$ such that $a \oplus b = r$. By definition, this is a remainder after adding a and b and dividing by m, so we can choose $q \in \mathbb{Z}$ such that $a + b = qm + r$. The left-hand side consists entirely of integer addition, which is commutative; that is, $a + b = b + a$. By substitution, $b + a = qm + r$. Recall that $r \in \mathbb{Z}_m$, and remainders of integer division are unique, so r is the remainder of dividing $b + a$ by m. By definition, $b \oplus a = r$. By substitution, $a \oplus b = b \oplus a$, and addition modulo m is commutative.

associative? Let $a, b, c \in \mathbb{Z}_m$. Let $\hat{r}, \check{r} \in \mathbb{Z}_m$ such that $a \oplus b = \hat{r}$ and $b \oplus c = \check{r}$, and let $\overrightarrow{r}, \overleftarrow{r} \in \mathbb{Z}_m$ such that $(a \oplus b) \oplus c = \overrightarrow{r}$ and $a \oplus (b \oplus c) = \overleftarrow{r}$. By definition, these are remainders after adding the numbers as ordinary integers and dividing by m, so we can choose $\hat{q}, \check{q}, \overrightarrow{r}, \overleftarrow{r} \in \mathbb{Z}$ such that

$$a+b = \hat{q}m+\hat{r} \quad , \quad b+c = \check{q}m+\check{r} \quad , \quad (a \oplus b)+c = \overrightarrow{q}m+\overrightarrow{r} \quad , \quad \text{and} \quad a+(b \oplus c) = \overleftarrow{q}m+\overleftarrow{r} \, .$$

By substitution and some rewriting,

$$\begin{aligned}(a \oplus b) \oplus c &= \overrightarrow{r} \\ &= [(a \oplus b) + c] - \overrightarrow{q}m \\ &= (\hat{r} + c) - \overrightarrow{q}m \\ &= ([(a + b) - \hat{q}m] + c) - \overrightarrow{q}m \\ &= [(a + b) + c] - \left(\hat{q} + \overrightarrow{q}\right) m \, .\end{aligned}$$

Similarly,

$$a \oplus (b \oplus c) = [a + (b + c)] - \left(\check{q} + \overleftarrow{q}\right) m \, .$$

By substitution,

$$[(a \oplus b) \oplus c]-[a \oplus (b \oplus c)] = \left([(a + b) + c] - \left(\hat{q} + \overrightarrow{q}\right) m\right)-\left([a + (b + c)] - \left(\check{q} + \overleftarrow{q}\right) m\right) \, .$$

The right-hand side consists entirely of integer addition, which is associative and commutative, so we can rewrite the equation as

$$[(a \oplus b) \oplus c] - [a \oplus (b \oplus c)] = ([(a + b) + c] - [a + (b + c)]) + \left[\left(\check{q} + \overleftarrow{q}\right) - \left(\hat{q} + \overrightarrow{q}\right)\right] m \, ,$$

and this simplifies to

$$[(a \oplus b) \oplus c] - [a \oplus (b \oplus c)] = \left[\left(\check{q} + \overleftarrow{q}\right) - \left(\hat{q} + \overrightarrow{q}\right)\right] m \, .$$

Hence, m divides the left hand side. The left hand side is a difference of elements of $\mathbb{Z}_m$; that is,

$$0 \quad \leq \quad (a \oplus b) \oplus c \, , \; a \oplus (b \oplus c) \; < m \, ,$$

and so

$$-m \quad < \quad [(a \oplus b) \oplus c] - [a \oplus (b \oplus c)] \quad < \quad m \ .$$

We showed just above that m divides that difference in the center, and the only multiple of m between $-m$ and m is 0. By process of elimination,

$$[(a \oplus b) \oplus c] - [a \oplus (b \oplus c)] = 0 \ ,$$

or in other words,

$$(a \oplus b) \oplus c = a \oplus (b \oplus c) \ ,$$

as desired.

identity? We claim that 0 is an additive identity for $\mathbb{Z}_m$. This is a natural guess, since 0 is an additive identity for $\mathbb{Z}$. First observe that, by definition, $0 \in \mathbb{Z}_m$. Now let $a \in \mathbb{Z}_m$. By definition of addition modulo m,

$$a \oplus 0 = \overline{a + 0} = \overline{a} \quad \text{and} \quad 0 \oplus a = \overline{0 + a} = \overline{a} \ .$$

Recall that $a \in \mathbb{Z}_m$. Elements of $\mathbb{Z}_m$ are the natural numbers $0, 1, \ldots, m-1$, so $\overline{a} = a$. Hence, $a \oplus 0 = a$ and $0 \oplus a = a$, and we have achieved our goal of showing that 0 is the additive identity of $\mathbb{Z}_m$.

invertible? Let $a \in \mathbb{Z}_m$. Don't make the mistake of thinking that $-a$ is the inverse of a: $-a$ is negative, and $\mathbb{Z}_m$ consists entirely of natural numbers. An element's inverse *has* to be in the same set as the original element.

Since $-a$ is out of the question, what else could be the inverse of a? If you remember how congruence works: any element of $\mathbb{Z}_m$ that is congruent to $-a$ will give the same result. Congruence is simply the result of adding multiples of m, so we could try $-a+m$ — or, written differently, $m-a$. Recall that $0 \leq a < m$, so $0 < m - a \leq m$. As long as $m - a < m$, we are OK, because

$$a + (m - a) = m + (a - a) = 1 \times m + 0 \ ,$$

so the remainder of dividing $a + (m - a)$ by m is 0; that is, $a \oplus (m - a) = 0$. Similarly, $(m - a) \oplus a = 0$, so $m - a$ is the inverse of a whenever $a \neq 0$.

Unfortunately, if $a = 0$, then $m - a = m \notin \mathbb{Z}_m$, and again we have a problem. But this isn't a big problem, because if $a = 0$ then $a \oplus a = 0 \oplus 0$, and $0 + 0 = 0$, which is its own remainder after division by m. By definition, 0 is its own inverse, and that covers all the cases.

Case 2. Is multiplication merely well-behaved?

We leave this to the exercises.

Case 3. Does multiplication distribute over addition?

Let $a, b, c \in \mathbb{Z}_m$. We need to show that

$$a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c) \ . \tag{3.1}$$

We will do this by showing that the left and right hand sides are equal.

To do this, choose $r, \hat{r}, \check{r}, \overleftarrow{r}, \overrightarrow{r} \in \mathbb{Z}_m$ such that

$$b \oplus c = r\ ,\ a \otimes b = \hat{r}\ ,\ a \otimes c = \check{r}\ ,\ a \otimes (b \oplus c) = \overleftarrow{r}\ ,\ \text{and}\ (a \otimes b) \oplus (a \otimes c) = \overrightarrow{r}\ .$$

Now choose quotients and remainders $q, \hat{q}, \check{q}, \overleftarrow{q}, \overrightarrow{q} \in \mathbb{Z}$ and such that, by substitution and rewriting,

$$\begin{aligned}
b + c &= qm + r\ , &&\Longrightarrow\ r = (b + c) - qm\ ;\\
a \times b &= \hat{q}m + \hat{r}\ , &&\Longrightarrow\ \hat{r} = (a \times b) - \hat{q}m\ ;\\
a \times c &= \check{q}m + \check{r}\ , &&\Longrightarrow\ \check{r} = (a \times c) - \check{q}m\ ;\\
a \times (b \oplus c) &= \overleftarrow{q}m + \overleftarrow{r}\ , &&\Longrightarrow\ \overleftarrow{r} = a \times (b \oplus c) - \overleftarrow{q}m\\
&&&\phantom{\Longrightarrow\ \overleftarrow{r}} = a \times r - \overleftarrow{q}m\\
&&&\phantom{\Longrightarrow\ \overleftarrow{r}} = a \times [(b + c) - qm] - \overleftarrow{q}m\\
&&&\phantom{\Longrightarrow\ \overleftarrow{r}} = a \times (b + c) - \left(aq + \overleftarrow{q}\right)m\ ;\\
(a \otimes b) + (a \otimes c) &= \overrightarrow{q}m + \overrightarrow{r}\ , &&\Longrightarrow\ \overrightarrow{r} = [(a \otimes b) + (a \otimes c)] - \overrightarrow{q}m\\
&&&\phantom{\Longrightarrow\ \overrightarrow{r}} = (\hat{r} + \check{r}) - \overrightarrow{q}m\\
&&&\phantom{\Longrightarrow\ \overrightarrow{r}} = [[(a \times b) - \hat{q}m] + [(a \times c) - \check{q}m]] - \overrightarrow{q}m\\
&&&\phantom{\Longrightarrow\ \overrightarrow{r}} = [(a \times b) + (a \times c)] - \left(\hat{q} + \check{q} + \overrightarrow{q}\right)m\ .
\end{aligned}$$

We can now substitute into the left hand side of (3.1) to obtain

$$a \otimes (b \oplus c) = \overleftarrow{r} = a \times (b + c) - \left(aq + \overleftarrow{q}\right)m\ .$$

The right hand side of this new equation consists of multiplication and addition of integers. Multiplication distributes over the integers in $\mathbb{Z}$, so

$$a \otimes (b \oplus c) = (a \times b + a \times c) - \left(aq + \overleftarrow{q}\right)m\ .$$

Rewrite this as

$$[a \otimes (b \oplus c)] + \left(aq + \overleftarrow{q}\right)m = a \times b + a \times c\ . \tag{3.2}$$

We turn to the right hand side of (3.1). Again by substitution, we have

$$(a \otimes b) \oplus (a \otimes c) = \overrightarrow{r} = (a \times b + a \times c) - \left(\hat{q} + \check{q} + \overrightarrow{q}\right)m\ .$$

Rewrite this as

$$[(a \otimes b) \oplus (a \otimes c)] + \left(\hat{q} + \check{q} + \overrightarrow{q}\right)m = a \times b + a \times c\ . \tag{3.3}$$

Observe that (3.2) and (3.3) have the same right-hand side, so by substitution and then a rewriting, we have

$$\begin{aligned}
[a \otimes (b \oplus c)] + \left(aq + \overleftarrow{q}\right)m &= [(a \otimes b) \oplus (a \otimes c)] + \left(\hat{q} + \check{q} + \overrightarrow{q}\right)m\\
[a \otimes (b \oplus c)] - [(a \otimes b) \oplus (a \otimes c)] &= \left(aq + \overleftarrow{q}\right)m - \left(\hat{q} + \check{q} + \overrightarrow{q}\right)m\\
&= \left[\left(aq + \overleftarrow{q}\right) - \left(\hat{q} + \check{q} + \overrightarrow{q}\right)\right]m\ .
\end{aligned}$$

The right hand side is a multiple of m, so the left hand side must also be a multiple of m. Hence

$$[a \otimes (b \oplus c)] - [(a \otimes b) \oplus (a \otimes c)] \quad \in \quad \{\ldots, -m, 0, m, 2m, \ldots\} \ . \tag{3.4}$$

Recall that

$$a \otimes (b \oplus c) \ , \ (a \otimes b) \oplus (a \otimes c) \quad \in \quad \mathbb{Z}_m \ ,$$

so by definition,

$$0 \quad \leq \quad a \otimes (b \oplus c) \ , \ (a \otimes b) \oplus (a \otimes c) \quad < \quad m \ ,$$

which implies that

$$-m \quad < \quad [a \otimes (b \oplus c)] - [(a \otimes b) \oplus (a \otimes c)] \quad < \quad m. \tag{3.5}$$

The only way that both (3.4) and (3.5) can be true is if

$$[a \otimes (b \oplus c)] - [(a \otimes b) \oplus (a \otimes c)] \quad = \quad 0 \ ,$$

and we have the desired property,

$$a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c) \ .$$

□

Now that we know that $\mathbb{Z}_m$ is a ring, what do you think the natural question should be?

Is $\mathbb{Z}_m$ a field?

If you remember the material in Section 1.5, you can actually work this out on your own! The main result that helps us here is Theorem 1.96.

Theorem 3.4. *Let $m \geq 2$. The ring $\mathbb{Z}_m$ is a field if and only if m is prime.*

Proof. By Theorem 3.3, $\mathbb{Z}_m$ is a ring.

First assume that $\mathbb{Z}_m$ is a field. By definition, every nonzero element of $\mathbb{Z}_m$ has a multiplicative inverse. By Theorem 1.96, a nonzero element a of $\mathbb{Z}_m$ has a multiplicative inverse if and only if $\gcd(a, m) = 1$. Hence, every nonzero $a \in \mathbb{Z}_m$ is relatively prime to m. That means none of $\{2, 3, \ldots, m-1\}$ is a divisor of m. The only natural divisors of m are thus 1 and m. We conclude that m is irreducible, or prime.

Conversely, assume that m is prime. By definition, it has only two natural divisors, 1 and m. Since none of $2, 3, \ldots, m-1$ divides the irreducible number m, we conclude that $\gcd(a, m) = 1$ for each nonzero $a \in \mathbb{Z}_m$. By Theorem 1.96, each nonzero $a \in \mathbb{Z}_m$ has a multiplicative inverse in $\mathbb{Z}_m$. By definition, $\mathbb{Z}_m$ is a field. □

Two important ring properties

The fact that $\mathbb{Z}_m$ is a ring should startle you, because an important property that we often use to solve equations is not necessarily true! Corollary 1.99 told us that the zero product property is true in $\mathbb{Z}_m$ only when m is prime. This means that, given an arbitrary ring R and two elements $r, s \in R$ such that $rs = 0$, we cannot conclude that $r = 0$ or $s = 0$: it's entirely possible that neither of them is!

Example 3.5 ((Reminder))**.** We have just shown that $\mathbb{Z}_6$ is a ring. Both $2, 3 \in \mathbb{Z}_6$, and neither is 0, but $2 \times 3 = 0$ in $\mathbb{Z}_6$.

Rings that satisfy the zero product property are called ***integral domains***. They are very important, but we will not focus very much on them in this course.

This troubles enough that we have to ask: are we guaranteed that $0 \times r = 0$? Fortunately, the answer is *yes*.

Theorem 3.6. *For any ring R, and for any $r \in R$, we have $0 \times r = 0$.*

Proof. Let R be a ring, and $r \in R$. By distribution, $r \times (0+0) = r \times 0 + r \times 0$. By the additive identity property, $0 + 0 = 0$. By substitution, $r \times 0 = r \times (0+0) = r \times 0 + r \times 0$ and $r \times 0 = r \times 0 + r \times 0$. Rings contain the inverses of their elements, so $-(r \times 0) \in R$. Apply substitution and various rings properties to rewrite

$$0 = -(r \times 0) + (r \times 0) = -(r \times 0) + [r \times 0 + r \times 0] = [-(r \times 0) + r \times 0] + r \times 0 = 0 + r \times 0 = r \times 0\,. \tag{3.6}$$

The ends of the chain tell us that $0 = r \times 0$. □

The other important ring property we need involves additive inverses. You are accustomed to writing $-2 = (-1) \times 2$ for the integers, and you may be tempted to do this for ring elements as well — but the issue with zero divisors should give you pause. Is it really true that, for any ring R and any $r \in R$, we can write $-r = (-1) \times r$?

Fortunately, this actually pans out.

Theorem 3.7. *For any ring R, and for any $r \in R$, we have $(-1) \times r = -r$ and $r \times (-1) = -r$.*

Proof. Let R be a ring, and $r \in R$. By definition of additive inverses, $-1 + 1 = 0$. By substitution, $(-1+1) \times r = 0 \times r$. By Theorem 3.6, $0 \times r = 0$, so by substitution, $(-1+1) \times r = 0$. By distribution, $(-1) \times r + 1 \times r = 0$. By the multiplicative identity property of a ring, $1 \times r = 0$, so by substitution, $(-1) \times r + r = 0$. By substitution, $[(-1) \times r + r] + (-r) = 0 + (-r)$. By the associative and identity properties of addition, we can rewrite this as $(-1) \times r + [r + (-r)] = -r$. By the inverse and identity properties of addition, we can rewrite this as $(-1) \times r + 0 = -r$, and thus as $(-1) \times r = -r$. This proves the first claim; the second is similar. □

Exercises

Exercise 3.8. In equation (3.6), each equals sign is justified by substitution or by a ring property, or both. Indicate which.

Exercise 3.9. Much like Theorem 3.6, it turns out that additive inverses are multiples of 1's additive inverse; that is, $-r = -1 \times r$. To prove this, we offer the following sequence of assertions. Explain why each assertion is true. Remember that your explanation must be either a ring property or the word "substitution".

(a) $1 + (-1) = 0$

(b) $r \times [1 + (-1)] = r \times 0$

(c) $r \times 0 = 0$

(d) $r \times [1 + (-1)] = 0$

(e) $r \times 1 + r \times (-1) = 0$

(f) $r \times 1 = r$

(g) $r + r \times (-1) = 0$

(h) $-r + [r + r \times (-1)] = -r + 0$

(i) $(-r + r) + r \times (-1) = -r + 0$

(j) $-r + r = 0$

(k) $0 + r \times (-1) = -r + 0$

(l) $0 + r \times (-1) = r \times (-1)$

(m) $-r + 0 = -r$

(n) $r \times (-1) = -r$

Exercise 3.10. Complete the proof of Theorem 3.3 by showing that multiplication in $\mathbb{Z}_m$ is closed, associative, and has an identity.

Exercise 3.11. Show that $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ are not fields.
Hint: By definition, you only have to find some element of the ring(s) that has no multiplicative inverse.

Exercise 3.12. Which of $\mathbb{Z}_{10}$, $\mathbb{Z}_{63}$, $\mathbb{Z}_{67}$, or $\mathbb{Z}_{117}$ is a field?

(a) What is the multiplicative inverse of 2 in that field?

(b) In the other rings, does 2 have a multiplicative inverse? If so, what is it? If not, why not?

Exercise 3.13. Show that the **zero product property** is always true in a field. That is, suppose $\mathbb{F}$ is a field, and let $a, b \in \mathbb{F}$. Show that if $ab = 0$, then $a = 0$ or $b = 0$.
Hint: Fields are more special than rings. Use what's special!

3.2 In the absence of division we have ideals

The previous section discussed a common structure that underlies the natural numbers, important sets of polynomials, and modular arithmetic. The operations involved are addition and multiplication. Mathematicians call this sort of thing ***generalization***: identifying which properties of one system ($\mathbb{Z}$) apply to a larger class of systems (rings).

Natural numbers and polynomials also have a Division Algorithm that computes a quotient and remainder. This is where modular arithmetic parts ways with the other rings we have studied: we do not generally perform division in $\mathbb{Z}_m$ — and when we do, it doesn't produce a quotient and remainder, because, as Theorem 3.4 points out, in these cases $\mathbb{Z}_m$ is a field, where division behaves as an operation, the same as it does in $\mathbb{Q}$ or $\mathbb{R}$.

Generalizing the Division Algorithm to rings is thus an interesting problem, and to understand how we can do it, we would do well to consider not so much the *process* of division as the *applications.*

Applications of division

For the natural numbers, three effects of division were

- common divisors (Section 1.3),
- prime numbers, also called irreducible (Section 1.4), and
- congruence (Section 1.5), which allowed us to build new rings ($\mathbb{Z}_m$) for applications (cryptography).

Polynomial division has the same effects (Sections 2.3, 2.4, and 2.5), and polynomial congruence likewise builds a new ring ($\mathbb{S}$) for applications (constructing the "imaginary" number). Both investigations start off with common divisors, which are based on the Euclidean algorithm. The Euclidean algorithm requires division, which — again — we don't have. Still, let's investigate what division does.

First, recall from Algorithm 1.1 that division is really repeated subtraction; we perform it until the remainder is smaller than the divisor. Subtraction is something we *do* have in a ring, and we *can* perform it repeatedly. Thus, the first attribute of division that we will try to generalize is repeated subtraction.

