A Number Theory Primer:

What Every Computer Scientist Should Know about Number Theory
(v0.13)

Victor Shoup

Chapter 1

Basic properties of the integers

This chapter discusses some of the basic properties of the integers, including the notions of divisibility and primality, unique factorization into primes, greatest common divisors, and least common multiples.

1.1 Divisibility and primality

A central concept in number theory is *divisibility*.

Consider the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. For $a, b \in \mathbb{Z}$, we say that a **divides** b if $az = b$ for some $z \in \mathbb{Z}$. If a divides b, we write $a \mid b$, and we may say that a is a **divisor** of b, or that b is a **multiple** of a, or that b is **divisible by** a. If a does not divide b, then we write $a \nmid b$.

We first state some simple facts about divisibility:

Theorem 1.1. *For all $a, b, c \in \mathbb{Z}$, we have*

(i) $a \mid a$, $1 \mid a$, *and* $a \mid 0$;

(ii) $0 \mid a$ *if and only if* $a = 0$;

(iii) $a \mid b$ *if and only if* $-a \mid b$ *if and only if* $a \mid -b$;

(iv) $a \mid b$ *and* $a \mid c$ *implies* $a \mid (b + c)$;

(v) $a \mid b$ *and* $b \mid c$ *implies* $a \mid c$.

Proof. These properties can be easily derived from the definition of divisibility, using elementary algebraic properties of the integers. For example, $a \mid a$ because we can write $a \cdot 1 = a$; $1 \mid a$ because we can write $1 \cdot a = a$; $a \mid 0$ because we can write $a \cdot 0 = 0$. We leave it as an easy exercise for the reader to verify the remaining properties. □

We make a simple observation. Suppose $a \mid b$ and $b \neq 0$. Then we have $1 \leq |a| \leq |b|$. To see this, suppose $az = b \neq 0$ for some integer z. Then $a \neq 0$ and $z \neq 0$, and it follows that $|a| \geq 1$ and $|z| \geq 1$. We conclude that $|a| \leq |a||z| = |b|$.

Using this observation, we may prove the following:

Theorem 1.2. *For all $a, b \in \mathbb{Z}$, we have $a \mid b$ and $b \mid a$ if and only if $a = \pm b$. In particular, for every $a \in \mathbb{Z}$, we have $a \mid 1$ if and only if $a = \pm 1$.*

Proof. Clearly, if $a = \pm b$, then $a \mid b$ and $b \mid a$. So let us assume that $a \mid b$ and $b \mid a$, and prove that $a = \pm b$. If either of a or b are zero, then the other must be zero as well. So assume that neither is zero. By the above observation, $a \mid b$ implies $|a| \leq |b|$, and $b \mid a$ implies $|b| \leq |a|$; thus, $|a| = |b|$, and so $a = \pm b$. That proves the first statement. The second statement follows from the first by setting $b := 1$, and noting that $1 \mid a$. $\square$

The product of any two non-zero integers is again non-zero. This implies the usual **cancellation law**: if a, b, and c are integers such that $a \neq 0$ and $ab = ac$, then we must have $b = c$; indeed, $ab = ac$ implies $a(b - c) = 0$, and so $a \neq 0$ implies $b - c = 0$, and hence $b = c$.

1.2 Division with remainder

One of the most important facts about the integers is the following fact regarding division with remainder. This fact underpins many other properties of the integers that we shall derive later, including the fact that every positive integer can be expressed uniquely as the product of primes.

Theorem 1.3 (Division with remainder property). *Let $a, b \in \mathbb{Z}$ with $b > 0$. Then there exist unique $q, r \in \mathbb{Z}$ such that $a = bq + r$ and $0 \leq r < b$.*

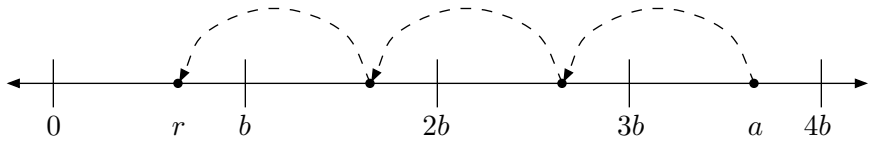
Proof. Consider the set S of non-negative integers of the form $a - bt$ with $t \in \mathbb{Z}$. This set is clearly non-empty; indeed, if $a \geq 0$, set $t := 0$, and if $a < 0$, set $t := a$. Since every non-empty set of non-negative integers contains a minimum, we define r to be the smallest element of S. By definition, r is of the form $r = a - bq$ for some $q \in \mathbb{Z}$, and $r \geq 0$. Also, we must have $r < b$, since otherwise, $r - b$ would be an element of S smaller than r, contradicting the minimality of r; indeed, if $r \geq b$, then we would have $0 \leq r - b = a - b(q + 1)$.

That proves the existence of r and q. For uniqueness, suppose that $a = bq + r$ and $a = bq' + r'$, where $0 \leq r < b$ and $0 \leq r' < b$. Then subtracting these two equations and rearranging terms, we obtain

$$r' - r = b(q - q').$$

Thus, $r' - r$ is a multiple of b; however, $0 \leq r < b$ and $0 \leq r' < b$ implies $|r' - r| < b$; therefore, the only possibility is $r' - r = 0$. Moreover, $0 = b(q - q')$ and $b \neq 0$ implies $q - q' = 0$. $\square$

Theorem 1.3 can be visualized as follows:



Starting with a, we subtract (or add, if a is negative) the value b until we end up with a number in the interval $[0, b)$.

Floors and ceilings. Let us briefly recall the usual **floor** and **ceiling** functions, denoted $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$, respectively. These are functions from $\mathbb{R}$ (the real numbers) to $\mathbb{Z}$. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the greatest integer $m \leq x$; equivalently, $\lfloor x \rfloor$ is the unique integer m such that $m \leq x < m + 1$, or put another way, such that $x = m + \epsilon$ for some $\epsilon \in [0, 1)$. Also, $\lceil x \rceil$ is the smallest integer $m \geq x$; equivalently, $\lceil x \rceil$ is the unique integer m such that $m - 1 < x \leq m$, or put another way, such that $x = m - \epsilon$ for some $\epsilon \in [0, 1)$.

The mod operator. Now let $a, b \in \mathbb{Z}$ with $b > 0$. If q and r are the unique integers from Theorem 1.3 that satisfy $a = bq + r$ and $0 \leq r < b$, we define

$$a \bmod b := r;$$

that is, $a \bmod b$ denotes the remainder in dividing a by b. It is clear that $b \mid a$ if and only if $a \bmod b = 0$. Dividing both sides of the equation $a = bq + r$ by b, we obtain $a/b = q + r/b$. Since $q \in \mathbb{Z}$ and $r/b \in [0, 1)$, we see that $q = \lfloor a/b \rfloor$. Thus,

$$(a \bmod b) = a - b\lfloor a/b \rfloor.$$

One can use this equation to extend the definition of $a \bmod b$ to all integers a and b, with $b \neq 0$; that is, for $b < 0$, we simply define $a \bmod b$ to be $a - b\lfloor a/b \rfloor$.

Theorem 1.3 may be generalized so that when dividing an integer a by a positive integer b, the remainder is placed in an interval other than $[0, b)$. Let x be any real number, and consider the interval $[x, x + b)$. As the reader may easily verify, this interval contains precisely b integers, namely, $\lceil x \rceil, \ldots, \lceil x \rceil + b - 1$. Applying Theorem 1.3 with $a - \lceil x \rceil$ in place of a, we obtain:

Theorem 1.4. *Let $a, b \in \mathbb{Z}$ with $b > 0$, and let $x \in \mathbb{R}$. Then there exist unique $q, r \in \mathbb{Z}$ such that $a = bq + r$ and $r \in [x, x + b)$.*

Div and mod in programming languages. Our definitions of division and remainder are similar to, but not exactly equivalent to, the corresponding operators used in modern programming languages. Typically, in such languages, an expression such as "`5/3`" evaluates to $\lfloor 5/3 \rfloor = 1$; similarly, "`5%3`" evaluates to $5 \bmod 2 = 1$. However, most programming languages treat negative numbers differently that we do here. For example, the evaluation of "`(-5)/3`" would discard the fractional part of $(-5)/3$, yielding -1, rather than $\lfloor (-5)/3 \rfloor = -2$. Also, "`(-5)%3`" would evaluate to $(-5) - (-1) \cdot 3 = -2$, rather than $(-5) - (-2) \cdot 3 = 1$.

Exercise 1.1. Let $a, b, d \in \mathbb{Z}$ with $d \neq 0$. Show that $a \mid b$ if and only if $da \mid db$.

Exercise 1.2. Let m be a positive integer. Show that for every real number $x \geq 1$, the number of multiples of m in the interval $[1, x]$ is $\lfloor x/m \rfloor$; in particular, for every integer $n \geq 1$, the number of multiples of m among $1, \ldots, n$ is $\lfloor n/m \rfloor$.

Exercise 1.3. Let $x \in \mathbb{R}$. Show that $2\lfloor x \rfloor \leq \lfloor 2x \rfloor \leq 2\lfloor x \rfloor + 1$.

Exercise 1.4. Let $x \in \mathbb{R}$ and $n \in \mathbb{Z}$ with $n > 0$. Show that $\lfloor \lfloor x \rfloor / n \rfloor = \lfloor x/n \rfloor$; in particular, $\lfloor \lfloor a/b \rfloor / c \rfloor = \lfloor a/bc \rfloor$ for all positive integers a, b, c.

Exercise 1.5. Let $a, b \in \mathbb{Z}$ with $b < 0$. Show that $(a \bmod b) \in (b, 0]$.

Exercise 1.6. Show that Theorem 1.4 also holds for the interval $(x, x + b]$. Does it hold in general for the intervals $[x, x + b]$ or $(x, x + b)$?

1.3 Greatest common divisors

For $a, b \in \mathbb{Z}$, we call $d \in \mathbb{Z}$ a **common divisor** of a and b if $d \mid a$ and $d \mid b$; moreover, we call such a d a **greatest common divisor** of a and b if d is non-negative and is divisible be every common divisor of a and b.

The reader may note that we have not defined the greatest common divisor to be the numerically largest common divisor, as is sometimes done. Rather, we have defined it in terms of the divisibility relation: it is "greatest" in the sense that it is a divisible by every other common divisor. Our definition is actually a stronger property, and is considered to be the more "mathemtically correct" one, in that it generalizes to other algebraic domains. Unfortunately, it is not painfully obvious from the definition that greatest common divisors even exist, and if they do, that they are uniquely defined. We shall, however, show below that for every $a, b \in \mathbb{Z}$, there exists a unique greatest common divisor of a and b, which we will denote by $\gcd(a, b)$.

We begin by showing that the greatest common divisor of a and b, if it exists, must be unique. To this end, suppose d and d' are greatest common divisors. Then by definition, since d is a greatest common divisor and d' is a common divisor, we must have $d' \mid d$. Similarly, we must gave $d \mid d'$. Therefore, by Theorem 1.2, we have $d = \pm d'$. By definition, greatest common divisors must be non-negative, and so we conclude that $d = d'$. That proves uniqueness. The reader will notice that in the uniqueness proof, we made essential use of the fact that greatest common divisors must be non-negative—in fact, this is precisely the reason why this requirement is included in our definition.

1.3.1 The existence of greatest common divisors and Euclid's algorithm

We shall now prove the existence of $\gcd(a, b)$. The proof, in fact, gives us an efficient algorithm to compute $\gcd(a, b)$. The algorithm is the well-known **Euclidean algorithm**.

Without loss of generality, we may assume that $a \geq b \geq 0$. We prove by induction on b that $\gcd(a, b)$ exists.

The base case is $b = 0$. We claim that $\gcd(a, 0) = a$. To see this, observe that by part (i) of Theorem 1.1, every integer is a divisor of 0. It follows that d is a common divisor of a and 0 if and only if it is a divisor of a. Certainly, a is a divisor of itself and it is non-negative, and from this, it is clear that a satisfies the definition of $\gcd(a, 0)$.

Everything in the above paragraph holds even if $a = 0$. In particular, $\gcd(0, 0) = 0$.

So now assume that $b > 0$. We also assume our induction hypothesis, which states that $\gcd(a', b')$ exists for all integers a', b' with $a' \geq b' \geq 0$ and $b' < b$. If we divide a by b, we obtain the quotient $q := \lfloor a/b \rfloor$ and the remainder $r := a \bmod b$, where $0 \leq r < b$. By our induction hypothesis, applied to $a' := b$ and $b' := r$, we see that $\gcd(b, r)$ exists. From the equation

$$a = bq + r,$$

it is easy to see that if an integer divides both b and r, then it also divides a; likewise, if an integer divides a and b, then it also divides r. It follows that the the set of common divisors of a and b is equal to the set of common divisors of b and r. Moreover, as $\gcd(b, r)$ is the unique non-negative common divisor that is divisible by each of these common divisors, and these common divisors are the same as the common divisors of a and b, we see that $\gcd(a, b)$ exists and is equal to $\gcd(b, r)$.