Another attribute of division — this may surprise you — is that it incorporates multiplication. Think here of the result: when we divide n by d, we obtain a quotient q and a remainder r such that $n - r = qd$. Right there you see not only subtraction, but division. Thus, the second attribute of division that we will try to generalize is divisibility, or, if you prefer, multiples.

With this in mind, let's try the following. Given a ring R, we say that an ***ideal subset*** of R is any nonempty set $I \subseteq R$ such that

- I is closed under subtraction; that is, $a - b \in I$ for all $a, b \in I$; and
- I ***absorbs*** multiplication from R; that is, $ra \in I$ for any $a \in I$ and any $r \in R$.

Let's work out what an "ideal subset" looks like in some of the rings we've studied.

Example 3.14. Suppose an ideal $I \subseteq \mathbb{Z}$ contains the numbers 15 and -18. Closed subtraction means that

- $15 - (-18) = 33 \in I$;
- $33 - 15 = 18 \in I$; and
- $18 - 15 = 3 \in I$.

This last discovery is interesting, because $3 = \gcd(15, -18)$. We have more to say about this later.

We haven't considered absorption yet. Since 3 is a divisor of every element of I we've seen so far, we'll look at its multiples. It implies that

- $0 \times 3 = 0 \in I$;
- $-1 \times 3 = -3 \in I$; and in fact

- $3z \in I$ for every $z \in \mathbb{Z}$.

Every multiple of 3 is in I. If neither 2 nor 1 is in I, then I is the set of all multiples of 3.

The last few words of Example 3.14 are more important than you might think at first glance, and we will return to this idea shortly.

Example 3.15. Suppose an ideal $J \subseteq \mathbb{Z}[x]$ contains the polynomials $x^3 + x^2$ and $x^2 - 1$.

- Absorption means that $x\left(x^2 - 1\right) = x^3 - x \in J$.
- Closure of subtraction means that $\left(x^3 + x^2\right) - \left(x^3 - x\right) = x^2 + x \in J$.
- Closure of subtraction also means that $\left(x^2 + x\right) - \left(x^2 - 1\right) = x + 1 \in J$.

Once again, we have just discovered $x + 1 = \gcd\left(x^3 + x^2, x^2 - 1\right)$. Now, $x + 1$ is irreducible! Perhaps J consists of the set of all multiples of $x + 1$. Notice the similarity to Example 3.14, where $3 = \gcd(15, -18)$.

It's important to point out that we don't actually have enough information to conclude that I is the set of multiples of 3, or that J is the set of multiples of $x + 1$. They may well contain other elements! Nevertheless, they point to an important class of ideals.

Generating ideals

In Examples 3.14 and 3.15 we started with two elements of a ring, and looked at what else might be an ideal. Of particular interest would be the smallest ideal that contains those elements. In general, if R is a ring and $r_1, \ldots, r_m \in R$, we say that the smallest ideal containing $r_1, \ldots, r_m$ is ***the ideal generated by*** $r_1, \ldots, r_m$, and write $\langle r_1, \ldots, r_m \rangle$ for short.

Example 3.16. Let $I = \langle 15, -18 \rangle$; that is, let I be the ideal generated by 15 and -18. Example 3.14 showed that $3 \in I$, or, $3 \in \langle 15, -18 \rangle$. Of course, $15, -18 \in \langle 3 \rangle$, because they are multiples of 3. Now, every element of I is either

- a multiple of $15 = 3 \times 5$, and thus in $\langle 3 \rangle$; or,
- a multiple of $18 = 3 \times 6$, and thus in $\langle 3 \rangle$; or,
- a difference two other elements of I.

It's starting to look as if $I = \langle 3 \rangle$.

Theorem 3.17. *Let R be a ring, and $a_1, \ldots, a_m \in R$. Then $\langle a_1, \ldots, a_m \rangle$ is the set of all sums of multiples of the a_i's. In symbols, $\langle a_1, \ldots, a_m \rangle = \{r_1 a_1 + \cdots + r_m a_m : r_i \in R\}$.*

Proof. We need to show that two sets are equal. That requires us to show that each is a subset of the other.

First, let $x \in \{r_1 a_1 + \cdots + r_m a_m : r_i \in R\}$. By definition, there exist $r_1, \ldots, r_m \in R$ such that $x = r_1 a_1 + \cdots + r_m a_m$. Rings satisfy the additive inverse property, so $-r_2, \ldots, -r_m \in R$. Consider the following chain of operations:

- Ideals absorb multiplication from R, so all of $r_1a_1, -r_2a_2, \ldots, -r_ma_m \in \langle a_1, \ldots, a_m\rangle$.
- Ideals are closed under subtraction, so
  - $r_1a_1 - (-r_2a_2) = r_1a_1 + r_2a_2 \in \langle a_1, \ldots, a_m\rangle$; and
  - $(r_1a_1 + r_2a_2) - (-r_3a_3) = r_1a_1 + r_2a_2 + r_3a_3 \in \langle a_1, \ldots, a_m\rangle$; and

    $\ldots$
  - $(r_1a_1 + r_2a_2 + \cdots + r_{m-1}a_{m-1}) - (-r_ma_m) = r_1a_1 + \cdots + r_ma_m \in \langle a_1, \ldots, a_m\rangle$.

This last element is x! We took an arbitrary x from $\{r_1a_1 + \cdots + r_ma_m : r_i \in R\}$, and we showed that $x \in \langle a_1, \ldots, a_m\rangle$. We have thus shown that

$$\{r_1a_1 + \cdots + r_ma_m : r_i \in R\} \subseteq \langle a_1, \ldots, a_m\rangle \ .$$

For the converse, we show two things. First, we show that each $a_1, \ldots, a_m$ is an element of $\{r_1a_1 + \cdots + r_ma_m : r_i \in R\}$. Then, we show that $\{r_1a_1 + \cdots + r_ma_m : r_i \in R\}$ is itself an ideal. We would then have

$$\{r_1a_1 + \cdots + r_ma_m : r_i \in R\} \supseteq \langle a_1, \ldots, a_m\rangle \ ,$$

so that each is a subset of the other, and they are equal.

For the first claim, recall that R has a multiplicative identity, which we write as 1, and an additive identity, which we write as 0. By definition of the set, and by Theorem 3.6,

$$\begin{gathered} 1 \cdot a_1 + 0 \cdot a_2 + \cdots + 0 \cdot a_m = a_1 \in \{r_1a_1 + \cdots + r_ma_m : r_i \in R\}, \\ \vdots \\ 0 \cdot a_1 + \cdots + 0 \cdot a_{m-1} + 1 \cdot a_m = a_m \in \{r_1a_1 + \cdots + r_ma_m : r_i \in R\}. \end{gathered}$$

So the first claim is proved.

For the second claim, we need to show that $\{r_1a_1 + \cdots + r_ma_m : r_i \in R\}$ is closed under subtraction and absorbs multiplication from R. To this end, take two arbitrary elements $b, c \in \{r_1a_1 + \cdots + r_ma_m : r_i \in R\}$. By definition, there exist $p_1, \ldots, p_m, q_1, \ldots, q_m \in R$ such that

$$b = p_1a_1 + \cdots + p_ma_m \quad \text{and} \quad c = q_1a_1 + \cdots + q_ma_m \ .$$

Their difference is

$$b - c = (p_1a_1 + \cdots + p_ma_m) - (q_1a_1 + \cdots + q_ma_m) \ .$$

By Exercise 3.9,

$$b - c = (p_1a_1 + \cdots + p_ma_m) + [(-q_1a_1) + \cdots + (-q_ma_m)] \ .$$

By the associative and commutative properties of ring addition,

$$b - c = (p_1 - q_1)\, a_1 + \cdots + (p_m - q_m)\, a_m \ .$$

Rings are closed under subtraction, so $p_i - q_i \in R$ for each i. By definition of the set, then,

$$b - c \in \{r_1a_1 + \cdots + r_ma_m : r_i \in R\} \ .$$

Meanwhile, let $r \in R$. By substitution,

$$rb = r\left(p_1a_1 + \cdots + p_m a_m\right) .$$

By the distributive property,

$$rb = r\left(p_1a_1\right) + \cdots + r\left(p_m a_m\right) .$$

By the associative property of ring multiplication,

$$rb = \left(rp_1\right)a_1 + \cdots + \left(rp_m\right)a_m .$$

Rings are closed under multiplicaiton, so $rp_i \in R$ for each i. By definition of the set, then,

$$rb \in \left\{r_1a_1 + \cdots + r_m a_m : r_i \in R\right\} .$$

We have shown that $\{r_1a_1 + \cdots + r_m a_m : r_i \in R\}$ is an ideal. As explained above, this completes the proof. □

Example 3.18. A simple example of an ideal generated by some elements is an ideal generated by only one element, called a ***principal ideal***. One example of a principal comes from the ring of integers: $\langle 3\rangle = \{3z : z \in \mathbb{Z}\}$. In the case of Example 3.16, it seems that $\langle 15, -18\rangle = \langle 3\rangle$. However, we still do not know this for certain.

More generally, let $d \in \mathbb{Z}$, and write $d\mathbb{Z} = \{dz : z \in \mathbb{Z}\}$; that is, $d\mathbb{Z}$ is the set of all integer multiples of d. Written another way, $d\mathbb{Z} = \langle d\rangle$. Theorem 3.17 tells us that $d\mathbb{Z}$ is a principal ideal.

Common divisors

Example 3.18 remarked that it looks as if $\langle 15, -18\rangle = \langle 3\rangle$, but we do not yet know this. Let's see if we can decide this. If you remember an observation from Example 3.14, $3 = \gcd(15, -18)$, so actually we have $\langle 15, -18\rangle = \langle \gcd(15, -18)\rangle$.

Is it always the case that $\langle r, s\rangle = \langle \gcd(r, s)\rangle$? You might think that we can't answer the question, as we haven't defined greatest common divisors for arbitrary rings, but in fact we can. The answer involves one theorem and one example.

Theorem 3.19. *Every ideal I of $\mathbb{Z}$ is principal, and the generator is the greatest common divisor of I's elements.*

Proof. Let I be an ideal of $\mathbb{Z}$, and let A be the set of I's positive elements. By the Well-Ordering Property, A has a minimal element; let's call it a. We claim that $I = \langle a\rangle$.

Let $n \in I$, and use the Division Algorithm to compute $q, r \in \mathbb{Z}$ such that $n = qa + r$ and $0 \leq r < a$. Rewrite $n = qa + r$ as $n - qa = r$. By absorption, $qa \in I$, and by closure of subtraction, $n - qa \in I$. By substitution, $r \in I$.

What do we know about r? We know that $r \in I$ and that $r \geq 0$. Is $r > 0$? If so, then by definition, $r \in A$. Recall that $r < a$; this is a problem, because a is supposed to be A's smallest element! This answer our question: we cannot have $r > 0$. Instead, $r = 0$.

We have shown that any $n \in I$ is divisible by a. Hence, a is a common divisor of I's elements. Of course, $a \in A$ and $A \subseteq I$ implies that $a \in I$ itself. No integer larger than a divides it, so a has to be the greatest common divisor of I's elements. □

If you look carefully at this proof, you'll notice that it relies on a property of integers that is not true for polynomials. Example 2.20 and Corollary 2.21 also give a hint on this: $\mathbb{Z}[x]$ is a ring, but its division algorithm does not guarantee what we need.

Example 3.20. The ideal $\langle 2, x\rangle$ is not principal in $\mathbb{Z}[x]$. To see why, suppose by way of contradiction that there were some $f \in \mathbb{Z}[x]$ such that $\langle 2, x\rangle = \langle f\rangle$. Then $f \mid 2$ and $f \mid x$. The only integers that divide both 2 and x are ± 1, so we are saying that $\langle 2, x\rangle = \langle 1\rangle$. By Theorem 3.17, this means that we can write $1 = 2p + xq$ for some $p, q \in \mathbb{Z}[x]$. If two polynomials are equal, each of their terms is equal, and there is only one term on the left-hand side of $1 = 2p + xq$, namely, 1 itself. So the constant term of $2p + xq$ must equal 1. But the constant term of $2p$ is a multiple of 2, and every term of xq is a multiple of x, so not constant. The polynomial $2p + xq$ has no constant terms! So $1 \neq 2p + xq$, and hence $\langle 1\rangle \neq \langle 2, x\rangle$. This was our only option to generate the entire ideal, so $\langle 2, x\rangle$ is not principal.

What a huge difference only one symbol can make! Rings like $\mathbb{Z}$, where every ideal is principal, are called ***principal ideal rings***, but as we see from $\mathbb{Z}[x]$, not every ring is a principal ideal ring.

Common roots

Speaking of polynomials, ideals have an important application to polynomials.

Theorem 3.21. *Let $f_1, \ldots, f_m$ be polynomials of $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, or $\mathbb{C}[x]$, and let $I = \langle f_1, \ldots, f_m\rangle$. Let a be any number such that $f_i(a) = 0$ for each i; that is, a is a common root of all the f's. Then a is a root of every $g \in I$.*

Proof. Let $g \in I$. By Theorem 3.17, we can find polynomials $p_1, \ldots, p_m$ such that $g = p_1 f_1 + \cdots + p_m f_m$. By substitution,

$$g(a) = p_1(a) f_1(a) + \cdots + p_m(a) f_m(a) = p_1(a) \times 0 + \cdots + p_m(a) \times 0 = 0\,;$$

that is, a is a root of g. □

Example 3.22. Let $f_1 = x^3 + x^2$ and $f_2 = x^2 - 1$. It is not hard to verify that $f_1(-1) = f_2(-1) = 0$; that is, -1 is a common root of f_1 and f_2. We saw in Example 3.15 that $g = x + 1 \in \left\langle x^3 + x^2, x^2 - 1\right\rangle$, so it must be that -1 is also a root of g — but this is obvious, since $g(-1) = -1 + 1 = 0$.

Theorem 3.21 is a very important tool in algebra, as it tells us that we can use propreties of ideals to analyze the roots common to all polynomials in the ideal.

Exercises

Exercise 3.23. Is $\langle 12, -18, 37\rangle$ a principal ideal in $\mathbb{Z}$? If so, what element generates it?

Exercise 3.24. Show that $0 \in I$ for any ideal I.

Exercise 3.25. Let I be an ideal of a ring. Show that if $a, b \in I$, then $a + b$ also in I.
Hint: You need the result of Exercise 3.24.

Exercise 3.26. Every ring R is itself an ideal. In fact, every ring is a principal ideal! Find an element $r \in R$ such that $R = \langle r \rangle$.
Hint: Look at some simple examples, such as $\mathbb{Z}$ and $\mathbb{Z}_6$.

Exercise 3.27. Suppose that a ring R is actually a field, like $\mathbb{Q}$ or $\mathbb{Z}_{17}$. Show that:

(a) The only two ideals possible are $\langle 0 \rangle$ and $\langle 1 \rangle$.

(b) R is a principal ideal ring.

Exercise 3.28. Show that $\mathbb{Q}[x]$ is a principal ideal ring.
Hint: Let I be any ideal of $\mathbb{Q}[x]$. Let D be the set of degrees of polynomials in I. You know that D has a minimal element d (why?) so choose from I any polynomial f of degree d. Now use a technique similar to that of Theorem 3.19 to show that f divides every element of I. By definition, then, $I = \langle f \rangle$.

Exercise 3.29. Are there any values of m for which $\mathbb{Z}_m$ is not a principal ideal ring?
Hint: Experiment with some small values of m. See if every ideal turns out to be principal for each value of m. If so, that should give you the insight you need to prove it in general. If not, then you've found a counterexample.
Another hint: Bézout's identity could prove helpful.

3.3 Cosets

Section 3.2 indicated that we would use ideals to obtain the same effects in rings that division has for integers and polynomials, but we listed three effects, and so far we've only really addressed one of them: common divisors. The second one, prime or irreducible numbers, we put off until Section 3.7. That leaves the question of congruence, which we now consider.

Congruence modulo an ideal

Both with the integers and with polynomials, we initially defined congruence using remainders, along these lines:

$$a \equiv b \pmod{m} \text{ if } a \text{ and } b \text{ have the same remainder after division by } m.$$

Abstract rings don't offer us division with remainder, so we have to take a different tack. Fortunately, we had already found a rather useful one: Theorems 1.82 and 2.53 tell us that

$$a \equiv b \pmod{m} \text{ if and only if } m \mid (a-b).$$

This gets us a little closer to something useful, though it still isn't as general as we want: for a principal ideal that is generated by only one element, say $I = \langle m \rangle$, it would be fine: just check that $m \mid (a-b)$. For an ideal that is generated by more than one element, such as $\langle 2, x \rangle$ from Example 3.20, then at the very least we'd have to check whether $2 \mid (a-b)$ and whether $m \mid (a-b)$.

Even that isn't enough. In Example 3.15, we found that the ideal $J = \left\langle x^3 + x^2, x^2 - 1 \right\rangle$ contained the polynomial $x+1$. Beacuse of this, we would want $2(x+1) = 2x+2$ and $3(x+1) = 3x+3$

to be congruent modulo J, but there's no way to detect that using divisibility by $x^3 + x^2$ or $x^2 - 1$. So we do something clever: *we state what we want, without saying how we get there.*

This is an important technique, and is a key step of abstraction. If we focus too much on *how* we compute something, we lose sight of the larger picture — a "miss the forest for the trees" affair. Instead, we try to back up and identify the simplest, most generic thing we want, without worrying whether we can actually accomplish it.

How, then, shall we define ***congruence in a ring*** R? Let I be an ideal of R, and let $a, b \in R$. We say that

$$a \equiv b \pmod I \text{ if } a - b \in I.$$

Consider some concrete examples to see how this works.

Example 3.30. Recall from Example 3.14 the ideal $I = \langle 18, -15 \rangle$. Are 22 and 3 congruent modulo I? What about 22 and 4?

Theorem 3.19 tells us that every ideal of $\mathbb{Z}$ is a principal ideal ring, generated by the elements' greatest common divisor. Well, I is an ideal of $\mathbb{Z}$, and $\gcd(18, -15) = 3$, so $I = \langle 3 \rangle$, as we pointed out at that time. If we look at 22 and 3, we see that $3 \nmid (22 - 3)$, so $22 - 3 \notin I$, so

$$22 \not\equiv 3 \pmod I .$$

On the other hand, $3 \mid (22 - 4)$, so $22 - 4 \in I$, so

$$22 \equiv 4 \pmod I .$$

Example 3.31. Recall from Example 3.15 the ideal $J = \left\langle x^3 + x^2, x^2 - 1 \right\rangle$. Are $3x + 3$ and $2x + 2$ congruent modulo I?

We are not quite so lucky with J as we were with I above — but we are not exactly unlucky, either, because we do know that $x + 1 \in J$ (and it wasn't hard to discover this). If we look at $3x + 3$ and $2x + 2$, we see that $x + 1 \mid [(3x + 3) - (2x + 2)]$, so $(3x + 3) - (2x + 2) \in J$, so

$$3x + 3 \equiv 2x + 2 \pmod J .$$

Example 3.32. Consider the ideal $K = \langle 2, x \rangle$. Are $x^2 - 4$ and $x^2 - 1$ congruent modulo K?

Again, we are not quite so lucky with K as we were with I, and we're also not quite as lucky as we were with J. However, if we look at $x^2 - 4$ and $x^2 - 1$, we see that

$$\left(x^2 - 4\right) - \left(x^2 - 1\right) = -3 .$$

Is $-3 \in K$?