That finishes the proof of the existence of $\gcd(a, b)$. This proof suggests the following recursive algorithm to actually compute it:

Algorithm Euclid(a, b)**:** On input a, b, where a and b are integers such that $a \geq b \geq 0$, compute $\gcd(a, b)$ as follows:

```
if b = 0
    then return a
    else return Euclid(b, a mod b)
```

Example 1.1. Suppose $a = 100$ and $b = 35$. We compute the numbers (a_i, b_i) that are the inputs to the ith recursive call to Euclid:

i	0	1	2	3
a_i	100	35	30	5
b_i	35	30	5	0

So we have $\gcd(a, b) = a_3 = 5$. $\square$

To actually implement this algorithm on a computer, one has to consider the size of the integers a and b. If they are small enough to fit into a singe machine "word" (typically 64- or 32-bits), then there is no issue. Otherwise, one needs to represent these large integers as vectors of machine words, and implement all of the basic arithmetic operations (addition, subtraction, multiplication, and division) in software. Some programming languages implement these operations as part of a standard library, while others do not.

In these notes, we will not worry about how these basic arithmetic operations are implemented. Rather, we will focus our attention to how many division steps are performed by Euclid's algorithm. Observe that when called on input (a, b), with $b > 0$, the algorithm performs one division step, and then calls itself recursively on input (a', b'), where $a' := b$ and $b' := a \bmod b$. In particular, $b > b'$. So in every recursive call, the second argument decreases by at least 1. It follows that the algorithm performs at most b division steps in total. However, it turns out that the algorithm terminates much faster than this:

Theorem 1.5. *On input (a, b), Euclid's algorithm performs $O(\log b)$ division steps.*

Proof. As observed above, if $b > 0$, the algorithm performs one division step, and then calls itself recursively on input (a', b'), where $a' := b$ and $b' := a \bmod b$, so that $b > b'$. If $b' > 0$, the algorithm performs another division step, and calls itself again on input (a'', b''), where $a'' := b'$ and $b'' := b \bmod b'$, so that $b' > b''$. Consider the quotient $q' := \lfloor b/b' \rfloor$. Since $b > b'$, we have $q' \geq 1$. Moreover, since $b = b'q' + b''$ and $b' > b''$, we have

$$b = b'q' + b'' \geq b' + b'' > 2b''.$$

This shows that after two division steps, the second argument to Euclid, is less than $b/2$, and so, after $2k$ division step, it is less than $b/2^k$. So if we choose k large enough so that $b/2^k \leq 1$, we can be sure that the number division steps is at most $2k$. Setting $k := \lceil \log_2 b \rceil$ does the job. $\square$

1.3.2 Bezout's Lemma and the extended Euclidean algorithm

For integers a and b, have seen not only that $\gcd(a, b)$ exists and is uniquely defined, but that we can efficiently compute it. Next, we prove the following remarkable fact, which will have numerous applications.

Theorem 1.6 (Bezout's Lemma). *Let $a, b \in \mathbb{Z}$ and $d := \gcd(a, b)$.*

(i) We have

$$as + bt = d \quad \text{for some } s, t \in \mathbb{Z}.$$

(ii) For every $d^* \in \mathbb{Z}$, we have

$$d \mid d^* \iff as^* + bt^* = d^* \quad \text{for some } s^*, t^* \in \mathbb{Z}.$$

We first observe that the part (ii) of Bezout's Lemma follows easily from part (i). Let $d^* \in \mathbb{Z}$ be given. On the one hand, if $d \mid d^*$, then $dz = d^*$ for some integer z, and $s^* := zs$ and $t^* := zt$,

where $as + bt = d$, do the job. On the other hand, if $as^* + bt^* = d^*$, then since $d \mid a$ and $d \mid b$, we must have $d \mid (as^* + bt^*) = d^*$.

For part (i) of Bezout's Lemma, we can again assume without loss of generality that that $a \geq b \geq 0$. We prove the statement by induction on b.

The base case is $b = 0$. In this case, $d = a$, and so, if we set $s := 1$ and $t := 0$, then $as + 0t = d$, as required.

So assume $b > 0$. We also assume our induction hypothesis, which states that for all integers a', b' with $a' \geq b' \geq 0$ and $b' < b$, there exist integers s' and t' with $a's' + b't' = d'$, where $d' := \gcd(a', b')$. We divide a by b, obtaining the quotient q and remainder r, where

$$a = bq + r \quad \text{and} \quad 0 \leq r < b.$$

Applying our induction hypothesis to $a' := b$ and $b' := r$, since $\gcd(a, b) = d = \gcd(b, r)$, we see there exist s', t' such that

$$bs' + rt' = d.$$

If we substitute r by $a - bq$ in this equation, and rearrange terms, we obtain

$$at' + b(s' - qt') = d.$$

Thus, $s := t'$ and $t := s' - qt'$ do the job.

That completes the proof of part (i) of Bezout's Lemma. The proof naturally suggests the following algorithm, which is called the **extended Euclidean algorithm**:

Algorithm ExtEuclid(a, b)**:** On input a, b, where a and b are integers such that $a \geq b \geq 0$, compute (d, s, t), where $d = \gcd(a, b)$ and s and t are integers such that $as + bt = d$, as follows:

```
if b = 0 then
    d ← a,  s ← 1,  t ← 0
else
    q ← ⌊a/b⌋,  r ← a mod b
    (d, s', t') ← ExtEuclid(b, r)
    s ← t',  t ← s' − qt'
return (d, s, t)
```

Example 1.2. Continuing with Example 1.1, we compute, in addition, the numbers s_i and t_i returned by the ith recursive call to Euclid, along with the corresponding quotient $q_i = \lfloor a_i/b_i \rfloor$:

i	0	1	2	3
a_i	100	35	30	5
b_i	35	30	5	0
s_i	-1	1	0	1
t_i	3	-1	1	0
q_i	2	1	6	

The rule being applied here is $s_i := t_{i+1}$ and $t_i := s_{i+1} - q_i t_{i+1}$. So we have $s = s_0 = -1$ and $t = t_0 = 3$. One can verify that $as + bt = 100 \cdot (-1) + 35 \cdot (3) = 5 = d$. $\square$

For $a, b \in \mathbb{Z}$, we say that $a, b \in \mathbb{Z}$ are **relatively prime** if $\gcd(a, b) = 1$, which is the same as saying that the only common divisors of a and b are ± 1. One may also say that a and b are **coprime**, which means the same thing as saying that they are relatively prime. Applying part (ii) of Bezout's Lemma with $d^* := 1$, we immediately obtain the following:

Theorem 1.7 (Corollary to Bezout's Lemma). *Let $a, b \in \mathbb{Z}$. Then we have:*

$$a \text{ and } b \text{ are relatively prime} \iff as + bt = 1 \text{ for some } s, t \in \mathbb{Z}.$$

This fact allows us to easily prove the following result, which will be useful in the next section, where we prove the unique factorization property for integers.

Theorem 1.8 (Coprime Lemma). *Let a, b, c be integers such that c divides ab, and a and c are relatively prime. Then c must divide b.*

Proof. Suppose that $c \mid ab$ and $\gcd(a, c) = 1$. By Theorem 1.7, we have $as + ct = 1$ for some $s, t \in \mathbb{Z}$. Multiplying this equation by b, we obtain

$$abs + cbt = b. \tag{1.1}$$

Since c divides ab by hypothesis, and since c clearly divides cbt, it follows that c divides the left-hand side of (1.1), and hence that c divides b. □

EXERCISE 1.7. Let a and b be integers, where either $a \neq 0$ or $b \neq 0$. Let $d := \gcd(a, b)$. Show that:

(a) $d > 0$;

(b) d is the numerically largest common divisor of a and b;

(c) d is the smallest positive integer that can be expressed as $as + bt$ for some integers s and t.

EXERCISE 1.8. Show that for all integers a, b, c, we have:

(a) $\gcd(a, b) = \gcd(b, a)$;

(b) $\gcd(a, b) = |a| \iff a \mid b$;

(c) $\gcd(a, 1) = 1$;

(d) $\gcd(ca, cb) = |c| \gcd(a, b)$.

EXERCISE 1.9. Suppose we run Algorithm ExtEuclid on input $(117, 67)$. Show the steps of the computation by giving the data corresponding to that shown in in the table in Example 1.2.

EXERCISE 1.10. Show that if $a \geq b > 0$, then the values s and t computed by ExtEuclid(a, b) satisfy

$$|s| \leq b/d \quad \text{and} \quad |t| \leq a/d.$$

Hint: prove by induction on b—be careful, you have to stop the induction before b gets to zero, so the last step to consider is when $b \mid a$.

EXERCISE 1.11. Suppose that a, b are integers with $d := \gcd(a, b) \neq 0$.

(a) Show that a/d and b/d are relatively prime.

(b) Show that if s and t are integers such that $as + bt = d$, then s and t are relatively prime.

EXERCISE 1.12. Let n be an integer. Show that if a, b are relatively prime integers, each of which divides n, then ab divides n.

EXERCISE 1.13. Let a, b, c be positive integers satisfying $\gcd(a, b) = 1$. Show that there is a number N, depending on a and b, such for all integers $c \geq N$, we can write $c = as + bt$ for *non-negative* integers s, t.

1.4 Unique factorization into primes

Let n be a positive integer. Trivially, 1 and n divide n. If $n > 1$ and no other positive integers besides 1 and n divide n, then we say n is **prime**. If $n > 1$ but n is not prime, then we say that n is **composite**. The number 1 is not considered to be either prime or composite. Evidently, n is composite if and only if $n = ab$ for some integers a, b with $1 < a < n$ and $1 < b < n$. The first few primes are

$$2, 3, 5, 7, 11, 13, 17, \ldots.$$

While it is possible to extend the definition of prime and composite to negative integers, we shall not do so in this text: *whenever we speak of a prime or composite number, we mean a positive integer.*

A basic fact is that every non-zero integer can be expressed as a signed product of primes in an essentially unique way. More precisely:

Theorem 1.9 (Fundamental theorem of arithmetic). *Every non-zero integer n can be expressed as*

$$n = \pm p_1^{e_1} \cdots p_r^{e_r},$$

where $p_1, \ldots, p_r$ are distinct primes and $e_1, \ldots, e_r$ are positive integers. Moreover, this expression is unique, up to a reordering of the primes.

Note that if $n = \pm 1$ in the above theorem, then $r = 0$, and the product of zero terms is interpreted (as usual) as 1.

The theorem intuitively says that the primes act as the "building blocks" out of which all non-zero integers can be formed by multiplication (and negation). The reader may be so familiar with this fact that he may feel it is somehow "self evident," requiring no proof; however, this feeling is simply a delusion, and most of the rest of this section and the next are devoted to developing a proof of this theorem. We shall give a quite leisurely proof, introducing a number of other very important tools and concepts along the way that will be useful later.

To prove Theorem 1.9, we may clearly assume that n is positive, since otherwise, we may multiply n by -1 and reduce to the case where n is positive.

The proof of the existence part of Theorem 1.9 is easy. This amounts to showing that every positive integer n can be expressed as a product (possibly empty) of primes. We may prove this by induction on n. If $n = 1$, the statement is true, as n is the product of zero primes. Now let $n > 1$, and assume that every positive integer smaller than n can be expressed as a product of primes. If n is a prime, then the statement is true, as n is the product of one prime. Assume, then, that n is composite, so that there exist $a, b \in \mathbb{Z}$ with $1 < a < n$, $1 < b < n$, and $n = ab$. By the induction hypothesis, both a and b can be expressed as a product of primes, and so the same holds for n.

The uniqueness part of Theorem 1.9 is a bit more challenging. However, using the results we have proven so far on greatest common divisors, the task is not so difficult.

Suppose that p is a prime and a is any integer. As the only divisors of p are ± 1 and $\pm p$, we have

$$\begin{aligned} p \mid a &\implies \gcd(a, p) = p, \text{ and} \\ p \nmid a &\implies \gcd(a, p) = 1. \end{aligned}$$

Combining this observation with Theorem 1.8, we have:

Theorem 1.10 (Euclid's Lemma). *Let p be prime, and let $a, b \in \mathbb{Z}$. Then $p \mid ab$ implies that $p \mid a$ or $p \mid b$.*

Proof. Assume that $p \mid ab$. If $p \mid a$, we are done, so assume that $p \nmid a$. By the above observation, $\gcd(a, p) = 1$, and so by Theorem 1.8, we have $p \mid b$. $\square$

An obvious corollary to Theorem 1.10 is that if $a_1, \ldots, a_k$ are integers, and if p is a prime that divides the product $a_1 \cdots a_k$, then $p \mid a_i$ for some $i = 1, \ldots, k$. This is easily proved by induction on k. For $k = 1$, the statement is trivially true. Now let $k > 1$, and assume that statement holds for $k - 1$. Then by Theorem 1.10, either $p \mid a_1$ or $p \mid a_2 \cdots a_k$; if $p \mid a_1$, we are done; otherwise, by induction, p divides one of $a_2, \ldots, a_k$.

Finishing the proof of Theorem 1.9. We are now in a position to prove the uniqueness part of Theorem 1.9, which we can state as follows: if $p_1, \ldots, p_r$ are primes (not necessarily distinct), and $q_1, \ldots, q_s$ are primes (also not necessarily distinct), such that

$$p_1 \cdots p_r = q_1 \cdots q_s, \tag{1.2}$$

then $(p_1, \ldots, p_r)$ is just a reordering of $(q_1, \ldots, q_s)$. We may prove this by induction on r. If $r = 0$, we must have $s = 0$ and we are done. Now suppose $r > 0$, and that the statement holds for $r - 1$. Since $r > 0$, we clearly must have $s > 0$. Also, as p_1 obviously divides the left-hand side of (1.2), it must also divide the right-hand side of (1.2); that is, $p_1 \mid q_1 \cdots q_s$. It follows from (the corollary to) Theorem 1.10 that $p_1 \mid q_j$ for some $j = 1, \ldots, s$, and moreover, since q_j is prime, we must have $p_1 = q_j$. Thus, we may cancel p_1 from the left-hand side of (1.2) and q_j from the right-hand side of (1.2), and the statement now follows from the induction hypothesis. That proves the uniqueness part of Theorem 1.9.