Suppose for a moment that it were. By definition, $2 \in K$, and absorption, $-2 \times 2 \in K$, and by closure of subtraction, $-3 - (-2 \times 2) = 1 \in K$. We saw in Example 3.20 that $1 \notin K$, so that forces $-3 \notin K$, and thus

$$x^2 - 4 \not\equiv x^2 - 1 \pmod K .$$

Notice how this solves our problem: we can "decide" whether a and b are congruent, by "deciding" whether $a - b$ is an element of I. How do we decide whether $a - b \in I$? Unfortunately, we had to adapt *ad hoc* techniques to each case, so in general, who knows? That's not the point.[1]

[1] The problem of deciding whether an ideal contains some element turns out to be a surprisingly difficult problem in general.

I realize this sounds like a dodge, but the definition is surprisingly useful. Before we turn to its uses, let's address a question that surely burns in your brain (or should): Is congruence modulo an ideal, just like congruence modulo an integer or congruence modulo a polynomial, an equivalence relation? Indeed it is!

Theorem 3.33. *Let R be a ring, and I an ideal. Congruence modulo I is an equivalence relation in R.*

Proof. We need to show that congruence modulo an ideal is reflexive, symmetric, and transitive. We show transitive, and leave the others to the exercises. Let $a, b, c \in R$, and suppose that $a \equiv b \pmod{I}$ and $b \equiv c \pmod{I}$. By definition, $a-b, b-c \in I$, and by Exercise 3.25 $(a-b)+(b-c) \in I$. By the definition of subtraction, along with the associative, additive identity, and additive inverse properties,

$$(a-b)+(b-c) = (a+(-b))+(b+(-c)) = a+(-b+b)+(-c) = a+0+(-c) = a-c\,,$$

and by substitution $a-c \in I$. By definition, $a \equiv c \pmod{I}$, as claimed. □

Cosets

Congruence of integers allowed us to create the rings of modular integer arithmetic, $\mathbb{Z}_m$. Congruence of polynomials likewise allowed us to create a ring of modular polynomial arithmetic, $\mathbb{S}$. We have used ideals to define congruence of ring elements; can we use this foundation to build new rings using congruence? Indeed, we can.

We'll use the integers to give us insight on the question. For an example, work with $m = 3$. Our goal is to use the ideal $3\mathbb{Z} = \{\ldots, -3, 0, 3, 6, \ldots\}$ to develop an analogue to the ring $\mathbb{Z}_3 = \{0, 1, 2\}$. Division by 3 is related to the ideal $3\mathbb{Z}$, in that elements of $3\mathbb{Z}$ have remainder 0, while non-elements of $3\mathbb{Z}$ have nonzero remainder. The remainders themselves are elements of $\mathbb{Z}_3$.

Consider $2 \in \mathbb{Z}_3$: which integers have a remainder of 2? That would be the set $\{\ldots, -1, 2, 5, 8, \ldots\}$. In some sense, any time we perform mular arithmetic, the number $2 \in \mathbb{Z}_3$ "stands in" for this set. Notice, moreover, that

$$\{\ldots, -1, 2, 5, 8, \ldots\} = \{\ldots, 2+(-3), 2+0, 2+3, 2+6, \ldots\} = \{2+t : t \in 3\mathbb{Z}\}\ .$$

Every integer in that set of remainders is simply a multiple of 3, offset by 2! This makes perfect sense, because division by 3, with a remainder of 2, has the form $n = 3q+2$: a multiple of 3, offset by 2. It makes sense to write this as follows:

$$\{\ldots, -1, 2, 5, 8, \ldots\} \quad = \quad \underbrace{2}_{\text{offset by 2}} \quad + \quad \underbrace{3\mathbb{Z}}_{\text{multiples of 3}}\quad .$$

In general, we can write $r + d\mathbb{Z}$ for the integer multiples of d offset by r:

$$r + d\mathbb{Z} \quad = \quad \{\ \ldots\ ,\ r+(-d)\ ,\ r+0\ ,\ r+d\ ,\ r+2d\ ,\ \ldots\ \}\ .$$

All these numbers have the same remainder[2] when dividing by d!

We can now write the same idea three different ways:

[2] That remainder might not be r. For instance, the elements of $4 + 3\mathbb{Z}$ have remainder 1 when we dividing them by 3.

- 22 and 4 have the same remainder after division by 3;
- $22 \equiv 4 \pmod 3$; and
- $22 + 3\mathbb{Z} = 4 + 3\mathbb{Z}$.

This relationship means that we can test each of them the same way, which we describe as an extension of Theorem 1.82:

Theorem 3.34. *Let $a, b, d \in \mathbb{Z}$, with $d \geq 2$. The following are equivalent.*

(A) a and b have the same remainder after division by d;

(B) $a \equiv b \pmod d$;

(C) $d \mid (a - b)$; and

(D) $a + d\mathbb{Z} = b + d\mathbb{Z}$.

Proof. (A) is equivalent to (B) by the definition in Section 1.5. (B) is equivalent to (C) by Theorem 1.82. It remains to show that (C) is equivalent to (D).

The statement, "$a + d\mathbb{Z} = b + d\mathbb{Z}$," means that the two sets $a + d\mathbb{Z}$ and $b + d\mathbb{Z}$ are equal. This is true if and only if every element of $a + d\mathbb{Z}$ is an element of $b + d\mathbb{Z}$ and vice versa. That is, every $x \in a + d\mathbb{Z}$ is an element of $b + d\mathbb{Z}$, and every $y \in b + d\mathbb{Z}$ is an element of $a + d\mathbb{Z}$. Any $x \in a + d\mathbb{Z}$ has the form $a + dz$ for some $z \in \mathbb{Z}$, and $x \in b + d\mathbb{Z}$ if and only if we can find some $z' \in \mathbb{Z}$ such that $x = b + dz'$, in which case $a + dz = b + dz'$, which we can rewrite as $a - b = d(z' - z)$. By definition of divisibility, $d \mid a - b$. The same is true for y. Hence, $a + d\mathbb{Z} = b + d\mathbb{Z}$ if and only if $d \mid (a - b)$, as desired. □

This is interesting for integers, but what about rings in general?

Let R be a ring, and I an ideal. For any $a \in R$, the ***coset*** of I with a, which we also call a's coset with I, or simply a's coset, is the set

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

(Notice that this looks just like $r + d\mathbb{Z}$, where a fills in for r and I fills in for $d\mathbb{Z}$.) Given a coset $a + I$, we say that a is I's ***offset***.

The ***quotient of** R **modulo** I* is the set of all cosets of I in R. In symbols, we would write,

$$R/I = \{r + I : r \in R\} \ .$$

Example 3.35. Let $R = \mathbb{Z}$ and $I = 4\mathbb{Z}$. By definition,

$$R/I \quad = \quad \{r + I : r \in R\} \quad = \quad \{r + 4\mathbb{Z} : r \in \mathbb{Z}\} \ .$$

Theorem 3.34 tells us that two cosets $a + 4\mathbb{Z}$ and $b + 4\mathbb{Z}$ are equal if and only if their offsets have the same remainder after division by 4. There are only four possible remainders, so there are only four distinct cosets of $4\mathbb{Z}$,

$$0 + 4\mathbb{Z} = 4\mathbb{Z}\ , \ 1 + 4\mathbb{Z}\ , \ 2 + 4\mathbb{Z}\ , \ 3 + 4\mathbb{Z}\ .$$

So

$$\mathbb{Z}/4\mathbb{Z} \quad = \quad \{\, 4\mathbb{Z}\,,\ 1+4\mathbb{Z}\,,\ 2+4\mathbb{Z}\,,\ 3+4\mathbb{Z}\,\} \ .$$

This looks a lot like $\mathbb{Z}_4 = \{0, 1, 2, 3\}$. Eventually we will prove that, for all practical purposes, they are in fact the same.

That said, remember that $2 + 4\mathbb{Z} = 6 + 4\mathbb{Z} = -10 + 4\mathbb{Z}$, since $4 \mid (2-6)$ and $4 \mid (6-(-10))$. So we could also write

$$\mathbb{Z}/4\mathbb{Z} \quad = \quad \{\, 4\mathbb{Z}\,,\ 1+4\mathbb{Z}\,,\ 6+4\mathbb{Z}\,,\ 3+4\mathbb{Z}\,\} \ .$$

In the same line, we could write

$$\mathbb{Z}/4\mathbb{Z} \quad = \quad \{\, 12+4\mathbb{Z}\,,\ -3+4\mathbb{Z}\,,\ 6+4\mathbb{Z}\,,\ 27+4\mathbb{Z}\,\} \ .$$

Example 3.36. Let $R = \mathbb{Z}[x]$ and $J = \left\langle x^3 + x^2, x^2 - 1 \right\rangle$. Recall from previous examples that $J = \langle x + 1 \rangle$. By substitution, then,

$$R/J \quad = \quad \mathbb{Z}[x] / \left\langle x^3 + x^2, x^2 - 1 \right\rangle \quad = \quad \mathbb{Z}[x] / \langle x+1 \rangle \ .$$

By Corollary 2.21, we can actually divide polynomials by $x + 1$, so by Exercise 3.43, two cosets $f + \langle x+1 \rangle$ and $g + \langle x+1 \rangle$ are the same if and only if f and g have the same remainder after division by $x + 1$. So the cosets of $\langle x+1 \rangle$ have the form $r + \langle x+1 \rangle$, where r is a remainder after division by $x + 1$. Since $\deg(x+1) = 1$, and the remainders have to have smaller degree, the remainders all have degree 0. That is, the remainders are all constants. Hence

$$R/J \quad = \quad \{\, \ldots\,,\ -1 + \langle x+1 \rangle\,,\ 0 + \langle x+1 \rangle\,,\ 1 + \langle x+1 \rangle\,,\ 2 + \langle x+1 \rangle\,,\ \ldots\,\} \ .$$

This looks a lot like $\mathbb{Z} = \{\ldots, -1, 0, 1, 2, \ldots\}$. Eventually we will prove that, for all practical purposes, they are in fact the same.

Example 3.37. Let $R = \mathbb{Z}[x]$ and $K = \langle 2, x \rangle$. Unlike the previous examples, we cannot identify a single generator of K, but with a bit of cleverness we can still write elements of R/K in a convenient form.

First, x is a monic polynomial, so by Corollary 2.21 we can divide any element of $\mathbb{Z}[x]$ by x and obtain a remainder of smaller degree. As with the previous example, this smaller-degree polynomial *must* be a constant. Thus, there exist $q, q', r, r' \in \mathbb{Z}[x]$ such that $f = qx + r$, $g = q'x + r'$, and $\deg(r), \deg(r') < \deg(x) = 1$. This forces $\deg(r) = \deg(r') = 0$, which means that r is a constant integer.

Now, 2 divides any even number to a remainder of zero, and any odd number to a remainder of 1. So we can choose $d, d', r'', r''' \in \mathbb{Z}$ such that $r = 2d + r''$ and $r' = 2d + r'''$, where $r \in \{0, 1\}$. By substitution, $f = qx + 2d + r''$ and $g = q'x + 2d' + r'''$.

Suppose that f and g have the same remainder after this process; that is, $r'' = r'''$. Then

$$f - g \quad = \quad \left(qx + 2d + \cancel{r''}\right) - \left(q'x + 2d' + \cancel{r'''}\right) \quad = \quad (q - q')\,x + 2\,(d - d') \quad \in \quad \langle 2, x \rangle \ .$$

We said above that this means $f \equiv g \pmod{K}$. Theorem 3.38 below shows that this means they lie in exactly the same coset. The only remainders possible after dividing by x and by 2 are 0 and 1, so we can write all cosets of R/K in the form $r + \langle 2, x \rangle$, where r is either 0 or an odd constant. That is,

$$R/K \quad = \quad \{\, 0 + \langle 2, x \rangle\ ,\ 1 + \langle 2, x \rangle\,\} \ .$$

We conclude to what is possibly the most important theorem of this chapter: deciding when two cosets are equal.

Theorem 3.38 (Coset equality)**.** *Let R be a ring, $r, s \in R$, and I an ideal of R.*

(A) $0 + I = I$.

(B) $r \in r + I$.

(C) The following are equivalent.

- *(i)* $r + I = s + I$.
- *(ii)* $r - s \in I$.
- *(iii)* $r \equiv s \pmod{I}$.

Proof. We leave the proof of (A) and (B) to Exercise 3.41. For (C), statements (ii) and (iii) are equivalent by definition. It will suffice to show that (i) and (iii) are equivalent. For the equivalence, we must as usual show that (i) $\Longrightarrow$ (ii) and (i) $\Longleftarrow$ (ii).

Assume first that $r + I = s + I$. From (B), we know that $r \in r + I$; by substitution, $r \in s + I$. By definition, $r = s + i$ for some $i \in I$. Rewrite this as $r - s = i$, and we see immediately that $r - s \in I$.

Conversely, assume that $r - s \in I$. We need to show that $r + I = s + I$. These are sets, so we need to show that two sets are equal, which requires us to show that every $x \in r + I$ is also in $s + I$, and that every $y \in s + I$ is also in $r + I$. Let $x \in r + I$; by definition, $x = r + i$ for some $i \in I$. Rewrite this as $r = x - i$. Recall that $r - s \in I$; by definition, $r - s = j$ for some $j \in I$. By substitution, $(x - i) - s = j$. Rewrite this as $x = s + (i + j)$. By Exercise 3.25, $i + j \in I$. By definition, $x \in s + I$. The proof that any $y \in s + I$ is also in $r + I$ is similar.

We have shown that (ii)$\Longleftrightarrow$(iii), as desired. □

Exercises

Exercise 3.39. List all the elements of $\mathbb{Z}/\langle 24, 30\rangle$. Compare to the elements of $\mathbb{Z}_{\gcd(24,30)}$.

Exercise 3.40. List at least ten elements of $\mathbb{Z}[x]/\left\langle x^2 + 2\right\rangle$.

Exercise 3.41. Show that for any ideal I of any ring R, and for any $r \in R$,

(a) $0 + I = I$; and

(b) $r \in r + I$.

Exercise 3.42. Determine if the following statements are true.

(a) $128 \equiv 17 \pmod{\langle 7\rangle}$

(b) $x^5 - 2x \equiv 4x^3 - 4x^2 \pmod{\left\langle x^3 + x^2, x^2 - 1\right\rangle}$

(c) $(x + 2) + \langle 2, x\rangle = \langle 2, x\rangle$

(d) $x^5 - 2x + 3 \equiv 7x^3 + 5 \pmod{\langle 2, x\rangle}$

Exercise 3.43. Suppose $f \in \mathbb{Z}[x]$ is monic. By Corollary 2.21, we can divide other polynomials $g \in \mathbb{Z}[x]$ by f, and obtain a quotient and remainder such that $g = fq + r$ and $\deg(r) < \deg(f)$. Use this to show that the following statements are equivalent:

(a) $g + \langle f \rangle = h + \langle f \rangle$;

(b) $f \mid (g - h)$; and

(c) g and h have the same remainder after division by f.

Hint: This looks intimidating, but it's really a direct proof from the definitions of cosets and of set equality. In fact, you can imitate the proof of Theorem 3.34.

3.4 Quotients of ideals are also rings

Just as remainders and congruence in $\mathbb{Z}$ and $\mathbb{Q}[x]$ allowed us to create the rings $\mathbb{Z}_m$ and $\mathbb{S}$, their analogues in a general ring — cosets and ideals — allow us to create new rings.

Let R be a set and I an ideal of R. Recall that R/I is the set of I's cosets in R:

$$R/I = \{ r + I : r \in I \} .$$

We show in due course that R/I is also a ring, but first we need to identify addition and multiplication operations for this ring. We do so as follows:

- Let $X, Y \in R/I$.
- By definition, X and Y are cosets, so we can find $x, y \in R$ such that $X = x + I$ and $Y = y + I$.
- The sum and product of X and Y should also be cosets. We define

$$X + Y = (x + y) + I \quad \text{and} \quad XY = xy + I .$$

In other words, the sum (or product) of two cosets is the coset whose offset is the sum (or product) of their offsets.

Example 3.44. Let $R = \mathbb{Z}$ and $I = \langle 4 \rangle = \{\ldots, -4, 0, 4, 8, \ldots\}$. We saw in Example 3.35 that the set R/I has four elements: $0 + I = I$, $1 + I$, $2 + I$, $3 + I$. By definition,

$+$	I	$1+I$	$2+I$	$3+I$
I	I	$1+I$	$2+I$	$3+I$
$1+I$	$1+I$	$2+I$	$3+I$	I
$2+I$	$2+I$	$3+I$	I	$1+I$
$3+I$	$3+I$	I	$1+I$	$2+I$

$\times$	I	$1+I$	$2+I$	$3+I$
I	I	I	I	I
$1+I$	I	$1+I$	$2+I$	$3+I$
$2+I$	I	$2+I$	I	$2+I$
$3+I$	I	$3+I$	$2+I$	$1+I$

If you look closely, you'll see some strange things are happening in the multiplication table: $(2 + I) + (2 + I) = I$, $(2 + I) \times (3 + I) = 2 + I\ldots$ really? Indeed they are: By definition,

$$(2 + I) + (2 + I) = 4 + I = I \quad \text{and} \quad (2 + I) \times (3 + I) = 6 + I ,$$

and $6 - 2 = 4 \in I$, so by Theorem 3.38, $6 + I = 2 + I$.

This multiplication table looks a lot like the multiplication table of $\mathbb{Z}_4$. Eventually we'll prove that they are in fact the same.

Example 3.45. Let $R = \mathbb{Z}[x]$ and $I = \langle 2, x\rangle$. We saw in Example 3.35 that the set R/I has two elements: $0 + I = I$, $1 + I$. By definition,

$+$	I	$1+I$
I	I	$1+I$
$1+I$	$1+I$	I

$\times$	I	$1+I$
I	I	I
$1+I$	I	$1+I$

If you look closely, you'll see some strange things are happening in the multiplication table: $(1 + I) + (1 + I) = I$, $I \times (1 + I) = I\ldots$ really? Indeed they are: By definition,

$$(1+I)+(1+I) = 2+I \quad \text{and} \quad I \times (1+I) = (0+I)\times(1+I) = 0+I\,,$$

and by Theorem 3.38, $0 + I = 2 + I = I$ because $0, 2 \in I$.

This multiplication table looks a lot like the multiplication table of $\mathbb{Z}_2$. Eventually we'll prove that they are in fact the same.

Example 3.46. Let $R = \mathbb{Z}_{100}$ and $I = \langle 4\rangle = \{0, 4, 8, \ldots, 96\}$. The set R/I has four elements: $0 + I = I$, $1 + I$, $2 + I$, $3 + I$. By definition,

$+$	$0+I$	$1+I$	$2+I$	$3+I$
$0+I$	$0+I$	$1+I$	$2+I$	$3+I$
$1+I$	$1+I$	$2+I$	$3+I$	$0+I$
$2+I$	$2+I$	$3+I$	$0+I$	$1+I$
$3+I$	$3+I$	$0+I$	$1+I$	$2+I$

$\times$	$0+I$	$1+I$	$2+I$	$3+I$
$0+I$	$0+I$	$0+I$	$0+I$	$0+I$
$1+I$	$0+I$	$1+I$	$2+I$	$3+I$
$2+I$	$0+I$	$2+I$	$0+I$	$2+I$
$3+I$	$0+I$	$3+I$	$2+I$	$1+I$

If you look closely, you'll see some strange things are happening in the multiplication table: $(2 + I) \times (3 + I) = 2 + I\ldots$ really? Indeed they are: By definition,

$$(2+I)\times(3+I) = 6+I\,,$$

and $6 - 2 = 4 \in I$, so by Theorem 3.38, $6 + I = 2 + I$.

Given such strange results,[3] we should ask ourselves two important questions:

1. whether these operations are "well-defined"; and if so,
2. whether we really are in a ring.

We need to think about the first question because a coset can have multiple representations. For instance, in the example above, we only wrote four cosets, but as we pointed out near the end, $6 + I = 2 + I$. We run the risk, then, that if we happen to choose a different offset for a coset, we'll give a different answer — but that shouldn't happen.