Exercise 1.14. Let n be a composite integer. Show that there exists a prime p dividing n, with $p \leq n^{1/2}$.

Exercise 1.15. Show that two integers are relatively prime if and only if there is no one prime that divides both of them.

Exercise 1.16. Let p be a prime and k an integer, with $0 < k < p$. Show that the binomial coefficient

$$\binom{p}{k} = \frac{p!}{k!(p-k)!},$$

which is an integer, is divisible by p.

Exercise 1.17. An integer a is called **square-free** if it is not divisible by the square of any integer greater than 1. Show that:

(a) a is square-free if and only if $a = \pm p_1 \cdots p_r$, where the p_i's are distinct primes;

(b) every positive integer n can be expressed uniquely as $n = ab^2$, where a and b are positive integers, and a is square-free.

1.5 Some consequences of unique factorization

The following theorem is a consequence of just the existence part of Theorem 1.9:

Theorem 1.11. *There are infinitely many primes.*

Proof. By way of contradiction, suppose that there were only finitely many primes; call them $p_1, \ldots, p_k$. Then set $M := \prod_{i=1}^{k} p_i$ and $N := M + 1$. Consider a prime p that divides N. There must be at least one such prime p, since $N \geq 2$, and every positive integer can be written as a product of primes. Clearly, p cannot equal any of the p_i's, since if it did, then p would divide M, and hence also divide $N - M = 1$, which is impossible. Therefore, the prime p is not among $p_1, \ldots, p_k$, which contradicts our assumption that these are the only primes. □

For each prime p, we may define the function ν_p, mapping non-zero integers to non-negative integers, as follows: for every integer $n \neq 0$, if $n = p^e m$, where $p \nmid m$, then $\nu_p(n) := e$. We may then write the factorization of n into primes as

$$n = \pm \prod_p p^{\nu_p(n)},$$

where the product is over all primes p; although syntactically this is an infinite product, all but finitely many of its terms are equal to 1, and so this expression makes sense.

Observe that if a and b are non-zero integers, then

$$\nu_p(a \cdot b) = \nu_p(a) + \nu_p(b) \quad \text{for all primes } p, \tag{1.3}$$

and

$$a \mid b \iff \nu_p(a) \leq \nu_p(b) \quad \text{for all primes } p. \tag{1.4}$$

From this, it is clear that

$$\gcd(a, b) = \prod_p p^{\min(\nu_p(a), \nu_p(b))}.$$

Least common multiples. For $a, b \in \mathbb{Z}$, a **common multiple** of a and b is an integer m such that $a \mid m$ and $b \mid m$; moreover, such an m is the **least common multiple** of a and b if m is non-negative and m divides all common multiples of a and b. It is easy to see that the least common multiple exists and is unique, and we denote the least common multiple of a and b by $\operatorname{lcm}(a, b)$. Indeed, for all $a, b \in \mathbb{Z}$, if either a or b are zero, the only common multiple of a and b is 0, and so $\operatorname{lcm}(a, b) = 0$; otherwise, if neither a nor b are zero, we have

$$\operatorname{lcm}(a, b) = \prod_p p^{\max(\nu_p(a), \nu_p(b))},$$

or equivalently, $\operatorname{lcm}(a, b)$ may be characterized as the smallest positive integer divisible by both a and b.

It is convenient to extend the domain of definition of ν_p to include 0, defining $\nu_p(0) := \infty$. If we interpret expressions involving "∞" appropriately,[1] then for arbitrary $a, b \in \mathbb{Z}$, both (1.3) and (1.4) hold, and in addition,

$$\nu_p(\gcd(a, b)) = \min(\nu_p(a), \nu_p(b)) \quad \text{and} \quad \nu_p(\operatorname{lcm}(a, b)) = \max(\nu_p(a), \nu_p(b))$$

for all primes p.

[1] The interpretation given to such expressions should be obvious: for example, for every $x \in \mathbb{R}$, we have $-\infty < x < \infty$, $x + \infty = \infty$, $x - \infty = -\infty$, $\infty + \infty = \infty$, and $(-\infty) + (-\infty) = -\infty$. Expressions such as $x \cdot (\pm\infty)$ also make sense, provided $x \neq 0$. However, the expressions $\infty - \infty$ and $0 \cdot \infty$ have no sensible interpretation.

Generalizing gcd's and lcm's to many integers. It is easy to generalize the notions of greatest common divisor and least common multiple from two integers to many integers. Let $a_1, \ldots, a_k$ be integers. We call $d \in \mathbb{Z}$ a common divisor of $a_1, \ldots, a_k$ if $d \mid a_i$ for $i = 1, \ldots, k$; moreover, we call such a d the greatest common divisor of $a_1, \ldots, a_k$ if d is non-negative and all other common divisors of $a_1, \ldots, a_k$ divide d. The greatest common divisor of $a_1, \ldots, a_k$ is denoted $\gcd(a_1, \ldots, a_k)$ and is the unique non-negative integer d satisfying

$$\nu_p(d) = \min(\nu_p(a_1), \ldots, \nu_p(a_k)) \text{ for all primes } p.$$

Analogously, we call $m \in \mathbb{Z}$ a common multiple of $a_1, \ldots, a_k$ if $a_i \mid m$ for all $i = 1, \ldots, k$; moreover, such an m is called the least common multiple of $a_1, \ldots, a_k$ if m divides all common multiples of $a_1, \ldots, a_k$. The least common multiple of $a_1, \ldots, a_k$ is denoted $\operatorname{lcm}(a_1, \ldots, a_k)$ and is the unique non-negative integer m satisfying

$$\nu_p(m) = \max(\nu_p(a_1), \ldots, \nu_p(a_k)) \text{ for all primes } p.$$

Finally, we say that the family $\{a_i\}_{i=1}^k$ is **pairwise relatively prime** if for all indices i, j with $i \neq j$, we have $\gcd(a_i, a_j) = 1$. Certainly, if $\{a_i\}_{i=1}^k$ is pairwise relatively prime, and $k > 1$, then $\gcd(a_1, \ldots, a_k) = 1$; however, $\gcd(a_1, \ldots, a_k) = 1$ does not imply that $\{a_i\}_{i=1}^k$ is pairwise relatively prime.

Rational numbers. Consider the rational numbers $\mathbb{Q} = \{a/b : a, b \in \mathbb{Z},\ b \neq 0\}$. Given any rational number a/b, if we set $d := \gcd(a, b)$, and define the integers $a_0 := a/d$ and $b_0 := b/d$, then we have $a/b = a_0/b_0$ and $\gcd(a_0, b_0) = 1$. Moreover, if $a_1/b_1 = a_0/b_0$, then we have $a_1 b_0 = a_0 b_1$, and so $b_0 \mid a_0 b_1$; also, since $\gcd(a_0, b_0) = 1$, we see that $b_0 \mid b_1$; writing $b_1 = b_0 c$, we see that $a_1 = a_0 c$. Thus, we can represent every rational number as a fraction in **lowest terms**, which means a fraction of the form a_0/b_0 where a_0 and b_0 are relatively prime; moreover, the values of a_0 and b_0 are uniquely determined up to sign, and every other fraction that represents the same rational number is of the form $a_0 c/b_0 c$, for some non-zero integer c.

Exercise 1.18. Let n be an integer. Generalizing Exercise 1.12, show that if $\{a_i\}_{i=1}^k$ is a pairwise relatively prime family of integers, where each a_i divides n, then their product $\prod_{i=1}^k a_i$ also divides n.

Exercise 1.19. Show that for all integers a, b, c, we have:

(a) $\operatorname{lcm}(a, b) = \operatorname{lcm}(b, a)$;

(b) $\operatorname{lcm}(a, b) = |a| \iff b \mid a$;

(c) $\operatorname{lcm}(a, a) = \operatorname{lcm}(a, 1) = |a|$;

(d) $\operatorname{lcm}(ca, cb) = |c| \operatorname{lcm}(a, b)$.

Exercise 1.20. Show that for all integers a, b, we have:

(a) $\gcd(a, b) \cdot \operatorname{lcm}(a, b) = |ab|$;

(b) $\gcd(a, b) = 1 \implies \operatorname{lcm}(a, b) = |ab|$.

Exercise 1.21. Let $a_1, \ldots, a_k \in \mathbb{Z}$ with $k > 1$. Show that:

$$\gcd(a_1, \ldots, a_k) = \gcd(a_1, \gcd(a_2, \ldots, a_k)) = \gcd(\gcd(a_1, \ldots, a_{k-1}), a_k);$$
$$\operatorname{lcm}(a_1, \ldots, a_k) = \operatorname{lcm}(a_1, \operatorname{lcm}(a_2, \ldots, a_k)) = \operatorname{lcm}(\operatorname{lcm}(a_1, \ldots, a_{k-1}), a_k).$$

EXERCISE 1.22. Let $a_1, \ldots, a_k \in \mathbb{Z}$ with $d := \gcd(a_1, \ldots, a_k)$. Show that there exist integers $s_1, \ldots, s_k$ such that $d = a_1 s_1 + \cdots + a_k s_k$.

EXERCISE 1.23. Show that if $\{a_i\}_{i=1}^k$ is a pairwise relatively prime family of integers, then $\mathrm{lcm}(a_1, \ldots, a_k) = |a_1 \cdots a_k|$.

EXERCISE 1.24. Show that every non-zero $x \in \mathbb{Q}$ can be expressed as

$$x = \pm p_1^{e_1} \cdots p_r^{e_r},$$

where the p_i's are distinct primes and the e_i's are non-zero integers, and that this expression in unique up to a reordering of the primes.

EXERCISE 1.25. Let n and k be positive integers, and suppose $x \in \mathbb{Q}$ such that $x^k = n$ for some $x \in \mathbb{Q}$. Show that $x \in \mathbb{Z}$. In other words, $\sqrt[k]{n}$ is either an integer or is irrational.

EXERCISE 1.26. Show that $\gcd(a + b, \mathrm{lcm}(a, b)) = \gcd(a, b)$ for all $a, b \in \mathbb{Z}$.

EXERCISE 1.27. Show that for every positive integer k, there exist k consecutive composite integers. Thus, there are arbitrarily large gaps between primes.

EXERCISE 1.28. Let p be a prime. Show that for all $a, b \in \mathbb{Z}$, we have $\nu_p(a+b) \geq \min\{\nu_p(a), \nu_p(b)\}$, and $\nu_p(a + b) = \nu_p(a)$ if $\nu_p(a) < \nu_p(b)$.

EXERCISE 1.29. For a given prime p, we may extend the domain of definition of ν_p from $\mathbb{Z}$ to $\mathbb{Q}$: for non-zero integers a, b, let us define $\nu_p(a/b) := \nu_p(a) - \nu_p(b)$. Show that:

(a) this definition of $\nu_p(a/b)$ is unambiguous, in the sense that it does not depend on the particular choice of a and b;

(b) for all $x, y \in \mathbb{Q}$, we have $\nu_p(xy) = \nu_p(x) + \nu_p(y)$;

(c) for all $x, y \in \mathbb{Q}$, we have $\nu_p(x+y) \geq \min\{\nu_p(x), \nu_p(y)\}$, and $\nu_p(x+y) = \nu_p(x)$ if $\nu_p(x) < \nu_p(y)$;

(d) for all non-zero $x \in \mathbb{Q}$, we have $x = \pm \prod_p p^{\nu_p(x)}$, where the product is over all primes, and all but a finite number of terms in the product are equal to 1;

(e) for all $x \in \mathbb{Q}$, we have $x \in \mathbb{Z}$ if and only if $\nu_p(x) \geq 0$ for all primes p.

EXERCISE 1.30. Let n be a positive integer, and let 2^k be the highest power of 2 in the set $S := \{1, \ldots, n\}$. Show that 2^k does not divide any other element in S.

EXERCISE 1.31. Let $n \in \mathbb{Z}$ with $n > 1$. Show that $\sum_{i=1}^n 1/i$ is not an integer.

EXERCISE 1.32. Let n be a positive integer, and let C_n denote the number of pairs of integers (a, b) with $a, b \in \{1, \ldots, n\}$ and $\gcd(a, b) = 1$, and let F_n be the number of *distinct* rational numbers a/b, where $0 \leq a < b \leq n$.

(a) Show that $F_n = (C_n + 1)/2$.

(b) Show that $C_n \geq n^2/4$. Hint: first show that $C_n \geq n^2(1 - \sum_{d \geq 2} 1/d^2)$, and then show that $\sum_{d \geq 2} 1/d^2 \leq 3/4$.

Chapter 2

Congruences

This chapter introduces the basic properties of congruences modulo n, along with the related notion of residue classes modulo n. Other items discussed include the Chinese remainder theorem and Fermat's little theorem.

2.1 Definitions and basic properties of congruences

Let n be a positive integer. For integers a and b, we say that a **is congruent to** b **modulo** n if $n \mid (a-b)$, and we write $a \equiv b \pmod{n}$. If $n \nmid (a-b)$, then we write $a \not\equiv b \pmod{n}$. Equivalently, $a \equiv b \pmod{n}$ if and only if $a = b + ny$ for some $y \in \mathbb{Z}$. The relation $a \equiv b \pmod{n}$ is called a **congruence relation**, or simply, a **congruence**. The number n appearing in such congruences is called the **modulus** of the congruence. Note that $a \equiv 0 \pmod{n}$ if and only if n divides a.