This may be a bit confusing, so to illustrate it, we'll use the multiplicaiton above. We examined $(2 + I) \times (3 + I)$; that gave us $6 + I$. However, $2 + I = 6 + I$, so in fact we are claiming that

$$\underbrace{6}_{2\times 3} + I = (2+I)\times(3+I) = \underbrace{(6+I)}_{\text{substitution}} \times (3+I) = \underbrace{18}_{6\times 3} + I\,.$$

How can that possibly be true, since $6 \neq 18$?

In this case, it's true for the same reason that $2 + I = 6 + I$: coset equality (Theorem 3.38). If we look at the difference between the offsets, $18 - 6 = 12$, we see that it is an element of $I = \langle 4\rangle$.

Is this true in general? The following theorem shows that it is.

[3] This shouldn't bother you if you played around with $\mathbb{Z}_4$, since $2 \times 3 = 6 \equiv 2 \pmod 4$ there, so $2 \times 3 = 2$.

Theorem 3.47. *Coset addition and multiplication are well-defined.*

Proof. Let $X, Y \in R/I$. By definition, there exist $x, y \in R$ such that $X = x + I$ and $Y = y + I$. Suppose that we can also find $x', y' \in R$ such that $X = x' + I$ and $Y = y' + I$. By substitution, $x + I = x' + I$, and by coset equality, $x - x' \in I$. Similarly, $y - y' \in I$. Keep these in mind as we complete the proof.

Is coset addition well-defined? We need $(x + y) + I = (x' + y') + I$; that is, we have to have the same result, no matter which offsets we choose. This is true if and only if $(x + y) - (x' + y') \in I$. If we consider the expression, we can write

$$(x + y) - (x' + y') = (x + y) + [(-1)(x' + y')] = (x + y) + [(-x') + (-y')] \ .$$

(Notice our reliance here on Theorem 3.9.) By the associative and commutative properties, we can rewrite the right-hand side as

$$(x + y) - (x' + y') = [x + (-x')] + [y + (-y')] = (x - x') + (y - y') \ .$$

Recall that $x - x', y - y' \in I$. By Exercise 3.25, their sum is also in I; that is, $(x - x') + (y - y') \in I$. By substitution, $(x + y) - (x' + y') \in I$. By coset equality (Theorem 3.38), $(x + y) + I = (x' + y') + I$. Hence addition of cosets is well-defined.

Is coset multiplication well-defined? We need $xy + I = x'y' + I$; that is, we have to have the same result, no matter which offsets we choose. Unlike the previous paragraph, we will prove here that the two sets contain each other. So, let $z \in xy + I$; by definition, we can choose $\hat{i} \in I$ such that $z = xy + \hat{i}$. By hypothesis, $x + I = x' + I$, and by coset equality $x \in x + I$, so by substitution $x \in x' + I$. By definition, $x = x' + i'$ for some $i' \in I$. Similarly, $y = y' + \bar{i}$ for some $\bar{i} \in I$. By substitution and ring properties,

$$z = xy + \hat{i} = (x' + i')\left(y' + \bar{i}\right) + \hat{i} = x'y' + \left(i'y' + \bar{i}x' + i'\bar{i} + \hat{i}\right) \ .$$

Consider each term of the sum in parentheses: by absorption, $i'y', \bar{i}x', i'\bar{i} \in I$, and by Exercise 3.25, $i'y' + \bar{i}x' + i'\bar{i} + \hat{i} \in I$. By definition, then, $z \in x'y' + I$. Since z was arbitrary in $xy + I$, every element of $xy + I$ lies in $x'y' + I$, and by definition $xy + I \subseteq x'y' + I$. The proof that $x'y' + I \subseteq xy + I$ is similar, so the two must in fact be equal. □

Now we consider whether the operations mean that quotient R/I really is a ring.

Theorem 3.48. *Let R be a ring, and I an ideal. If we define addition and multiplication as above, then the quotient R/I satisfies the property of a ring.*

(Because of this, from now on we call R/I the ***quotient ring of** R **modulo** I*.)

Proof. We have to prove ten properties: 5 for addition, 4 for multiplication, and distribution. This is a lot, but they're relatively easy; you basically apply and re-apply definitions. We'll show four of them here, and leave the rest to the reader to complete in the exercises. For all the properties we show, let $X, Y, Z \in R/I$. By definition, we can find $x, y \in R$ such that $X = x + I$, $Y = y + I$, and $Z = z + I$.

Addition is closed: By definition of the operation, $X + Y = (x + y) + I$. By closure of addition in R, we know that $x + y \in R$. Elements of R/I consist of cosets of I whose offsets are in R. By definition, $(x + y) + I$ is a coset, so $(x + y) + I \in R/I$.

Addition is commutative: We leave this to Exercise 3.50.

Addition is associative: We need to show that $X + (Y + Z) = (X + Y) + Z$. Substituting the definition of coset addition for the inner sum, we have $X + (Y + Z) = X + [(y + z) + I]$ and $(X + Y) + Z = [(x + y) + I] + Z$. Substituting the definition of coset addition for the outer sum, we have $X + [(y + z) + I] = [x + (y + z)] + I$ and $[(x + y) + I] + Z = [(x + y) + z] + I$. Addition is associative in R, so we can rewrite the right-hand side of the first equation as $[x + (y + z)] + I = [(x + y) + z] + I$. This is the right-hand side of the second equation, so we can chain our equations together:

$$\begin{aligned} X + (Y + Z) &= X + [(y + z) + I] \\ &= [x + (y + z)] + I \\ &= [(x + y) + z] + I \\ &= [(x + y) + I] + Z \\ &= (X + Y) + Z\ . \end{aligned}$$

(Be sure you understand why each of those equalities is true!) We link the first and last expressions in this chain of equalities to conclude that $X + (Y + Z) = (X + Y) + Z$.

Addition has an identity: We need to find an element of R/I that acts as an additive identity. Since R/I's elements are cosets, we need to find a coset. The natural coset to consider is $0 + I$, or just plain I if you prefer (by Theorem 3.38), but we don't. By definition of coset addition and the additive identity property of R,

$$X + (0 + I) = (x + 0) + I = x + I\ ,$$

and similarly

$$(0 + I) + X = (0 + x) + I = x + I\ ,$$

so $0 + I$ is indeed an additive identity of R/I.

Addition is invertible: We leave this to Exercise 3.51.

Multiplication is closed: We leave this to Exercise 3.52.

Multiplication is commutative: We leave this to Exercise 3.53.

Multiplication is associative: We leave this to Exercise 3.54.

Multiplication has an identity: We leave this to Exercises 3.55.

Multiplication distributes over addition: We need to show that $X\,(Y + Z) = XY + XZ$.

We'll start with the left-hand side. By definition of coset addition, $X\,(Y + Z) = X\,[(y + z) + I]$. By definition of coset multiplication, $X\,[(y + z) + I] = [x\,(y + z)] + I$. By the distributive property of R's operations, $[x\,(y + z)] + I = (xy + xz) + I$.

Now we look at the right-hand side. By definition of coset multiplication, $XY + XZ = (xy + I) + (xz + I)$. By definition of coset addition, $(xy + I) + (xz + I) = (xy + xz) + I$.

The right-hand sides of the last equations in the previous two paragraphs are the same, so we can chain our equations together:

$$\begin{aligned}X(Y+Z) &= X[(y+z)+I]\\ &= [x(y+z)]+I\\ &= (xy+xz)+I\\ &= (xy+I)+(xz+I)\\ &= XY+XZ\ .\end{aligned}$$

(Be sure you understand why each of those equalities is true!) We link the first and last expressions in this chain of equalities to conclude that $X(Y+Z)=XY+XZ$. □

Exercises

Exercise 3.49. Let $R=\mathbb{Z}_{14}$ and $I=\langle 7\rangle$. Compute addition and multiplications tables for R/I. Look carefully at the multiplication table: is R/I an integral domain? What about a field?

Exercise 3.50. Show that addition in a quotient ring is commutative.

Exercise 3.51. Show that addition in a quotient ring is invertible.

Exercise 3.52. Show that multiplication in a quotient ring is closed.

Exercise 3.53. Show that multiplication in a quotient ring is commutative.

Exercise 3.54. Show that multiplication in a quotient ring is associative.

Exercise 3.55. Show that multiplication in a quotient ring has an identity.

3.5 Polynomial rings

So far we have worked with polynomials whose coefficients were integers or rational numbers: $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$. Those behaved like rings.

What if we wanted to study polynomials whose coefficients are real numbers? Or, what if their coefficients come from a ring like $\mathbb{Z}_m$? Organizing those polynomials in sets also gives us a ring.

Theorem 3.56. *Let R be any ring, and $R[x]$ the sets of all polynomials whose coefficients are elements of R. This, too, is a ring, where addition and multiplication are defined in a manner analagous to their definition in $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, etc.*

The proof is long and somewhat tedious, but there's nothing particularly *hard* about it. If you're like most of us, you could always use more practice reading mathematics, so take the time to read through each part of the proof. Make sure you understand each step. If you don't, *ask!* There's no shame in asking about something like this — though you should of course make sure that your question isn't answered simply by looking up a definition.

Proof. Let $f, g, h \in R[x]$. By definition, there exists $mr \in \mathbb{N}$ and $a_1, \ldots, a_m, b_1, \ldots, b_m, c_1, \ldots, c_m \in R$ such that $f = \sum_{i=0}^{m} a_i x^i$, $g = \sum_{i=0}^{m} b_i x^i$, and $h = \sum_{i=0}^{m} c_i x^i$. (It is possible that one or more of $\deg(f) < 0$, $\deg(g) < 0$, $\deg(h) < 0$.) We prove the ten properties of a ring.

Addition is closed: By definition, $f + g = \sum_{i=0}^{m} a_i x^i + \sum_{i=0}^{m} b_i x^i = \sum_{i=0}^{m} (a_i + b_i) x^i$. Addition is closed in R, so $a_i + b_i \in R$, and each term of $f + g$ has a coefficient in R. By definition, $f + g \in R[x]$.

Addition is commutative: As above, $f + g = \sum_{i=0}^{m} (a_i + b_i) x^i$. Addition is commutative in R, so $a_i + b_i = b_i + a_i$. By substitution, $f + g = \sum_{i=0}^{m} (b_i + a_i) x^i$. On the other hand, we know by definition that $g + f = \sum_{i=0}^{m} (b_i + a_i) x^i$. By substitution, $f + g = g + f$.

Addition is associative: We need to show that $(f + g) + h = f + (g + h)$. By definition, $f + g = \sum_{i=0}^{m} (a_i + b_i) x^i$, so by substitution, $(f + g) + h = \left[\sum_{i=0}^{m} (a_i + b_i) x^i\right] + h$. By definition,

$$(f + g) + h = \left[\sum_{i=0}^{m} (a_i + b_i) x^i\right] + h = \sum_{i=0}^{m} [(a_i + b_i) + c_i] x^i \ . \tag{3.7}$$

Similar reasoning gives us

$$f + (g + h) = f + \left[\sum_{i=0}^{m} (b_i + c_i) x^i\right] = \sum_{i=0}^{m} [a_i + (b_i + c_i)] x^i \ . \tag{3.8}$$

The right-hand sides of equations (3.7) and (3.7) look nearly identical. Fortunately, addition is associative in R, and this tells us that they *are* equal. By substitution, then, $(f + g) + h = f + (g + h)$.

Addition has an identity: We claim that 0, the zero polynomial (whose coefficients are all 0), is the identity of $R[x]$. To do that, we need to show that $f + 0 = f$ and $0 + f = f$ for all $f \in R[x]$. By definition,

$$f + 0 = \left(\sum_{i=0}^{m} a_i x^i\right) + \sum_{i=0}^{m} 0 \cdot x^i = \sum_{i=0}^{m} (a_i + 0) x^i \ .$$

Since 0 is an identity in R, we know that each $a_i + 0 = a_i$, so $f + 0 = f$. Similar reasoning shows that $0 + f = f$, so that the zero polynomial is an identity in $R[x]$.

Addition is invertible: We claim that the additive inverse of f is the polynomial whose coefficients are the opposites of f's coefficients. We will use $-f$ as a shorthand; that is,

$$-f = \sum_{i=0}^{m} (-a_i) x^i \ .$$

To show that it is indeed an inverse, we need to show that $f + (-f) = 0$ and $(-f) + f = 0$. By definition,

$$f + (-f) = \left(\sum_{i=0}^{m} a_i x^i\right) + \left(\sum_{i=0}^{m} (-a_i) x^i\right) = \sum_{i=0}^{m} [a_i + (-a_i)] x^i = \sum_{i=0}^{m} 0 \cdot x^i = 0 \ .$$

Similar reasoning shows that $(-f) + f = 0$.

Multiplication is closed: By definition,

$$fg = \left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{i=0}^{m} b_i x^i\right) = \sum_{i=0}^{2m} \left(\sum_{j+k=i} a_j b_k\right) x^i \ .$$

(Remember where we first addressed this in equation (2.1) on page 83.) Multiplication is closed in R, so each $a_j b_k \in R$. Addition is closed in R, so each sum $\sum_{j+k=i} a_j b_k \in R$, so that each term of fg has a coefficient in R. By definition, $fg \in R[x]$.

Multiplication is associative: By definition,

$$\begin{aligned}(fg)h &= \left[\left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{i=0}^{m} b_i x^i\right)\right]\left(\sum_{i=0}^{m} c_i x^i\right)\\ &= \left[\sum_{i=0}^{2m}\left(\sum_{j+k=i} a_j b_k\right)x^i\right]\left(\sum_{i=0}^{m} c_i x^i\right)\\ &= \sum_{i=0}^{3m}\left\{\sum_{\ell+\ell'=i}\left[\left(\sum_{j+k=\ell} a_j b_k\right)c_{\ell'}\right]\right\}x^i .\end{aligned}$$

Multiplication distributes over addition, so

$$(fg)h = \sum_{i=0}^{3m}\left\{\sum_{\ell+\ell'=i}\left[\sum_{j+k=\ell}(a_j b_k)\,c_{\ell'}\right]\right\}x^i .$$

We can combine the sum of sums as follows:

$$(fg)h = \sum_{i=0}^{3m}\left\{\sum_{(j+k)+\ell'=i}(a_j b_k)\,c_{\ell'}\right\}x^i . \tag{3.9}$$

Similar reasoning gives us,

$$f(gh) = \sum_{i=0}^{3m}\left\{\sum_{\ell+(j+k)=i} a_\ell\,(b_j c_k)\right\}x^i .$$

The trick here is to notice that we can rename ℓ, j, k in this last equation to j, k, ℓ', respectively, and the equation is still true. That is,

$$f(gh) = \sum_{i=0}^{3m}\left\{\sum_{j+(k+\ell')=i} a_j\,(b_k c_{\ell'})\right\}x^i .$$

This looks an awful lot like equation (3.9), except for the parentheses in the wrong place. But, multiplication is associative in R, so the two are in fact the same.

Multiplication has an identity: We claim that 1, the polynomial whose constant term is 1 and whose other coefficients are all 0, is the identity of $R[x]$. To see why, observe that

$$\begin{aligned}f \times 1 &= \left(\sum_{i=0}^{m} a_i x^i\right)(0 \cdot x^m + \cdots + 0 \cdot x + 1)\\ &= (a_m \cdot 0)\,x^{2m} + (a_m \cdot 0 + a_{m-1} \cdot 0)\,x^{2m-1} + \cdots + (a_1 \cdot 1 + a_0 \cdot 0)\,x + a_0 \cdot 1 ,\end{aligned}$$

which simplifies as $f \times 1 = f$. Similarly, $1 \times f = f$, so 1 is indeed the multiplicative identity of $R[x]$.

Multiplication distributes over addition: By definition,

$$\begin{aligned} f(g+h) &= \left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{i=0}^{m} b_i x^i + \sum_{i=0}^{m} c_i x^i\right) \\ &= \left(\sum_{i=0}^{m} a_i x^i\right)\left[\sum_{i=0}^{m} (b_i + c_i)\, x^i\right] \\ &= \sum_{i=0}^{2m}\left[\sum_{j+k=i} a_j (b_k + c_k)\right] x^i . \end{aligned}$$

Multiplication distributes over addition in R, so

$$f(g+h) = \sum_{i=0}^{2m}\left[\sum_{j+k=i} \left(a_j b_k + a_j c_k\right)\right] x^i . \tag{3.10}$$

On the other hand,

$$\begin{aligned} fg + fh &= \left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{i=0}^{m} b_i x^i\right) + \left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{i=0}^{m} c_i x^i\right) \\ &= \left[\sum_{i=0}^{2m}\left(\sum_{j+k=i} a_j b_k\right) x^i\right] + \left[\sum_{i=0}^{2m}\left(\sum_{j+k=i} a_j c_k\right) x^i\right] \\ &= \sum_{i=0}^{2m}\left[\sum_{j+k=i} \left(a_j b_k + a_j c_k\right)\right] x^i . \end{aligned}$$

The right hand side of this equation is identical to the right hand side of equation (3.10), so $f(g+h) = fg + fh$, as desired. □

Example 3.57. The ring $\mathbb{Z}_3[x]$ has 18 polynomials of degree 2:

$$\begin{array}{cccc} x^2 & 2x^2 & x^2 + x & 2x^2 + x \\ x^2 + 1 & 2x^2 + 1 & x^2 + x + 1 & 2x^2 + x + 1 \\ x^2 + 2 & 2x^2 + 2 & x^2 + x + 2 & 2x^2 + x + 2 \\ & & x^2 + 2x & 2x^2 + 2x \\ & & x^2 + 2x + 1 & 2x^2 + 2x + 1 \\ & & x^2 + 2x + 2 & 2x^2 + 2x + 2 \end{array}$$

Other degree-2 polynomials you might imagine simplify to one of these. For instance, $-17x^2 + 7$ simplifies to $x^2 + 2$, thanks to congruence modulo 3.

Polynomial division

You will recall that polynomial division in $\mathbb{Q}[x]$ worked out fine (Theorem 2.18) but not quite for $\mathbb{Z}[x]$ (Corollary 2.21). This was because division of rational numbers is a proper operation — two rationals in, one rational out — but division of integers is not a proper operation — two integers in, two integers out. The same parallel carries over for the general ring setting: if a ring has a proper division operation, then the proof of Theorem 2.18 carries over with only minor modifications.

But what sorts of rings have a "proper division operation"? Certainly any ring that behaves like $\mathbb{Q}$, where every nonzero element has a multiplicative inverse: fields, of course! For instance, we replace quotients of rational numbers by the product of a field element and its multiplicative inverse. Writing and proving a theorem for this becomes a mere exercise in copying Theorem 2.18 and replacing $\mathbb{Q}$ with a symbol for an arbitrary field.

Theorem 3.58 (The division theorem for polynomial rings over a field)**.** *Let $\mathbb{F}$ be a field, and $f, d \in \mathbb{F}[x]$ where $d \neq 0$. There exist $q, r \in \mathbb{F}[x]$ such that*

$$f = qd + r \quad \textit{and} \quad \textit{either } r = 0 \textit{ or } \deg(r) < \deg(d) .$$

In addition, q and r are uniquely determined by f and d.

Algorithm 3.1 on the following page, which is based on Algorithm 1.1 on page 16, produces the result we want.

Example 3.59. Suppose $\mathbb{F} = \mathbb{Z}_7$, which is a field by Theorem 3.4. We apply Algorithm 3.1 to $f = x^5 + 2x^2 + 1$ and $d = 2x^3 + x$.

- In step 1 we set $r = x^5 + 2x^2 + 1$ and $q = 0$.
- Since $r \neq 0$ and $\deg(r) = 5 > 3 = \deg(d)$, we perform step 2.
  - We set $t = 1 \times 2^{-1} \times x^{5-3} = 4x^2$.
  - Add t to q, resulting in $q = 4x^2$.
  - Subtract td from r, resulting in

    $$r = \left(x^5 + 2x^2 + 1\right) - 4x^2\left(2x^3 + x\right) = 3x^3 + 2x^2 + 1 .$$

    (Recall that $-4 = 3$ in $\mathbb{Z}_7$.)
- Since $r \neq 0$ and $\deg(r) = 3 = \deg(d)$, we perform step 2.
  - We set $t = 3 \times 2^{-1} \times x^{3-3} = 5$.
  - Add t to q, resulting in $q = 4x^2 + 5$.
  - Subtract td from r, resulting in

    $$r = \left(3x^3 + 2x^2 + 1\right) - (5)\left(2x^3 + x\right) = 2x^2 + 2x + 1 .$$

    (Recall that $-5 = 2$ in $\mathbb{Z}_7$.)

Algorithm 3.1 Polynomial division

```
inputs
  • $f, d \in \mathbb{F}[x]$, where $\mathbb{F}$ is a field
outputs
  • $q, d \in \mathbb{F}[x]$ such that
      – $f = qd + r$, and
      – either $r = 0$ or $\deg(r) < \deg(d)$
do
  1. let $r = f$, $q = 0$
  2. while $r \neq 0$ and $\deg(r) \geq \deg(d)$
      (a) let $t = \mathrm{lc}(r) \times \mathrm{lc}(d)^{-1} \times x^{\deg(r) - \deg(d)}$
      (b) add $t$ to $q$
      (c) subtract $td$ from $r$
  3. return $q$ and $r$
```

- At this point $r \neq 0$ but $\deg(r) = 2 < \deg(d)$, so we proceed to step 3 and return q and r.

It is easy to verify that

$$\begin{aligned} qd + r &= \left(4x^2 + 5\right)\left(2x^3 + x\right) + \left(2x^2 + 2x + 1\right) \\ &= \left(8x^5 + \cancelto{0}{14x^3} + 5x\right) + \left(2x^2 + 2x + 1\right) \\ &= x^5 + 2x^2 + \cancelto{0}{7x} + 1 \\ &= f\ . \end{aligned}$$

Proof of Theorem 3.58. If Algorithm 3.1 terminates correctly, the resulting q and r will satisfy Theorem 3.58, so we prove that Algorithm 3.1 terminates correctly.

Termination? If $f = 0$, then step 1 sets $q = 0$ and $r = f = 0$, so nothing happens at step 2, and step 3 returns $q = r = 0$, in which case

$$qd + r = 0 = f\ .$$

Not only has the algorithm terminated, we see that the output is correct.

Otherwise, $f \neq 0$. Step 1 sets $q = 0$ and $r = f$. If $\deg(f) < \deg(d)$, then nothing happens at step 2, and step 3 returns $q = 0$ and $r = f$, in which case

$$qd + r = f\ ,$$

and $\deg(r) = \deg(f) < \deg(d)$. Not only has the algorithm terminated, we see that the output is correct.

That leaves the case $f \neq 0$ and $\deg(f) \geq \deg(d)$. We claim that every time we perform step 2, the degree of r decreases. To see why, notice that we choose t such that, by substitution,

$$\begin{aligned} t \times \mathrm{lt}(d) &= \left[\mathrm{lc}(r) \times \mathrm{lc}(d)^{-1} \times x^{\deg(r)-\deg(d)}\right] \times \left[\mathrm{lc}(d) \times x^{\deg(d)}\right] \\ &= \mathrm{lc}(r) \times \left[\mathrm{lc}(d)^{-1} \times \mathrm{lc}(d)\right] \times x^{\deg(r)-\deg(d)+\deg(d)} \\ &= \mathrm{lt}(r) \ . \end{aligned}$$

Subtracting td from r thus cancels $\mathrm{lt}(r)$, leaving us with a polynomial of smaller degree.

Recall that the degree of a polynomial is a natural number. If we denote the degrees of r on each pass through the loop of step 2 as $n_0, n_1, \ldots$, then $n_0 > n_1 > \cdots$. This is a nonincreasing sequence of natural numbers. By Theorem 1.28, this sequence must eventually stabilize, so we cannot perform step 2 indefinitely. Eventually we must pass on to step 3, which terminates the algorithm.

Correctness? We have two things to prove: that $f = qd + r$, and that $r = 0$ or $\deg(r) < \deg(d)$. We consider the second one first.

- To show that $r = 0$ or $\deg(r) < \deg(d)$ we have two subcases.
  - If the returned value is $r = 0$, then we are done.
  - Otherwise, the condition on step 2 requires the algorithm to continue as long as $\deg(r) \geq \deg(d)$. We now know the algorithm terminates, so the loop cannot continue indefinitely, so the values returned in step 3 satisfy $\deg(r) < \deg(d)$.
- To show that $f = qd + r$ we again have two subcases.
  - If the algorithm does not perform step 2, then we saw already that $f = qd + r$.
  - Otherwise, enumerate each t computed in step 2(a) of the algorithm as $t_0, t_1, \ldots, t_{\mathrm{last}}$. The algorithm returns

    $$q = t_0 + t_1 + \cdots t_{\mathrm{last}} \quad \text{and} \quad r = f - t_1 d - t_2 d - \cdots - t_{\mathrm{last}} d \ .$$

    By substitution,

    $$\begin{aligned} qd + r &= (t_0 + \cdots + t_{\mathrm{last}})\, d + (f - t_1 d - \cdots - t_{\mathrm{last}} d) \\ &= (t_0 d + \cdots + t_{\mathrm{last}} d) + (f - t_1 d - \cdots - t_{\mathrm{last}} d) \\ &= f \ . \end{aligned}$$

We still have to show that q and r are unique. Suppose that in addition to q and r, we can find $\hat{q}, \hat{r} \in \mathbb{Q}[x]$ that satisfy the theorem. By substitution,

$$qd + r = \hat{q}d + \hat{r} \ .$$

Rewrite as

$$(q - \hat{q})\, d = \hat{r} - r \ .$$

By Theorem 2.4, either $\hat{r} - r = 0$ or $\deg(\hat{r} - r) \leq \max(\deg(\hat{r}), \deg(r)) < \deg d$. Similarly,[4] $q - \hat{q} = 0$ or $\deg((q - \hat{q})d) = \deg(q - \hat{q}) + \deg(d) \geq \deg(d)$. The degree of the left hand side cannot be smaller than the degree of the right hand side; they have to be equal. We conclude that $\hat{r} - r = 0$ and $q - \hat{q} = 0$; or, $r = \hat{r}$ and $q = \hat{q}$.

Regardless of the situation, the outputs of Algorithm 3.1 satisfy the stated requirements. The algorithm terminates correctly. As per the discussion at the beginning of the proof, this proves Theorem 3.58. □

A neat consequence of Theorem 3.58 is that polynomials over a field always form principal ideal rings.

Theorem 3.60. *If $\mathbb{F}$ be a field, then $\mathbb{F}[x]$ is a principal ideal ring.*

Proof. Let $\mathbb{F}$ be a field, and let I be any idea of $\mathbb{F}[x]$. Consider the following steps:

- Let $D = \{\deg(f) : f \in I \setminus \{0\}\}$; that is, D is the set of degrees of all nonzero polynomials.
- The degree of a polynomial is a natural number. By the Well-Ordering Principle, it has a minimum element; call that element d.
- Let $g \in I$ be any polynomial of degree d.

By absorption, any multiple of g is also in I, so $\langle g \rangle \subseteq I$.

We claim that $I = \langle g \rangle$. To see why, let $f \in I$, and apply Theorem 3.58 to compute $q, d \in \mathbb{F}[x]$ such that $f = qg + r$ and either $r = 0$ or $\deg(r) < \deg(f)$. Rewrite $f = qg + r$ as $r = f - qg$. By absorption, $qg \in I$, and by closure of subtraction, $f - qg \in I$, so in fact $r \in I$. If $r = 0$, then $f = qg$, which implies $f \in \langle g \rangle$, and we are fine. Otherwise, $\deg(r) < \deg(g)$, so we have a found in I a nonzero polynomial, r, whose degree is smaller than the degree of g.

This contradicts the construction of g, whose degree is minimal! The only assumption we made that was not well-founded was the assumption that $r \neq 0$. It follows that $r = 0$. As we wrote above, this implies $f \in \langle g \rangle$. Since f was arbitrary in I, we see that $I \subseteq \langle g \rangle$. We already showed that $\langle g \rangle \subseteq I$, so we must actually have $\langle g \rangle = I$. □

The proof of Theorem 3.60 uses a neat trick that is important enough to highlight.[5]

Lemma 3.61. *Let $\mathbb{F}$ be a field, I an ideal of $\mathbb{F}[x]$, and g the generator of I. Let $f \in \mathbb{F}[x]$, and r the remainder of dividing f by g. Then $f \in I$ if and only if $r \in I$.*

Proof. Let q be the quotient associated with dividing f by g, so that $f = qg + r$. By absorption, $qg \in I$. If $r \in I$, then Exercise 3.25 tells us that $qg + r \in I$, and by substitution, $f \in I$ as well. Conversely, if $f \in I$, then by closure of subtraction, $f - qg \in I$, and by substitution, $r \in I$ as well. □

[4] Once again, the zero product property has an implied role here; see if you can spot it!

[5] In fact, this lemma is an very important tool in higher algebra.

Quotient rings from polynomial rings

Once we have polynomial rings, we can make ideals from them, and thus quotient rings. We will show that the remainders from polynomial division make it easy to think about quotient rings — just as it helps with congruence.

Example 3.62. Let $R = \mathbb{Z}_3[x]$ and $I = \langle x^2+1 \rangle$. The notation "$\mathbb{Z}_3[x]$" means the set of polynomials whose coefficients are elements of $\mathbb{Z}_3$; by Theorem 3.56, this set is also a ring.

Consider two elements of R/I. They have the form $f+I$ and $g+I$, where f and g are polynomials whose coefficients are elements of $\mathbb{Z}_3$. By Theorem 3.38, $f+I = g+I$ if and only if $f-g \in I$. But how do we decide whether $f-g \in I$? By Lemma 3.61, $f-g \in I$ if and only if its remainder after division by x^2+1 is 0.

We can also take a different route. Rather than divide $f-g$, let's divide each of f and g by x^2+1, finding quotients q_f, q_g and remainders r_f, r_g. If $f+I = g+I$, then $f-g \in I$, and by substitution,

$$f-g = \left[q_f\left(x^2+1\right)+r_f\right]-\left[q_g\left(x^2+1\right)+r_g\right] = \left(q_f-q_g\right)\left(x^2+1\right)+\left(r_f-r_g\right) .$$

Rewrite this as

$$r_f-r_g = (f-g)-\left(q_f-q_g\right)\left(x^2+1\right) .$$

By absorption and closure of subtraction, the right-hand side is in I, and that forces the left-hand side to be in I. If $r_f - r_g = 0$, then the remainders of dividing f and g by x^2+1 are the same. Otherwise

$$\deg\left(r_f-r_g\right) \quad \leq \quad \max\left\{\deg\left(r_f\right), \deg\left(r_g\right)\right\} \quad < \quad \deg\left(x^2+1\right) ,$$

but this cannot be, because $r_f - r_g \in I$, and $\left(x^2+1\right) \nmid \left(r_f-r_g\right)$ if they are nonzero and of smaller degree. Hence $r_f - r_g = 0$.

In other words, we can find whether two cosets are the same by dividing their offsets by x^2+1. For instance, suppose

$$\begin{aligned} p &= x^7+x^5+4x^3+2x^2+5x+6 , \\ q &= x^5+2x^3+x^2+2x+5 , \text{ and} \\ r &= x^7+x^5+3x^2+3x+5 . \end{aligned}$$

After dividing by x^2+1, we find that

$$p = (x+4)+I \quad , \quad q = (x+4)+I \quad , \text{ and} \quad r = (3x+2)+I .$$

Hence, $p+I = q+I$ and $p \equiv q \pmod{I}$, while $p+I \neq r+I$ and $p \not\equiv q \pmod{I}$.

This holds true in general.

Theorem 3.63 (Coset equality for $\mathbb{F}[x]$)**.** *Let $\mathbb{F}$ be a field, and I an ideal of $\mathbb{F}[x]$. Let g be the generator of I, and let $f, h \in \mathbb{F}[x]$. The cosets $f+I$ and $h+I$ are identical if and only if the remainders of dividing f and h by g are identical.*

Proof. The proof is nearly identical to the discussion in Example 3.62, so be sure you understand that first.

Throughout the proof, let $q_f, r_f \in \mathbb{F}[x]$ and $q_h, r_h \in \mathbb{F}[x]$ be the quotients and remainders of dividing f and h by g, respectively.

Assume that $h+I$ and $h+I$ are identical. By coset equality, $f-h \in I$. By Theorem 3.60, $\mathbb{F}[x]$ is a principal ideal ring, so $I = \langle g \rangle$, so $f-h$ is a multiple of g. Let $q_{f-h}, r_{f-h} \in \mathbb{F}[x]$ be the quotient and remainder of dividing $f-h$ by g. We just said that $f-h$ is a multiple of g, so $r_{f-h} = 0$; that is, $f-h = q_{f-h}g$. By substitution,

$$f-h = \left(q_f g + r_f\right) - \left(q_h g + r_h\right) = \left(q_f - q_h\right) g + \left(r_f - r_h\right) .$$

Also by substitution,

$$q_{f-h} g = \left(q_f - q_h\right) g + \left(r_f - r_h\right) .$$

Rewrite this as

$$\left[q_{f-h} - \left(q_f - q_h\right)\right] g = r_f - r_h .$$

If $r_f - r_h \neq 0$, then the left-hand side of the equation is not 0, so the two sides have equal degree. However,

$$\begin{aligned} \deg\left(\left[q_{f-h} - \left(q_f - q_h\right)\right] g\right) &= \deg\left[q_{f-h} - \left(q_f - q_h\right)\right] + \deg(g) \\ &\geq \deg(g) \\ &> \max\left\{\deg\left(r_f\right), \deg\left(r_h\right)\right\} \\ &= \deg\left(r_f - r_h\right) , \end{aligned}$$

so the two sides must have different degrees, a contradiction. Hence $r_f - r_h = 0$, and the remainders from dividing f and h by g are the same.

Conversely, assume that the remainders of dividing f and h by g are the same. We can rewrite the equation $f = q_f g + r_f$ as $f - r_f = q_f g$. By absorption, $q_f g \in I$, so by substitution, $f - r_f \in I$, and by coset equality, $f + I = r_f + I$. Similarly, $h + I = r_h + I$. By substitution, $r_f + I = r_h + I$. Also by substitution, $f + I = h + I$, as claimed. □

Exercises

Exercise 3.64 (The Freshman's Dream)**.** Show that $(x+1)^2 = x^2 + 1$ if we perform the arithmetic in $\mathbb{Z}_2[x]$.
Remark: The name of this exercise is inspired by the unfortunate phenomenon where freshmen who supposedly know algebra think it's always true.

Exercise 3.65. List all the degree-3 polynomials of $\mathbb{Z}_2[x]$. (There are 8 of them.)

Exercise 3.66. There are exactly 4 elements of $\mathbb{Z}_2[x] / \left\langle x^2 + x + 1 \right\rangle$. List them all.

Exercise 3.67. In Exercise 3.66, you found four elements of $\mathbb{Z}_2[x] / \left\langle x^2 + x + 1 \right\rangle$. Construct addition and multiplication tables for this ring.

Exercise 3.68. Which of the following polynomials are congruent modulo $\langle 3x + 2 \rangle$ in $\mathbb{Z}_5[x]$, if any?

- $3x^7 + 3x^5 + 4x^3 + 2x^2 + 4$
- $3x^5 + 4x^4 + 4x^3 + 6x^2 + 4x + 4$
- $4x^9 + 4x^7 + 2x^6 + 4x^4 + 5x^2 + 3x + 5$

Exercise 3.69. The following exercises illustrate how polynomial roots behave differently when the field is different.

(a) Find a root of $x^2 + 1$ in $\mathbb{Z}_2[x]$.

(b) Show that $x^2 + x + 1$ does not have a root in $\mathbb{Z}_2[x]$.

(c) In Exercise 3.67, you constructed addition and multiplication tables for $\mathbb{Z}[x] / \langle x^2 + x + 1 \rangle$. Use this table to show that $\mathbb{Z}[x] / \langle x^2 + x + 1 \rangle$ contains a root of $x^2 + x + 1$.

Exercise 3.70. Show that the Factor Theorem (Exercise 2.25) remains true regardless of the underlying field. That is, let $\mathbb{F}$ be any field. Show that if $f \in \mathbb{F}[x]$ and $s \in \mathbb{F}$ is a root of f, then $x - s$ is a factor of f.
Hint: The proof should be more or less identical.

Exercise 3.71. Show that the Remainder Theorem (Exercise 2.26) remains true regardless of the underlying field. That is, let $\mathbb{F}$ be any field. Show that if $f \in \mathbb{F}[x]$ and $s \in \mathbb{F}$, then the remainder of dividing f by $x - s$ is $f(s)$.
Hint: The proof should be more or less identical.

Exercise 3.72. Prove the Euclidean Algorithm works in a polynomial ring $\mathbb{F}[x]$, where $\mathbb{F}$ is any field.
Hint: Follow the proof of Theorem 2.34, changing only what you absolutely have to change.

Exercise 3.73. Show that if $\mathbb{F}$ is a field, then $\mathbb{F}[x]$ is a principal ideal ring.
Hint: For all practical purposes, this is identical to Exercise 3.28; just change what needs changing.

Exercise 3.74. Polynomial division does not work when the coefficients come from an arbitrary ring. You've already seen this with $\mathbb{Z}[x]$. However, we were able to adapt division to work in $\mathbb{Z}[x]$; see Corollary 2.21. This does not work in an arbitrary ring. Even though $\mathbb{Z}$ is a not a field, it still satisfies an important property which all fields satisfy, but arbitrary rings do not. What is it, and why does it matter?
Hint: There is more than one way to answer this. For a more specific hint, try to divide $2x + 2$ by itself in $\mathbb{Z}_6$. — More specifically, see if you can find more than one quotient. Once you can do that, ask yourself why. It might help to review the introduction to $\mathbb{Z}_m$.