If we view the modulus n as fixed, then the following theorem says that the binary relation "$\cdot \equiv \cdot \pmod{n}$" is an equivalence relation on the set $\mathbb{Z}$.

Theorem 2.1 (Equivalence Property). *Let n be a positive integer. For all $a, b, c \in \mathbb{Z}$, we have:*

(i) $a \equiv a \pmod{n}$*;*

(ii) $a \equiv b \pmod{n}$ *implies* $b \equiv a \pmod{n}$*;*

(iii) $a \equiv b \pmod{n}$ *and* $b \equiv c \pmod{n}$ *implies* $a \equiv c \pmod{n}$.

Proof. For (i), observe that n divides $0 = a - a$. For (ii), observe that if n divides $a - b$, then it also divides $-(a-b) = b - a$. For (iii), observe that if n divides $a - b$ and $b - c$, then it also divides $(a-b) + (b-c) = a - c$. $\square$

Another key property of congruences is that they are "compatible" with integer addition and multiplication, in the following sense:

Theorem 2.2 (Compatibility Property). *Let $a, a', b, b', n \in \mathbb{Z}$ with $n > 0$. If*

$$a \equiv a' \pmod{n} \quad \textit{and} \quad b \equiv b' \pmod{n},$$

then

$$a + b \equiv a' + b' \pmod{n} \quad \textit{and} \quad a \cdot b \equiv a' \cdot b' \pmod{n}.$$

Proof. Suppose that $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$. This means that there exist integers x and y such that $a = a' + nx$ and $b = b' + ny$. Therefore,

$$a + b = a' + b' + n(x + y),$$

which proves the first congruence of the theorem, and

$$ab = (a' + nx)(b' + ny) = a'b' + n(a'y + b'x + nxy),$$

which proves the second congruence. □

Theorems 1.3 and 1.4 can be restated in terms of congruences (with a as given, and $b := n$):

Theorem 2.3. *Let $a, n \in \mathbb{Z}$ with $n > 0$. Then there exists a unique integer r such that $a \equiv r \pmod{n}$ and $0 \leq r < n$, namely, $r := a \bmod n$. More generally, for every $x \in \mathbb{R}$, there exists a unique integer $r \in [x, x + n)$ such that $a \equiv r \pmod{n}$.*

We also have:

Theorem 2.4. *Let $a, b, n \in \mathbb{Z}$ with $n > 0$. Then we have:*

$$a \equiv b \pmod{n} \iff (a \bmod n) = (b \bmod n). \tag{2.1}$$

Proof. By the existence part of Theorem 2.3, we have $a \equiv r \pmod{n}$ and $b \equiv s \pmod{n}$, where $r := a \bmod n$ and $s := b \bmod n$. We want to show $a \equiv b \pmod{n} \iff r = s$.

On the one hand, if $r = s$, then (using Theorem 2.1) we have $a \equiv b \pmod{n}$. On the other hand, if $a \equiv b \pmod{n}$, then (again using Theorem 2.1), we have $a \equiv s \pmod{n}$; moreover, by the uniqueness part of Theorem 2.3, we have $r = s$. □

Note that we are using the notation "mod" in two different ways here: on the left-hand side of (2.1), as a part of a congruence relation (e.g., $17 \equiv -3 \pmod{5}$), and on the right-hand side, as a binary operator (e.g., $(17 \bmod 5) = 2 = (-3 \bmod m)$). The reader should be aware that these two uses are similar, but not quite the same.

Theorems 2.1 and 2.2 are deceptively powerful: they allow one to work with congruence relations modulo n much as one would with ordinary equalities. One can add to, subtract from, or multiply both sides of a congruence modulo n by the same integer. Also, we have the following **substitution principle**: roughly speaking, if $a \equiv a' \pmod{n}$, we can substitute a' for a in any arithmetic expression involving a, without changing the value of the expression mod n. More precisely, suppose $E(v)$ is any arithmetic expression in a variable v, built up from v and integer constants using addition, subtraction, and multiplication operators; then $a \equiv a' \pmod{n}$ implies $E(a) \equiv E(a') \pmod{n}$.

To see why this "substitution principle" works, consider the expression

$$E(v) := (v + 1)(v + 2).$$

Suppose $a \equiv a' \pmod{n}$. We want to show: $E(a) \equiv E(a') \pmod{n}$. Appying Theorem 2.2 three times, we have:

$$\begin{aligned} a + 1 &\equiv a' + 1 \pmod{n} \\ a + 2 &\equiv a' + 2 \pmod{n} \\ (a + 1)(a + 2) &\equiv (a' + 1)(a' + 2) \pmod{n}. \end{aligned}$$

This type of calculation generalizes to arbitrary expressions.

Example 2.1. We can use the above principles to greatly simplify modular computations. Suppose we wish to compute $4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \bmod 17$. We could do this by first computing $a := 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9$, and then divide a by 17 to get $a \bmod 17$. However, a itself will be a rather large number, and we can simplify the computation by reducing intermediate results mod 17 as we go. Observe that $4 \cdot 5 = 20$, and $20 \equiv 3 \pmod{17}$. Therefore, using the above "substitution principle", we can replace $4 \cdot 5$ by 3 in the expression $4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9$, without changing its value mod 17:

$$\underline{4 \cdot 5} \cdot 6 \cdot 7 \cdot 8 \cdot 9 \equiv \underline{3} \cdot 6 \cdot 7 \cdot 8 \cdot 9 \pmod{17}.$$

Applying this substitution strategy repeatedly, we can compute:

$$\begin{aligned}
&\underbrace{4 \cdot 5}_{4 \cdot 5 \equiv 3 \pmod{17}} \cdot 6 \cdot 7 \cdot 8 \cdot 9 \\
&\underbrace{3 \cdot 6}_{3 \cdot 6 \equiv 1 \pmod{17}} \cdot 7 \cdot 8 \cdot 9 \\
&1 \cdot 7 \cdot 8 \cdot 9 \\
&\underbrace{7 \cdot 8}_{7 \cdot 8 \equiv 5 \pmod{17}} \cdot 9 \\
&\underbrace{5 \cdot 9}_{5 \cdot 9 \equiv 11 \pmod{17}}
\end{aligned}$$

We conclude that $4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \equiv 11 \pmod{17}$, or in other words, $4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \bmod 17 = 11$. □