3.6 Isomorphism

In Examples 3.44 and 3.45 we saw that some quotient rings have addition and multiplication tables that look exactly like the addition and multiplication tables of rings we're more familiar with. In that case, $\mathbb{Z}/\langle 4\rangle$ looked like $\mathbb{Z}_4$,

$$\underbrace{\begin{array}{c|cccc} + & I & 1+I & 2+I & 3+I \\ \hline I & I & 1+I & 2+I & 3+I \\ 1+I & 1+I & 2+I & 3+I & I \\ 2+I & 2+I & 3+I & I & 1+I \\ 3+I & 3+I & I & 1+I & 2+I \end{array} \qquad \begin{array}{c|cccc} \times & I & 1+I & 2+I & 3+I \\ \hline I & I & I & I & I \\ 1+I & I & 1+I & 2+I & 3+I \\ 2+I & I & 2+I & I & 2+I \\ 3+I & I & 3+I & 2+I & 1+I \end{array}}_{\mathbb{Z}/\langle 4\rangle}$$

$$\overbrace{\begin{array}{c|cccc} + & I & 1+I & 2+I & 3+I \\ \hline I & I & 1+I & 2+I & 3+I \\ 1+I & 1+I & 2+I & 3+I & I \\ 2+I & 2+I & 3+I & I & 1+I \\ 3+I & 3+I & I & 1+I & 2+I \end{array} \qquad \begin{array}{c|cccc} \times & I & 1+I & 2+I & 3+I \\ \hline I & I & I & I & I \\ 1+I & I & 1+I & 2+I & 3+I \\ 2+I & I & 2+I & I & 2+I \\ 3+I & I & 3+I & 2+I & 1+I \end{array}}^{\mathbb{Z}_4}$$

...and $\mathbb{Z}[x]/\langle 2,x\rangle$ looked like $\mathbb{Z}_2$,

$$\underbrace{\begin{array}{c|cc} + & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 1 & 0 \end{array} \qquad \begin{array}{c|cc} \times & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \end{array}}_{\mathbb{Z}_2} \Leftrightarrow \underbrace{\begin{array}{c|cc} + & I & 1+I \\ \hline I & I & 1+I \\ 1+I & 1+I & I \end{array} \qquad \begin{array}{c|cc} \times & I & 1+I \\ \hline I & I & I \\ 1+I & I & 1+I \end{array}}_{\mathbb{Z}[x]/\langle 2,x\rangle} .$$

They have the same number of elements, and their addition and multiplication tables have the same shape. Mathematicians call this property "isomorphism", from the Greek words for "identical form." We use the ≅ symbol as a shorthand, which means that we can write

$$\mathbb{Z}/\langle 4\rangle \cong \mathbb{Z}_4 \quad \text{and} \quad \mathbb{Z}[x]/\langle 2,x\rangle \cong \mathbb{Z}_2 .$$

It's not a challenge to compare the addition and multiplication tables of two small rings like $\mathbb{Z}_4$ and $\mathbb{Z}/\langle 4\rangle$. However, this is infeasible when the rings are large, and impossible when they are infinite. We need another technique to determine whether rings like this are isomorphic.

The isomorphism function

An important tool that mathematicians use to study infinite sets is the ***function***. Way back in Section 1.2 we stated that a function f from a set S to a set T is a subset of $S \times T$ such that if $(s,t),(s,u) \in f$, then $t = u$. A more traditional way of writing this is that $f(s) = t$ and $f(s) = u$ only if $t = u$. In "plain English," we're saying that we can always predict the result of applying f to an object.

For example, you have probably seen the function $f(x) = x^2$. Regardless of the value of x we start with, f will give us *only* one result: $(-2,4)$, $(1,1)$, $(2,4)$, and so forth. Contrast this to the relation $y^2 = x$. For every positive value of x, the relation gives us two y-values: $(2,4)$ and $(-2,4)$, $(1,1)$ and $(-1,1)$, and so forth. This behavior is non-deterministic, and we don't generally like it.

Functions by themselves aren't enough to define isomorphism, but we can use them to check for isomorphism in the following way. Let R and S be two rings that we want to check for isomorphism. First determine a function f that maps from R to S.

1. Check whether the sets have the same size:

   (a) f doesn't confuse any two elements in R with the same element in S; the technical phrase is that f is ***one-to-one***. Distinct inputs have distinct outputs; put another way, if two things seem to give you the same result, they must be the same thing.

   In symbols, write this as follows: for every $a, b \in R$, if $f(a) = f(b)$, then $a = b$. (We call a the ***preimage*** of b.)

   (b) f misses nothing in S; the technical phrase is that f maps ***onto*** S. In symbols, we write this as follows: for every $c \in S$, we can find $a \in R$ such that $f(a) = c$.

2. Check the corresponding elements add or multiply to corresponding elements. In symbols, we write this as follows:

$$\underbrace{f(a+b)}_{\text{what } a+b \text{ corresponds to}} = \underbrace{f(a)}_{\text{what } a \text{ corresponds to}} + \underbrace{f(b)}_{\text{what } b \text{ corresponds to}}$$

   and

$$\underbrace{f(ab)}_{\text{what } ab \text{ corresponds to}} = \underbrace{f(a)}_{\text{what } a \text{ corresponds to}} \cdot \underbrace{f(b)}_{\text{what } b \text{ corresponds to}} .$$

All told, we have four things to check.

Example 3.75. Is $\mathbb{Z}_4 \cong \mathbb{Z}/\langle 4\rangle$? We already know that it is, but let's try using this four-step definition. First we need a function that maps from $\mathbb{Z}_4$ to $\mathbb{Z}/\langle 4\rangle$. Let's try this one:

$$f : \mathbb{Z}_4 \longrightarrow \mathbb{Z}/\langle 4\rangle \quad \text{by} \quad f(a) = a + \langle 4\rangle \ .$$

one-to-one? Let $a, b \in \mathbb{Z}_4$, and assume that $f(a) = f(b)$. By substitution, $a + \langle 4\rangle = b + \langle 4\rangle$. By coset equality, $a - b \in \langle 4\rangle$. By definition of a principal ideal, $4 \mid (a - b)$. By definition of congruence, $a \equiv b \pmod 4$. By definition of $\mathbb{Z}_4$, $a = b$.

onto? Let A be any coset of $\mathbb{Z}/\langle 4\rangle$. We saw in Example 3.35 that $A = a + \langle 4\rangle$ for some $a \in \{0, 1, 2, 3\}$. By definition of f, we see that $f(a) = a + \langle 4\rangle$. Hence every element of $\mathbb{Z}[x]/\langle x+1\rangle$ has a preimage in $\mathbb{Z}$.

preserves addition? Let $a, b \in \mathbb{Z}_4$. By definition of $\mathbb{Z}_4$, $a + b = c$ where c is the remainder of dividing $a + b$ by 4. Hence, we need $f(c) = f(a+b) = f(a) + f(b)$. By definition of the function, $f(a) = a + \langle 4\rangle$, $f(b) = b + \langle 4\rangle$, and $f(c) = c + \langle 4\rangle$. By definition of coset arithmetic, $f(a) + f(b) = (a+b) + \langle 4\rangle$. To finish the proof, we work backwards. We need to show that $(a+b) + \langle 4\rangle = c + \langle 4\rangle$. This is true only if $[(a+b) - c] \in \langle 4\rangle$, which is true only if $4 \mid [(a+b) - c]$; and this is true only if $a + b \equiv c \pmod 4$, which is true only if $a+b$ and c have the same remainder after dividing by 4 — but we defined c as the remainder of dividing $a + b$ by 4, so we are done.

preserves multiplication? Let $a, b \in \mathbb{Z}_4$. By definition of $\mathbb{Z}_4$, $ab = c$ where c is the remainder of dividing ab by 4. Hence, we need $f(c) = f(ab) = f(a) f(b)$. By definition of the function, $f(a) = a + \langle 4\rangle$, $f(b) = b + \langle 4\rangle$, and $f(c) = c + \langle 4\rangle$. By definition of coset arithmetic, $f(a) f(b) = (ab) + \langle 4\rangle$. To finish the proof, we work backwards. We need to show that $(ab) + \langle 4\rangle = c + \langle 4\rangle$. This is true only if $(ab - c) \in \langle 4\rangle$, which is true only if $4 \mid (ab - c)$; and this is true only if $ab \equiv c \pmod 4$, which is true only if ab and c have the same remainder after dividing by 4 — but we defined c as the remainder of dividing ab by 4, so we are done.

The next example recalls a different similarity.

Example 3.76. In Example 3.36, we determined that the cosets of $\mathbb{Z}[x]/\langle x+1\rangle$ have the form

$$\{\ldots, -1+\langle x+1\rangle, \langle x+1\rangle, 1+\langle x+1\rangle, 2+\langle x+1\rangle, \ldots\},$$

and we remarked that it looked an awful lot like $\mathbb{Z}$. Unfortunately, we cannot build comprehensive addition and multiplication tables for an infinite set. *Fortunately*, we now have the isomorphism tool to help us out. First, we need a function that maps from $\mathbb{Z}$ to $\mathbb{Z}[x]/\langle x+1\rangle$. Let's try this one:

$$f:\mathbb{Z}\to\mathbb{Z}[x]/\langle x+1\rangle \quad \text{by} \quad f(a)=a+\langle x+1\rangle .$$

one-to-one? Let $a,b\in\mathbb{Z}$, and assume that $f(a)=f(b)$. By substitution, $a+\langle x+1\rangle = b+\langle x+1\rangle$. By coset equality, $a-b\in\langle x+1\rangle$. By definition of a principal ideal, $(x+1)\mid(a-b)$. If $a,b\in\mathbb{Z}$, they are constant scalars, and $\deg(a)=\deg(b)=0$, so $\deg(a-b)=0$, whereas $\deg(x+1)=1$. A polynomial like $x+1$ can only divide a smaller-degree polynomial like $a-b$ if the second polynomial is zero. That is, $a-b=0$, or, $a=b$.

onto? Let A be any coset of $\mathbb{Z}[x]/\langle x+1\rangle$. From Example 3.36 we know that $A=a+\langle x+1\rangle$ for some $a\in\mathbb{Z}$. By definition of f, $f(a)=a+\langle x+1\rangle$. Hence every element of $\mathbb{Z}[x]/\langle x+1\rangle$ has a preimage in $\mathbb{Z}$.

preserves addition? Let $a,b\in\mathbb{Z}$. We need $f(a+b)=f(a)+f(b)$. By definition, $f(a)=a+\langle x+1\rangle$, $f(b)=b+\langle x+1\rangle$, and $f(a+b)=(a+b)+\langle x+1\rangle$. By definition of coset arithmetic, $f(a)+f(b)=(a+b)+\langle x+1\rangle$. By substitution, $f(a+b)=f(a)+f(b)$.

preserves multiplication? Let $a,b\in\mathbb{Z}$. We need $f(ab)=f(a)\cdot f(b)$. By definition, $f(a)=a+\langle x+1\rangle$, $f(b)=b+\langle x+1\rangle$, and $f(ab)=ab+\langle x+1\rangle$. By definition of coset arithmetic, $f(a)\cdot f(b)=ab+\langle x+1\rangle$. By substitution, $f(ab)=f(a)\cdot f(b)$.

Is the isomorphism relation like equality or congruence?

Now that we have shown that $\mathbb{Z}_4\cong\mathbb{Z}/\langle 4\rangle$, you might be tempted to write $\mathbb{Z}/\langle 4\rangle\cong\mathbb{Z}_4$. Unfortunately, we can't do that quite yet. The definition of isomorphism requires that we identify a function that maps from the first ring to the second, and the function f that we found in Example 3.75 maps from $\mathbb{Z}_4$ to $\mathbb{Z}/\langle 4\rangle$, not the other way around! In this case, it's not too hard to find an isomorphism from $\mathbb{Z}/\langle 4\rangle$ to $\mathbb{Z}_4$, but is that true in general?

Here's something else you might like to do. Suppose that R,S,T are all rings, and you know that $R\cong S$ and $S\cong T$. You may be tempted to write $R\cong T$; after all, if R "has the same shape" as S, and S "has the same shape" as T, shouldn't R "have the same shape" as T? Intuitively, this makes sense, but it can happen in mathematics that things that *look* intuitive don't pan out when we investigate them further. (Recall, for instance, Exercise 3.64, "The Freshman's Dream.")

What we're asking is whether the isomorphism relation behaves like equality, or, whether the isomorphism relation is an equivalence relation. We talked about that back in Section 1.1; then revisited it for Theorem 1.86 in Section 1.5. If you forgot it, it might be advisable to review the specifics of the definition, but the ideas are all in the theorem below.

Theorem 3.77. *Isomorphism is an equivalence relation.*

Proof. We have to show that isomorphism is reflexive, symmetric, and transitive. Let R be a ring.

reflexive? We need to show that $R \cong R$. First we need a function from R to itself. The most obvious candidate is the identity function, $\iota : R \longrightarrow R$ by $\iota(r) = r$. If this is an isomorphism, then we are done. But is it? Let's check the properties.

To show that ι is one-to-one, let $a, b \in R$ and assume that $\iota(a) = \iota(b)$. By definition of ι, $\iota(a) = a$ and $\iota(b) = b$. By substitution, $a = b$, so ι has not confused two different inputs for the same output, so ι is one-to-one.

To show that ι is onto, let $b \in R$. We need to find $a \in R$ such that $\iota(a) = b$. By definition, $\iota(a) = a$, so if we choose $a = b$ then we have what we need. Every element of R has a preimage in R under ι, so ι is onto.

To show that ι preserves the operations, let $a, b \in R$. By definition, $\iota(a) = a$, $\iota(b) = b$, $\iota(a+b) = a+b$, and $\iota(ab) = ab$. By substitution, $\iota(a+b) = \iota(a) + \iota(b)$ and $\iota(ab) = \iota(a)\iota(b)$; that is, ι preserves the operations. We already showed that ι is one-to-one and onto, so ι is an isomorphism, and $R \cong R$.

symmetric? Let S be a ring, and assume that $R \cong S$. We need to show that $S \cong R$, also. For that, we need to find an isomorphism that maps from S to R. Why not use information we already have? We already know that there's a isomorphism f that maps from R to S. If we could reverse f and preserve the isomorphism property, then we'd be done. But how do we "reverse" a function f? If it has an inverse function, that will do the trick. *Does* f have an inverse function? After all, only one-to-one funtions have inverse functions. Recall that the definition of an isomorphism like f is that it must be one-to-one: so, *yes*, f has an inverse function, f^{-1}, and it maps from S to R, just as we need. It remains to show that f^{-1} satisfies the isomorphism properties.

To show that f^{-1} is one-to-one, let $c, d \in S$ and assume that $f^{-1}(c) = f^{-1}(d)$. By definition of an inverse function, there exist $a, b \in R$ such that $f(a) = c$, $f(b) = d$, and thus $a = f^{-1}(c)$ and $b = f^{-1}(d)$. By substitution, $a = b$. But f is a function, so $f(a) = f(b)$, which implies that $c = d$. The inverse function f^{-1} has not confused two different inputs for the same output, so f^{-1} is one-to-one.

To show that f is onto, let $a \in R$. We need to find $c \in S$ such that $f^{-1}(c) = a$. Let $c = f(a)$; by definition of an inverse function, $f^{-1}(c) = a$. Every element of R has a preimage in S under f^{-1}, so f^{-1} is onto.

To show that f^{-1} preserves the operations, let $c, d \in S$. We need to show that $f^{-1}(c+d) = f^{-1}(c) + f^{-1}(d)$ and $f^{-1}(cd) = f^{-1}(c) f^{-1}(d)$. To do this, let $a = f^{-1}(c)$ and $b = f^{-1}(d)$. By definition of an inverse function, $f(a) = c$ and $f(b) = d$. Recall that f is an isomorphism, so $f(a+b) = f(a) + f(b)$ and $f(ab) = f(a) \cdot f(b)$. By substitution, $f(a+b) = c+d$ and $f(ab) = cd$. By definition of an inverse function, $a + b = f^{-1}(c+d)$ and $ab = f^{-1}(cd)$. By a chain of subsitutions, then,

$$f^{-1}(c+d) = a+b = f^{-1}(c) + f^{-1}(d) \quad \text{and} \quad f^{-1}(cd) = ab = f^{-1}(c) f^{-1}(d) \ ;$$

that is, f^{-1} preserves the operations. We already showed that f^{-1} is one-to-one and onto, so f^{-1} is an isomorphism.

transitive? We leave this to Exercise 3.85.

We have show that isomorphism is reflexive, symmetric, and transitive. It satisfies the requirements of an equivalence relation. □

The Isomorphism Theorem

Finding an isomorphism from R to S can be difficult. In fact, it might well be that R is not in fact isomorphic to S. Amazingly, we can often use a ring that is not isomorphic to another to find a ring that is.

Example 3.78. $\mathbb{Z}$ is not isomorphic to $\mathbb{Z}_4$, as the former is infinite, while the latter is very finite. Nevertheless, there is a similarity between their operations. For instance, $2 + 3 = 5$ in $\mathbb{Z}$, and while $5 \notin \mathbb{Z}_4$, it is the case that $2 + 3 = 1$ in $\mathbb{Z}_4$, and $5 \equiv 1 \pmod 4$.

By now you know that congruence is only a hop, skip, and a jump away from coset equality: in this case,we can translate the action above as

$$(2 + \langle 4\rangle) + (3 + \langle 4\rangle) \quad = \quad 5 + \langle 4\rangle \quad = \quad 1 + \langle 4\rangle \ .$$

As it happens, we found that the quotient ring $\mathbb{Z}/\langle 4\rangle$ turned out isomorphic to $\mathbb{Z}_4$.

One of the key requirements of our theorem will be that the function preserve the operations; this is so important that we give this property a name, ***homomorphism***. With this in hand, we can give isomorphism a simpler definition: an isomorphism is a homomorphism that is one-to-one and onto.

Theorem 3.79 (The Isomorphism Theorem)**.** *Let R and S be rings, and suppose that we can find a homomorphism $f : R \to S$ that maps onto S, but might not satisfy the one-to-one property.*

(A) The set $K = \{r \in R : f(r) = 0\}$ is an ideal.

(B) The quotient ring R/K is isomorphic to S. In symbols, $R/K \cong S$.

(C) We can find an onto homomorphism $\nu : R \to R/K$ and an isomorphism μ from R/K to S such that f is the composition of μ with ν; that is, $f = \mu \circ \nu$.

We call K the ***kernel*** of a homorphism. It's an important object of study in higher algebra. Meanwhile, the relationship between f, ν, and μ satisfies the following diagram:

$$\begin{tikzcd}
R \arrow[rr, "f"] \arrow[dr, "\nu"'] & & S \\
& R/K \arrow[ur, "\mu"'] &
\end{tikzcd}$$

The idea is that if you pick an element $r \in R$, you get the same $s \in S$ regardless of the arrows you choose to take: both $f(r) = s$ and $(\mu \circ \nu)(r) = \mu(\nu(r)) = s$. Mathematicians would say that "this diagram commutes."

Example 3.80. Before we prove the theorem, let's look at an example of how it can work for us. Example 3.76 went through a lot of work to show that $\mathbb{Z}[x]/\langle x+1\rangle \cong \mathbb{Z}$. Let's try a different approach.

Let $\varphi : \mathbb{Z}[x] \to \mathbb{Z}$ by

$$\varphi(a_n x^n + \cdots + a_1 x + a_0) = a_0 \ ;$$

that is, φ keeps only the the constant term of a polynomial. This is an onto homomorphism:

onto? Let $a \in \mathbb{Z}$. Certainly $a \in \mathbb{Z}[x]$, as well. By definition, $\varphi(a) = a$.

preserves addition? Let $f, g \in \mathbb{Z}[x]$, and choose $a_i, b_i \in \mathbb{Z}$ such that $f = a_m x^m + \cdots + a_1 x + a_0$ and $g = b_m x^m + \cdots + b_1 x + b_0$. By polynomial addition and definition of f, we know that

$$\varphi(f+g) = \varphi((a_m + b_m)x^m + \cdots + (a_1 + b_1)x + (a_0 + b_0)) = a_0 + b_0 \,,$$

while

$$\varphi(f) + \varphi(g) = \varphi(a_m x^m + \cdots + a_1 x + a_0) + \varphi(b_m x^m + \cdots + b_1 x + b_0) = a_0 + b_0 \,,$$

and by substitution, $\varphi(f+g) = \varphi(f) + \varphi(g)$.

preserves multiplication? Similar to *preserves addition;* see Exercise 3.87.

The hypothesis of the Isomorphism Theorem is satisfied! So, what is φ's kernel? Let f be in the kernel of φ; by definition, $\varphi(f)$ is f's constant term, and also by definition, $\varphi(f) = 0$, so by substitution, f's constant term is 0. In other words, f is a polynomial with no constant term. By Exercise , $f \in \langle x \rangle$, so the kernel of φ is $\langle x \rangle$.

By the Isomorphism Theorem, then, $\mathbb{Z}[x]/\langle x \rangle \cong \mathbb{Z}$.

Proof of Theorem 3.79. We leave (A) to the exercises, and show (C) before (B). We actually get (B) for free, since the isomorphism μ of (C) shows that $R/K \cong S$.

Define $\nu : R \to R/K$ by $\nu(r) = r + K$. We claim that ν is a homomorphism, and that ν is onto R/K. To see that ν is a homomorphism, let $a, b \in R$. We need to show that $\nu(a+b) = \nu(a) + \nu(b)$. By definition, $\nu(a) = a + K$, $\nu(b) = b + K$, and $\nu(a+b) = (a+b) + K$. By the definition of coset addition, $(a+K) + (b+K) = (a+b) + K$. By substitution, $\nu(a) + \nu(b) = \nu(a+b)$, so ν is indeed a homomorphism. To see that ν is onto, let $A \in R/K$. By definition of coset, there exists $a \in R$ such that $A = a + K$. By definition, $\nu(a) = a + K = A$, so every coset of R/K has a preimage in R, and ν is thus an onto homomorphism.

Define $\mu : R/K \to S$ in the following way: for any $A \in R/K$, choose an offset $a \in R$ such that $A = a + K$, then set $\mu(A) = f(a)$. That is, μ maps A to whatever f would map A's offset.

Before we show that μ is the isomorphism we want, we have to make sure that it's actually a function. After all, any coset $A \in R/K$ can have multiple representations: in $\mathbb{Z}/\langle 4 \rangle$, for example, we can write the coset $3 + \langle 4 \rangle$ as $7 + \langle 4 \rangle$, $-13 + \langle 4 \rangle$, or an infinite number of other ways, too! If $f(3) \neq f(7)$ or $f(7) \neq f(-13)$, then μ would not be a function, as it would give us different answers for the same coset, depending on which offset we chose!

Proving that μ is in fact a function is called showing μ is ***well-defined:*** we have to show that for any coset, μ produces one and only one result. Let $A \in R/K$ be a coset, and suppose we can write it two different ways: $A = a + K = b + K$ for some $a, b \in R$. We need to show that $\mu(A)$ has only one value, regardless of whether we look at $\mu(a+K)$ or $\mu(b+K)$. Since μ is defined via f, we need to show that $f(a) = f(b)$.

How do we show that? Our difficulty is that $a + K = b + K$. By coset equality, $a - b \in K$. By definition, we can find $k \in K$ such that $a - b = k$. Rewrite this as $a = b + k$. Now apply definitions and properties to reveal that

$$f(a) = f(b+k) = f(b) + f(k) = f(b) + 0 = f(b) \ . \tag{3.11}$$

Linking the ends of the chain together, we have $f(a) = f(b)$. It doesn't matter how we write A: whether we write it as $a + K$ or as $b + K$, we still have $\mu(A) = f(a) = f(b)$, so μ is well-defined.

We still have to show that μ is a homormophism. To show that it's one-to-one, let $A, B \in R/K$, and assume that $\mu(A) = \mu(B)$. By definition, there exist $a, b \in R$ such that $A = a+K$ and $B = b+K$. By substitution, $\mu(a+K) = \mu(b+K)$. By definition, $\mu(a+K) = f(a)$ and $\mu(b+K) = f(b)$. By substitution, $f(a) = f(b)$. Rewrite as $f(a) - f(b) = 0$. By Lemma 3.81 below, $f(a-b) = 0$. By definition, $a - b \in K$: it maps to 0, and K contains all elements of R that map to 0. By coset equality, $a + K = b + K$. By substitution, $A = B$. We have now shown that μ is one-to-one.

To show that it's onto, let $s \in S$. Recall that f is itself onto, so there exists $r \in R$ such that $f(r) = s$. By definition, $\mu(r+K) = f(r) = s$, so s has a preimage in R/K under μ, and μ maps onto S.

To show that it preserves the operations, let $A, B \in R/K$. We need to show that $\mu(A+B) = \mu(A) + \mu(B)$ and $\mu(AB) = \mu(A)\mu(B)$. By definition, there exist $a, b \in R$ such that $A = a + K$ and $B = b + K$. The definition of coset arithmetic tells us that $A + B = (a+b) + K$ and $AB = ab + K$. By definition, $\mu(A+B) = f(a+b)$ and $\mu(AB) = f(ab)$. Recall that f is a homomorphism. By definition, $f(a+b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$. By definition, $f(a) = \mu(a+K)$ and $f(b) = \mu(b+K)$. Linking the ends of the chains of equations, as well as making use of a final substitution, we obtain

$$\mu(A+B) = \mu(A) + \mu(B) \quad \text{and} \quad \mu(AB) = \mu(A)\mu(B) \ .$$

We have shown that μ is a homomorphism. We already showed that it was one-to-one and onto, so μ is an isomorphism.

Finally, we show that $f = \mu \circ \nu$. Let $r \in R$, and $s = f(r)$. Applying definitions obtains

$$(\mu \circ \nu)(s) = \mu(\nu(s)) = \mu(s+K) = f(s) \ ,$$

as desired. □

Lemma 3.81. *If f is a homomorphism from a ring R to a ring S, then*

(A) $f(0) = 0$*;*

(B) $f(-a) = -f(a)$*; and*

(C) $f(a-b) = f(a) - f(b)$*.*

Proof. (A) By the identity property, $0 + 0 = 0$. By substitution, $f(0+0) = f(0)$. By the homormophism property, $f(0+0) = f(0) + f(0)$. By subsitution, $f(0) + f(0) = f(0)$. Add $-f(0)$ to both sides, and we have $-f(0) + [f(0) + f(0)] = -f(0) + f(0)$. By the associative property, we can rewrite the left-hand side as $[-f(0) + f(0)] + f(0) = -f(0) + f(0)$. By the inverse property, we can rewrite the equation as $0 + f(0) = 0$, and by the identity property we can rewrite it as $f(0) = 0$, as desired.

(B) By the inverse property, $a + (-a) = 0$. By substitution and (A), $f(a+(-a)) = f(0) = 0$, so $f(a+(-a)) = 0$. By the homomorphism property, $f(a) + f(-a) = 0$. Rewrite as $f(-a) = -f(-a)$, and we are done.

(C) By the definition of subtraction, Exercise 3.9 and Theorem 3.7, $f(a) - f(b) = f(a) + [-f(b)]$. By part (B), $f(a) + [-f(b)] = f(a) + f(-b)$. By the homorphism property, $f(a) + f(-b) = f(a+(-b))$. As above, $f(a+(-b)) = f(a-b)$. Linking the ends of this chain of equations together, we have $f(a) - f(b) = f(a-b)$. □

Exercises

Exercise 3.82. State the definition or property that justifies each equality of the chain of equations (3.11).

Exercise 3.83. Show that $\mathbb{Z}_2 \cong \mathbb{Z}[x] / \langle 2, x \rangle$.
Hint: Imitate Examples 3.75 and 3.76

Exercise 3.84. Let R be a ring, and $r \in R$.

(a) If $1 \times r = 0$, then $r = 0$. In other words, the Zero Product Rule always applies to 1, even if it does not generally apply in R.
Hint: What do we already know is special about 1?

(b) If f is a homomorphism from R to a ring S, then $f(1) = 1$.
Hint: Try something similar to Lemma 3.81(A), only using multiplication instead of addition. You'll also need part (a) of this exercise.

Exercise 3.85. Show that isomorphism is transitive. That is, show that if R, S, and T are rings with $R \cong S$ and $S \cong T$, then $R \cong T$.
Hint: Use the isomorphisms $f : R \to S$ and $g : S \to T$ to build a new isomorphism $h : R \to T$. Don't forget to prove the four isomorphism properties.

Exercise 3.86. Let f be a homomorphism from a ring R to a ring S, and let K be its kernel; that is,

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

Show that K is an ideal of R.
Hint: This really ought to spill out from applying the definitions of homomorphism, ideal, and kernel.

Exercise 3.87. Show that $f : \mathbb{Z}[x] \to \mathbb{Z}_2$ by $f(a_m x^m + \cdots + a_1 x + a_0) = a_0$ preserves multiplication.

Exercise 3.88. Let $m \geq 2$, and let $f : \mathbb{Z}[x] \to \mathbb{Z}_2$ by $f(a_m x^m + \cdots + a_1 x + a_0) = r$, where r is the remainder of dividing a_0 by m.

(a) Show that f is an onto homomorphism.

(b) Show that the kernel of f is $\langle m, x \rangle$.

(c) Explain how this proves that $\mathbb{Z}[x]/\langle m,x \rangle \cong \mathbb{Z}_2$.

3.7 Factorization

Section 2.4 introduced the idea of units and associates in order to bring some precision to the factorization of polynomials. These ideas carry over to an arbitrary ring. Given a ring R and nonzero elements $a, r, s, t \in R$, we say that:

- r is a ***unit*** of R if it has a multiplicative inverse in R;

- a is an ***associate*** of r if there is a unit $u \in R$ such that $a = ur$;
- r ***factors*** in R if $r = st$ and neither s nor t is a unit; and
- r is ***irreducible*** over R if it is not a unit of R and does not factor in R.

Example 3.89. You might think that $6 = 2 \times 3$ means that "6 factors," but in fact it depends on the ring we're working in. If we consider 6 as an element of of $\mathbb{Z}$, then 6 factors indeed, because neither 2 nor 3 is a unit.

However, if we consider 6 as an element of $\mathbb{Q}$, then 6 does not factor, because every expression $6 = ab$ involves units. For instance, $6 = 2 \times 3$, but 2 is a unit with inverse $1/2$, and 3 is also a unit with inverse $1/3$. On the other hand, 6 is not irreducible, either, because 6 is itself a unit with inverse $1/6$.

Example 3.90. In a similar way, whether $2x + 2 = 2(x + 1)$ is a factorization depends on the ring we're working it. If we consider $2x+2$ as an element of $\mathbb{Z}[x]$, then it factors indeed, because neither 2 nor $x + 1$ is a unit.

However, if we consider $2x + 2$ as an element of $\mathbb{Q}[x]$, then $2x + 2$ does not factor, because every expression $2x + 2 = pq$ involves units. For instance, $2x + 2 = 2(x + 1)$, but 2 is a unit with inverse $1/2$.

More generally, if $2x + 2 = pq$, then by substitution, $1 = \deg(2x + 2) = \deg(pq) = \deg(p) + \deg(q)$. Linking the ends of the chain, we have $1 = \deg(p) + \deg(q)$. This is possible only if $\deg(p) = 0$ or $\deg(q) = 0$, but in that case p or q is a nonzero constant polynomial, which always has a multiplicative inverse in $\mathbb{Q}[x]$, and is thus a unit.

Sometimes, irreducible ≠ prime

At this point we encounter a distinction with a difference. Recall from Theorem 1.66 that a prime number, or an irreducible number, always satisfied Euclid's Lemma,

$$p \mid ab \quad \Longrightarrow \quad p \mid a \text{ or } p \mid b \ .$$

Theorem 2.45 gave us a similar result for polynomials. You might think, then, that this is true in general.

Example 3.91. Let $R = \left\{a + b\sqrt{-5} : a, b \in \mathbb{Z}\right\}$. You will show in Exercise 3.101 that R is a ring. However, something strange happens. To explain why takes several steps.

We first claim that $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$ are irreducible. To see why, suppose there exist $a + b\sqrt{-5}, c + d\sqrt{-5} \in R$ such that

$$1 + \sqrt{-5} = \left(a + b\sqrt{-5}\right)\left(c + d\sqrt{-5}\right) \ .$$

Expanding the right-hand side gives us

$$1 + \sqrt{-5} = (ac - 5bd) + (ad + bc)\sqrt{-5} \ .$$

Rewrite this again as

$$1 - (ac - 5bd) = \sqrt{-5} \times [(ad + bc) - 1] \ .$$

The left- and right-hand sides are equal, but the left-hand side is an integer, while the right-hand side is not, unless both are 0:

$$1-(ac-5db)=0 \quad \text{and} \quad (ad+bc)-1=0\ .$$

Rewrite as

$$1=(ac-5db) \quad \text{and} \quad (ad+bc)=1\ .$$

Suppose we know a and b. (We don't, but pretend we do; that's what "suppose" means.) We can solve for c and d using a technique similar to elimination:

$$\begin{array}{rcl} ac-5bd &=& 1\\ ad+\ \ bc &=& 1\\ \hline \end{array} \implies \begin{array}{rcl} acd-5bd^2 &=& d\\ -\quad acd+\ \ bc^2 &=& c\\ \hline -b\left(c^2+5d^2\right) &=& d-c \end{array} \implies b\left(c^2+5d^2\right)=c-d\ .$$

Look at the left- and right-hand sides of that last equation. For any integer, $c^2+5d^2 \geq c^2 \geq c$, with equality only if $c=1$ and $d=0$. Those values give us $c+d\sqrt{-5}=1$, and 1 is a unit, so we don't care about that case. Hence $c \neq 1$ or $d \neq 0$, and $c^2+5d^2 > c$, which means $b\left(c^2+5d^2\right) > c$. The only way we can have $b\left(c^2+5d^2\right)=c-d$ is if $d<0$, but then $|d|=-d$, $b>0$, and

$$b\left(c^2+5d^2\right) \geq c^2+5d^2 > c+5\,|d| > c-d\ ,$$

a contradiction. We conclude that $1+\sqrt{-5}$ does not in fact factor. The proof for $1-\sqrt{-5}$ is similar.

Notice that $\left(1+\sqrt{-5}\right)\left(1-\sqrt{-5}\right)=6$. As you well know, $2\times3=6$. Are 2 and 3 irreducible in R? Indeed they are. Suppose there exist $a+b\sqrt{-5}, c+d\sqrt{-5} \in R$ such that $3=\left(a+b\sqrt{-5}\right)\left(c+d\sqrt{-5}\right)$. Expanding the right-hand side gives us

$$3=(ac-5bd)+(ad+bc)\sqrt{-5}\ .$$

Rewrite this again as

$$3-(ac-5bd)=(ad+bc)\sqrt{-5}\ .$$

As before, the left- and right-hand sides are equal if and only if they are both 0, so

$$3-(ac-5bd)=0 \quad \text{and} \quad ad+bc=0\ .$$

Rewrite this again as

$$ac-5bd=3 \quad \text{and} \quad ad+bc=0\ .$$

Once again, suppose we know a and b. We can solve for c and d using a technique similar to elimination:

$$\begin{array}{rcl} ac-5bd &=& 3\\ ad+\ \ bc &=& 0\\ \hline \end{array} \implies \begin{array}{rcl} acd-5bd^2 &=& 3d\\ -\quad acd+\ \ bc^2 &=& 0\\ \hline -b\left(c^2+5d^2\right) &=& 3d \end{array} \implies b\left(c^2+5d^2\right)=-3d\ .$$

Right away we see that $b<0$. However,

$$\left|b\left(c^2+d^2\right)\right|=|b|\left(c^2+d^2\right) \geq |b|\,d^2 \quad \text{and} \quad |-3d|=3\,|d|\ .$$

The left-hand sides of each must be equal, so

$$3\,|d| \geq |b|\,d^2 \ .$$

This is possible only if $d \in \{0, \pm 1\}$, and if $d = \pm 1$ then $3 \geq |b|$, so $b \in \{0, \pm 1, \pm 2, \pm 3\}$. Recall that what we actually have to satisfy is $b\left(c^2 + 5d^2\right) = -3d$; which values of b, c, and d satisfy this? By testing each, we find that only one combination works:

$$(b, c, d) \quad = \quad (0, c, 0) \ .$$

are the only possibilities. The factorization now becomes

$$3 = a \times c \ .$$

Recall that a and c are integers, but 3 is irreducible as an integer. We conclude that it does not factor in R, either. The proof that 2 does not factor in R is similar.

We have now show that $1 \pm \sqrt{-5}$, 2, and 3 are irreducible in R. Recall that

$$\left(1 + \sqrt{-5}\right) \times \left(1 - \sqrt{-5}\right) = 6 = 2 \times 3 \ .$$

Neither factor on the left divides either factor on the right, and vice-versa. This contradicts Euclid's Lemma in this ring.

Prime elements of a ring

We have seen that Euclid's Lemma is valid in $\mathbb{Z}$ and $\mathbb{Q}[x]$ (Theorems 1.66 and 2.45), but not for an arbitrary ring. Given that it is a very useful property, mathematicians made the following choice. They kept the name "irreducible" to mean what we've called it all along, but gave the word "prime" a different meaning in ring theory. In a ring R, we say that a nonzero $r \in R$ is ***prime*** if it is not a unit and it satisfies Euclid's Lemma; that is, for any $a, b \in R$ such that $r \mid ab$, at least one of $r \mid a$ or $r \mid b$ must be true.

Now that we've distinguished these two ideas, how do we decide when they are the same in a ring?

Theorem 3.92. *"Prime" and "irreducible" mean the same thing in $\mathbb{Z}$, $\mathbb{Z}_m$ if m is a prime integer, and $\mathbb{F}[x]$ whenever $\mathbb{F}$ is a field.*

Proof. Theorem 1.66 proves this for $\mathbb{Z}$. Theorem 2.45 proves it for $\mathbb{Q}[x]$, but the proof works for $\mathbb{F}[x]$ as well, so long as we also prove the Euclidean Algorithm for $\mathbb{F}[x]$, which you should have done in Exercise 3.72.

That leaves $\mathbb{Z}_m$. Let $m \geq 2$ be a prime integer and $a \in \mathbb{Z}_m$ be nonzero. Every nonzero element of $\mathbb{Z}_m$ is a unit (Theorem 1.96). Both prime and irreducible elements of a ring are nonunits, so $\mathbb{Z}_m$ has neither prime nor irreducible elements, so it satisfies the claim "vacuously" (there is nothing in $\mathbb{Z}_m$ to contradict it). □

Does $\mathbb{Z}_m$ guarantee that primes are irreducible even when m is not prime?

Example 3.93. Primes and irreducibles need not be the same thing in $\mathbb{Z}_m$. For example, suppose $m = 6$. We claim that 2 is prime, but not irreducible.

It is certainly not a unit, as $\gcd(2, 6) \neq 1$.

To see that 2 is prime, suppose $2 \mid ab$ in $\mathbb{Z}_6$. Choose $q \in \mathbb{Z}_6$ such that $2q = ab$ in $\mathbb{Z}_6$. By Theorem 1.82, $6 \mid (2q - ab)$ in $\mathbb{Z}$. Choose $r \in \mathbb{Z}$ such that $6r = 2q - ab$ and rewrite as $ab = 2(q - 3r)$, so $2 \mid ab$ in $\mathbb{Z}$. We know that 2 is irreducible in $\mathbb{Z}$, and by Euclid's Lemma it is prime in $\mathbb{Z}$, so $2 \mid a$ or $2 \mid b$ in $\mathbb{Z}$. We'll say that $2 \mid a$; choose $s \in \mathbb{Z}$ such that $2s = a$. Then $2s - a = 0$, and $6 \mid (2s - a)$, so $2s = a$ in $\mathbb{Z}_6$, as well. In other words, $2 \mid a$, satisfying the definition of prime.

However, $2 = 2 \times 4$ in $\mathbb{Z}_6$, and neither 2 nor 4 is a unit, so 2 is not irreducible.

As it happens, $\mathbb{Z}_6$ has no irreducible elements at all: 1 and 5 are units, while $2 = 2 \times 4$, $3 = 3 \times 3$, and $4 = 2 \times 2$. These are strange factorizations, but that's what happens sometimes. If $\mathbb{Z}_m$ does have irreducible elements, however, those elements are in fact prime.

Theorem 3.94. *Irreducible elements of $\mathbb{Z}_m$ are also prime.*

(The real issue in $\mathbb{Z}_m$, then, is whether it even has irreducible elements!)

Proof. Let $a \in \mathbb{Z}_m$, and assume a is irreducible. Assume further that there exist $b, c \in \mathbb{Z}_m$ such that $a \mid bc$. Choose $q \in \mathbb{Z}_m$ such that $aq = bc$ in $\mathbb{Z}_m$. By Theorem 1.82, $m \mid (aq - bc)$ in $\mathbb{Z}$. Choose $r \in \mathbb{Z}$ such that $mr = aq - bc$ in $\mathbb{Z}$.

First let $d = \gcd(a, m)$. If $d = 1$, then by Theorem 1.96, a would be a unit, and units are by definition not irreducible. So $d \neq 1$. We now consider two cases.

Case 1. Suppose a is irreducible in $\mathbb{Z}$, as well; that is, $d = a$. Choose $s \in \mathbb{Z}$ such that $m = as$. By substitution, $(as)r = aq - bc$ in $\mathbb{Z}$. Rewrite as $a(sr - q) = bc$ in $\mathbb{Z}$. By definition, $a \mid bc$ in $\mathbb{Z}$. By Euclid's Lemma, $a \mid b$ or $a \mid c$ in $\mathbb{Z}$. Say that $a \mid b$ in $\mathbb{Z}$, and choose $t \in \mathbb{Z}$ such that $at = b$ in $\mathbb{Z}$. Recall that $1 < a, b < m$, so $1 < t < m$, as well, so $t \in \mathbb{Z}_m$. Moreover, $at - b = 0$, so $m \mid (at - b)$, so $at = b$ in $\mathbb{Z}_m$. By definition, $a \mid b$ in $\mathbb{Z}_m$. Recall that $b, c \in \mathbb{Z}_m$ are any pair such that $a \mid bc$ in $\mathbb{Z}_m$. By definition, a is prime.

Case 2. Suppose a is not irreducible in $\mathbb{Z}$; that is, it factors as $a = xy$ for some $x, y \in \mathbb{Z}$. Without loss of generality, we may assume that $1 < x, y < a$, so that $x, y \in \mathbb{Z}_m$, as well, so $a = xy$ in $\mathbb{Z}_m$, as well. By hypothesis, a is irreducible in $\mathbb{Z}_m$, so if $a = xy$ in $\mathbb{Z}_m$, one of x or y is a unit. Without loss of generality, assume that y is a unit.

It cannot be that x is also a unit, because then a would be a unit itself, with inverse $x^{-1}y^{-1}$. So x is not a unit, and it cannot factor; otherwise, a would factor, as well. So x is itself irreducible in $\mathbb{Z}_m$.

So far, we have shown that if $a = xy$ in $\mathbb{Z}$, where x is irreducible in $\mathbb{Z}$, then x must also be irreducible in $\mathbb{Z}_m$. By Case 1, x is prime. By substitution, $x \mid bc$ in $\mathbb{Z}_m$, and by the definition of prime, $x \mid b$ or $x \mid c$ in $\mathbb{Z}_m$. Say that $x \mid b$ in $\mathbb{Z}_m$. Choose $s \in \mathbb{Z}_m$ such that $xs = b$ in $\mathbb{Z}_m$. Recall that $a = xy$ in $\mathbb{Z}_m$, where y is a unit. Rewrite as $ay^{-1} = x$ in $\mathbb{Z}_m$. By substitution, $\left(ay^{-1}\right)s = b$ in $\mathbb{Z}_m$. By the associative property, $a\left(y^{-1}s\right) = b$ in $\mathbb{Z}_m$. By closure of multiplication, $y^{-1}s \in \mathbb{Z}_m$, and by definition of divisibility, $a \mid b$ in $\mathbb{Z}_m$.

Recall that $b, c \in \mathbb{Z}_m$ are any pair such that $a \mid bc$ in $\mathbb{Z}_m$. By definition, a is prime.

□

Let's sum up what we've seen so far.

1. Irreducible elements are not always prime, as we saw with $1+\sqrt{-5}$.
2. Prime elements are not always irreducible, as we saw in $\mathbb{Z}_6$.
3. However, prime and irreducible often are the same thing, as we saw in $\mathbb{Z}$, $\mathbb{F}[x]$, and in any $\mathbb{Z}_m$ that actually has irreducible elements.

Principal ideals of irreducible elements

We've spent some time looking at principal ideals, so let's look at how principal ideals generated by prime or irreducible elements behave. Since "prime" and "irreducible" are both defined in terms of divisibility, let's look first at how divisibility relates to principal ideals.

Example 3.95. First let's consider how divisibility and principal ideals interact in a ring that's easy to consider, such as $\mathbb{Z}$. We know that $2 \mid 6$. The corresponding principal ideals are

$$\langle 2\rangle = \{\ldots,-2,0,2,4,\ldots\} \quad \text{and} \quad \langle 6\rangle = \{\ldots,-6,0,6,12,\ldots\} \ .$$

Every element of $\langle 6\rangle$ is also in $\langle 2\rangle$, but not vice versa. In this particular case, $\langle 6\rangle \subsetneq \langle 2\rangle$. (We could also write $\langle 6\rangle \subseteq \langle 2\rangle$.)

In a similar way, $\langle 2\rangle \subsetneq \langle a\rangle$ only if $a \mid 2$ and $a \neq 2$. By 2's irreducibility, that means $a = \pm 1\ldots$ but $\langle 1\rangle = \mathbb{Z}$. So $\langle 2\rangle$ is in some sense a "maximal" ideal, and this is due to its being irreducible.

Let's see how this works in general.

Lemma 3.96. *Let R be a ring, and $a, b \in R$. Then $a \mid b$ if and only if $\langle b\rangle \subseteq \langle a\rangle$.*

Proof. Assume $a \mid b$, and let $x \in \langle b\rangle$. By Theorem 3.17, we can choose $r \in R$ such that $x = rb$. By definition of divisibility, we can find $q \in R$ such that $b = sa$. By substitution, $x = r(sa) = (rs)a$. By closure of multiplication, $rs \in R$. By absorption and Theorem 3.17, $(rs)a \in \langle a\rangle$. By substitution, $x \in \langle a\rangle$. As x was an arbitrary element of $\langle b\rangle$, we have shown that $\langle b\rangle \subseteq \langle a\rangle$.

Conversely, suppose $\langle b\rangle \subseteq \langle a\rangle$. By definition of subset, $b \in \langle a\rangle$. By Theorem 3.17, $b = ra$ for some $r \in R$. By definition of divisibility, $a \mid b$. □

How is this result useful?

Theorem 3.97. *Let R be a principal ideal ring, and $r \in R$. If r is irreducible, then $\langle r\rangle$ is* **maximal** *in the sense that the only ideal that contains $\langle r\rangle$ is $\langle 1\rangle = R$: the ring itself.*

Proof. Assume r is irreducible. Let I be any idea of R, and assume that $\langle r\rangle \subseteq I$. Assume that $\langle r\rangle \neq I$. By definition of subset, r itself is in I. Recall that R is a principal ideal ring, so $I = \langle i\rangle$ for some $i \in R$. By Lemma 3.96, $i \mid r$. Choose $q \in R$ such that $r = iq$.

By definition of irreducible, one of i or q is a unit. If i is a unit, then by Exercise 3.103 below, $\langle i\rangle = R$. Otherwise, q is a unit, and i is an associate of of r. By Exercise below, $\langle i\rangle = \langle r\rangle$.

We have shown that if r is irreducible and I is any ideal that contains $\langle r\rangle$, then $I = \langle r\rangle$ or $I = R$. That leaves no room for an ideal to "squeeze in between" $\langle r\rangle$ and R! Hence $\langle r\rangle$ is maximal in the sense defined by the theorem. □

A ***principal ideal domain*** is a principal ideal ring that satisfies the Zero Product Property.

Corollary 3.98. *Let R be a principal ideal domain, and $r \in R$ a nonzero element. The following are equivalent.*

(A) The element r is irreducible.

(B) The ideal $\langle r \rangle$ is maximal (in the sense defined above).

(C) The quotient ring $R/\langle r \rangle$ is a field.

In fact, (A)$\Longrightarrow$(B)$\Longleftrightarrow$(C) even in a principal ideal ring.

Proof. We have already shown that (A)$\Longrightarrow$(B) in Theorem 3.97. We will show that (B)$\Longrightarrow$(A), and then that (B)$\Longleftrightarrow$(C).

(B) $\Longrightarrow$(A)? Assume (B); that is, the quotient ring $R/\langle r \rangle$ is maximal. To show that r is irreducible, assume that there exist $a, b \in R$ such that $r = ab$. We must show that one of a or b is a unit. If a is a unit, then we are done, so assume that a is not a unit. We must show that b is a unit.

By absorption, $r \in \langle a \rangle$. By Theorem 3.17, every element $x \in \langle r \rangle$ has the form $x = qr$ for some $q \in R$, so by substitution $x = q(ab)$. By the associative and commutative properties, $x = (qb)a$. By closure of multiplication, $qa \in R$; by absorption, $(qb)a \in \langle a \rangle$; by substitution, $x \in \langle a \rangle$. As $x \in \langle r \rangle$ was arbitrary, $\langle r \rangle \subseteq \langle a \rangle$.

If $\langle r \rangle = \langle a \rangle$, then $a \in \langle r \rangle$. By Theorem 3.17, $a = sr$ for some $s \in R$. By substitution, $sr = s(ab)$. By the associative and commutative properties, $sr = (sb)a$. By susbtitution, $a = (sb)a$. Rewrite as $a(1 - sb) = 0$. Recall that R is a principal ideal domain; it satisfies the Zero Product Property, so $a = 0$ or $1 - sb = 0$. If $a = 0$, then $r = 0$, a contradiction. Hence $1 - sb = 0$. Rewrite as $1 = sb$, so that b is a unit, as desired: r is irreducible.

(B)$\Longleftrightarrow$(C)? First assume (B); that is, the ideal $\langle r \rangle$ is maximal. Let $X \in R/\langle r \rangle$ be nonzero. By definition, there exists $x \in R$ such that $X = x + \langle r \rangle$. By coset equality, $x \notin \langle r \rangle$. Let $I = \langle x, r \rangle$. By construction, $\langle r \rangle \subsetneq I$. By definition of maximal, $I = R$. By Exercise 3.26, $I = \langle 1 \rangle$, and by substitution, $\langle x, r \rangle = \langle 1 \rangle$. By definition, $1 \in \langle x, r \rangle$. By Theorem 3.17, there exist $a, b \in R$ such that $1 = ax + br$. Rewrite as $1 - ax = br$. By definition, $r \mid (1 - ax)$. By coset equality, $1 + \langle r \rangle = ax + \langle r \rangle$. By coset arithmetic, $ax + \langle r \rangle = (a + \langle r \rangle)(x + \langle r \rangle)$. By substitution, $1 + \langle r \rangle = (a + \langle r \rangle)(x + \langle r \rangle)$. By definition, $a + \langle r \rangle$ is the multiplicative inverse of $x + \langle r \rangle = X$, so X is a unit. As X was an arbitrary nonzero element of $R/\langle r \rangle$, we conclude that $R/\langle r \rangle$ is a field.

Conversely, assume (C); that is, the quotient ring $R/\langle r \rangle$ is a field. Let I be any ring of R such that $\langle r \rangle \subsetneq I$. By hypothesis, R is a principal ideal ring, so $I = \langle i \rangle$ for some $i \in R$. If $i \in \langle r \rangle$, then $I = \langle i \rangle \subseteq \langle r \rangle$, a contradiction, so $i \notin \langle r \rangle$. By coset equality, $i + \langle r \rangle \neq \langle r \rangle$; that is, $i + \langle r \rangle$ is nonzero in $R/\langle r \rangle$. Recall that $R/\langle r \rangle$ is a field; choose a multiplicative inverse of $i + \langle r \rangle$ and suppose that we can write it as $j + \langle r \rangle \in R/\langle r \rangle$. By definition, $1 + \langle r \rangle = (i + \langle r \rangle)(j + \langle r \rangle)$. By coset arithmetic, $1 + \langle r \rangle = ij + \langle r \rangle$. By coset equality, $1 - ij \in \langle r \rangle$. By Theorem 3.17, $1 - ij = rs$ for some $s \in R$. Rewrite as $1 = ij + rs$. Recall that $r \in \langle i \rangle$, so by absorption both $ij, rs \in I$. By Exercise 3.25, $ij + rs \in I$. By substitution, $1 \in I$. By Exercise 3.26, $I = R$. Recall that I was any ring of R such that $\langle r \rangle \subsetneq I$; we saw that $I = R$ regardless. By definition, $\langle r \rangle$ is maximal. □

The requirement that R be a principal ideal domain is essential for (A)$\Longleftrightarrow$(B), but it's not essential for (B)$\Longleftrightarrow$(C) to have a principal ideal ring; that just makes the proof more uniform (and arguably less abstract). We outline a more general proof in the exercises.

Constructing the complex numbers

The end of Section 2.5 showed that we could construct the complex number using polynomial congruence with x^2+1. Here we construct a ring that behaves like the complex numbers using the same idea, only in the terminology of ideals. In addition, we use isomorphism to prove that we really have constructed $\mathbb{C}$, or at least something equivalent to it.

Theorem 3.99. *Let $R=\mathbb{R}[x]$ and $I=\langle x^2+1\rangle$. Then $R/I\cong\mathbb{C}$.*

Proof. Let $a\in\mathbb{R}$; then $a^2+1\geq 1>0$, so by the Factor Theorem (Exercise 3.70), x^2+1 does not factor in $\mathbb{R}[x]$. By definition, x^2+1 is irreducible in R. By Corollary 3.98, R/I is a field.

Let $f:\mathbb{R}[x]\to\mathbb{C}$ by

$$f(a_nx^n+\cdots+a_1x+a_0) \quad=\quad a_ni^n+\cdots+a_1i+a_0\ .$$

We claim that f is a homomorphism. To see why, let $p,q\in\mathbb{R}[x]$. By definition, we can choose $a_m,\ldots,a_0,b_m,\ldots,b_0\in\mathbb{R}$ such that $p=a_mx^m+\cdots+a_1x+a_0$ and $q=b_mx^m+\cdots+b_1x+b_0$. By polynomial arithmetic and the definition of f, we have

$$\begin{aligned}f(p+q)&=f((a_m+b_m)x^m+\cdots+(a_1+b_1)x+(a_0+b_0))\\&=(a_m+b_m)i^m+\cdots+(a_1+b_1)i+(a_0+b_0)\end{aligned}$$

while

$$f(p)+f(q)=(a_mx^m+\cdots+a_1x+a_0)+(b_mx^m+\cdots+b_1x+b_0)\ .$$

By the associative, commutative, and distributive properties of a ring, the last expressions of the previous two equations are equal, so $f(p+q)=f(p)+f(q)$. Similarly,

$$f(pq)=f\left(\sum_{j=1}^{2m}\left(\sum_{k+\ell=j}a_kb_\ell\right)x^j\right)=\sum_{j=1}^{2m}\left(\sum_{k+\ell=j}a_kb_\ell\right)i^j$$

while

$$\begin{aligned}f(p)f(q)&=(a_mi^m+\cdots+a_1i+a_0)(b_mi^m+\cdots+b_1i+b_0)\\&=\sum_{j=1}^{2m}\left(\sum_{k+\ell=j}a_kb_\ell\right)i^j\ .\end{aligned}$$

By the associative, commutative, and distributive properties of a ring, the last expressions of the previous two equations are equal, so $f(pq)=f(p)f(q)$, and f is indeed a homomorphism.

Recall that the kernel of f is the set of all elements of $\mathbb{R}[x]$ that f maps to 0 in $\mathbb{C}$. By definition, $f(0)=0$, but also $f(x^2+1)=i^2+1=-1+1=0$, so x^2+1 is in the kernel of f. Since f is a homomorphism, *every* multiple of x^2+1 is in the kernel, as

$$f\left(g\times\left(x^2+1\right)\right)=f(g)\times f\left(x^2+1\right)=f(g)\times 0=0\ .$$

On the other hand, for any $h \in \mathbb{R}[x]$ that is not a multiple of x^2+1, then by the Division Theorem we can write $h = (x^2+1) \cdot q + r$ for some $q, r \in \mathbb{R}[x]$, with $r \neq 0$ and $\deg(r) < 2$. Applying the homomorphism property, we have

$$\begin{aligned} f(h) &= f\left(\left(x^2+1\right) \cdot q + r\right) \\ &= f\left(x^2+1\right) \cdot f(q) + f(r) \\ &= 0 \cdot f(q) + f(r) \\ &= f(r) \ . \end{aligned}$$

Since $\deg(r) < 2$, we can choose $c, d \in \mathbb{R}$ such that $r = cx + d$, and by definition of f, we have

$$f(r) = f(cx+d) = ci + d \neq 0 \ .$$

By definition, r is not in the kernel, so h is not in the kernel.

The kernel of f thus consists exclusively of multiples of x^2+1; that is, the kernel of f is $K = \langle x^2+1 \rangle$. By the Isomorphism Theorem, there exists an isomorphism μ from R/K to $\mathbb{C}$, so that $R/K \cong \mathbb{C}$. □

In other words, if we look at the offsets of $\langle x^2+1 \rangle$, all of which can be written in the form $ax + b$ for $a, b \in \mathbb{R}$, we are really looking at $\mathbb{C}$ itself. In addition, x^2+1 is equivalent to 0, which means x^2 is equivalent to -1, which means x is equivalent to i, the square root of -1. In short, we have again constructed the imaginary number, but this time we constructed it in the process of constructing all of $\mathbb{C}$… and we used ideals to do it, which allows us to conclude that

The "imaginary" number is most certainly "real".

We have now laid the essential foundation of higher algebra, which seems like as good a place as any to wrap things up.

Exercises

Exercise 3.100. In the ring $\mathbb{Z}_{12}$, show that:

(a) 1, 5, 7, and 11 are units;

(b) 2 and 10 are associates;

(c) 2 and 10 are irreducible;

(d) 3 is not irreducible.

Exercise 3.101. Let $R = \left\{a + b\sqrt{-5} : a, b \in \mathbb{Z}\right\}$.

(a) Simplify $\left(1 + 2\sqrt{-5}\right) + \left(-3 + \sqrt{-5}\right)$ and $\left(1 + 2\sqrt{-5}\right) \times \left(-3 + \sqrt{-5}\right)$.

(b) Show that R is a ring under ordinary addition and multiplication.

(c) Show that $1 - \sqrt{-5}$ is irreducible in R.

(d) Show that 2 is irreducible in R.

Exercise 3.102. Show that $\mathbb{Z}_{14}$ has a prime element that is not irreducible. What does that tell you about the irreducible elements of $\mathbb{Z}_{14}$?

Exercise 3.103. Let R be a ring, and $u \in R$. Show that u is a unit if and only if $\langle u \rangle = R$.
Hint: Exercise 3.26 would be useful.

Exercise 3.104. Let $R = \mathbb{R}[x]$ and $I = \left\langle x^2 + 1 \right\rangle$. Simplify the following expressions of R/I. Offsets should have minimal degree.

(A) $(x + I) \cdot (x + I)$

(B) $\left[\left(x^3 + 1\right) + I\right] + [(x - 1) + I]$

(C) $\left[\left(x^2 + 1\right) + I\right] \times \left[\left(x^2 - 1\right) + I\right]$

Exercise 3.105. It is possible to show that the quotient ring of a maximal ideal is always a field, regardless of the original ring. The following lines sketch out a proof; fill in the details.

First, show that if R is a ring whose only ideals are $\{0\}$ and R itself, then it is a field. (Exercise 3.26 will help.)

Now, let R be any ring, and I any ideal of R. Let $Q = R/I$. Suppose A is an ideal of Q; show that there is some ideal J of R such that $A = \{j + I : j \in J\}$ and $I \subseteq J$. *Hint:* Every ideal contains the zero element, so A contains I, and $I = 0 + I$, so $0 \in J$. It remains to show that J is an ideal.

Finally, suppose that I is maximal in R. What does that say about any J that contains I? What does that say about any ideal A of $Q = R/I$? What does that say about Q itself?

Index