Example 2.2. A convenient "rule of thumb" that one often uses to test for divisibility by 3 is to add up the digits of a number, and test if the the sum of digits is itself divisible by 3. For example, if $a = 25614$, we add the digits of a, obtaining $2 + 5 + 6 + 1 + 4 = 18$; since 18 is divisible by 3, we conclude that a is divisible by 3 as well. We can justify this rule using congruences. Let a be a positive integer whose base-10 representation is $a = (a_{k-1} \cdots a_1 a_0)_{10}$, so $a = \sum_{i=0}^{k-1} a_i 10^i$. Let b be the sum of the decimal digits of a; that is, let $b := \sum_{i=0}^{k-1} a_i$. We will show that $a \equiv b \pmod{3}$. This will justify the divisibility-by-3 rule, since then we see that $a \equiv 0 \pmod{3}$ (i.e., a is divisible by 3) if and only if $b \equiv 0 \pmod{3}$ (i.e., b is divisible by 3). To show that $a \equiv b \pmod{3}$, we first observe that $10 \equiv 1 \pmod{3}$. Then, we calculate

$$a = \sum_{i=0}^{k-1} a_i 10^i \equiv \sum_{i=0}^{k-1} a_i \cdot 1 \pmod{3}.$$

Here, we have used the above "substitution principle": since, $10 \equiv 1 \pmod{3}$, we can substitute 1 for 10 in the above congruence mod 3. See also Exercises 2.5 and 2.6. □

Example 2.3. Let us find the set of solutions z to the congruence

$$8z - 4 \equiv 5z - 2 \pmod{7}. \tag{2.2}$$

Suppose z is a solution to (2.2). If we add 4 and subtract $5z$ on both sides of (2.2), we see that z is a solution to

$$3z \equiv 2 \pmod{7}. \tag{2.3}$$

Conversely, if z is a solution to (2.3), then by subtracting 4 and adding $5z$ on both sides of (2.3), we see that z is a solution to (2.2). Thus, (2.2) and (2.3) have the same solution set.

Next, suppose z is a solution to (2.3). We would like to divide both sides of (2.3) by 3, to get z by itself on the left-hand side. We cannot do this directly, but since $5 \cdot 3 = 15 \equiv 1 \pmod 7$, we can achieve the same effect by multiplying both sides of (2.3) by 5. If we do this, we obtain

$$15z \equiv 10 \pmod 7. \tag{2.4}$$

Now, we apply the above "substitution principle": since, $15 \equiv 1 \pmod 7$, we can substitute 1 for 15 in (2.4). Likewise, since $10 \equiv 3 \pmod 7$, we can substitute 3 for 10. Making both of these substitutions, we obtain

$$z \equiv 3 \pmod 7. \tag{2.5}$$

Thus, any solution z to (2.3) is a solution to (2.5). Conversely, if z is a solution to (2.5), then by multiplying both sides of (2.5) by 3, we get $3z \equiv 9 \pmod 7$, and since $9 \equiv 2 \pmod 7$, we see that z is a solution to (2.3). Thus, (2.2), (2.3), and (2.5) all have the same solution set.

Therefore, z is a solution to (2.2) if and only if $z \equiv 3 \pmod 7$. That is, the solutions to (2.2) are precisely those integers that are congruent to 3 modulo 7, which we can list as follows:

$$\ldots, -18, -11, -4, 3, 10, 17, 24, \ldots \quad \square$$

In the next section, we shall give a systematic treatment of the problem of solving linear congruences, such as the one appearing in the previous example.

2.1.1 Application: computer arithmetic

Computers store integers as words, which are binary strings of fixed length. Congruences allow one to better understand how arithmetic on these words actually works.

Suppose we have a very minimalistic computer with just 4-bit words. On such a machine, the type `unsigned` in the `C` language would represent the integers $0, \ldots, 15$, encoded in binary as follows:

0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

When one adds or subtracts two such 4-bit integers, the result is always a 4-bit integer: the hardware just ignores any carries out of or borrows into the high-order bit position. For example, if we add the numbers 9 and 13, we get 22, which in binary is the 5-bit number 10110. The computer hardware will just throw away the left-most bit, resulting in the 4-bit number 0110, which is the binary representation of 6. Thus, on this machine, if we add 9 and 13, we get 6 instead of 22. Notice that while 6 is not equal to 22, it is congruent to 22 modulo 16, i.e., $6 \equiv 22 \pmod{16}$. This is no coincidence. When the hardware throws away the high-order bit, this is the same as subtracting off the value 16.

In general, if our computer has n-bit words, then such words can represent the numbers $0, \ldots, 2^n - 1$, encoded in binary as n-bit strings. Moreover, unsigned arithmetic (addition, subtraction, and multiplication) is really just arithmetic mod 2^n.

As for signed arithmetic, if the machine uses 2's complement arithmetic (as is the case for essentially all computers used today), then nothing really changes. Here again is the table for encoding 4-bit integers, where we also include the 2's complement encoding of signed integers:

0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	2	3	4	5	6	7	-8	-7	-6	-5	-4	-3	-2	-1

Note that $15 \equiv -1 \pmod{16}$, and $14 \equiv -2 \pmod{16}$, and so on. This is also not a coincidence. More generally, for n-bit 2's complement arithmetic, for each $x \in [1 \,..\, 2^{n-1}]$, we encode the negative integer $-x$ as the binary encoding of the positive integer $(-x) \bmod 2^n$. In particular, -1 is encoded as the binary encoding of $(-1) \bmod 2^n = 2^n - 1$, which is the bit string $11 \cdots 1$ of all 1's. Addition, subtraction, and multiplication of n-bit 2's complement integers is carried out using exactly the same hardware as for n-bit unsigned integers: all results are just computed mod 2^n.

Consider again our 4-bit machine, and suppose the type `int` represents the integers $-8, -7, \ldots, 6, 7$ in 2's complement. Suppose the `a`, `b`, `c`, `d` are variable of type `int` whose values are

```
a=3, b=4, c=4, d=5
```

and our `C` program computes

```
int e = a*b-c*d;
```

Technically speaking, the evaluation of `a*b-c*d` overflows and the result is not well defined. Nevertheless, it is more likely than not that the program actually assigns to `e` the correct value, which is $-8 = 3 \cdot 4 - 4 \cdot 5$. The reason is that the hardware will compute the result mod 16, and since the correct result itself fits in the range $-8, -7, \ldots, 6, 7$, the computed result and the correct result must match.[1]

See Exercise 2.7 for more on 2's complement arithmetic.

EXERCISE 2.1. Let $a, b, n \in \mathbb{Z}$ with $n > 0$. Show that $a \equiv b \pmod{n}$ if and only if $(a \bmod n) = (b \bmod n)$.

EXERCISE 2.2. Let $a, b, n, n' \in \mathbb{Z}$ with $n > 0$, $n' > 0$, and $n' \mid n$. Show that if $a \equiv b \pmod{n}$, then $a \equiv b \pmod{n'}$.

EXERCISE 2.3. Let $a, b, n, n' \in \mathbb{Z}$ with $n > 0$, $n' > 0$, and $\gcd(n, n') = 1$. Show that if $a \equiv b \pmod{n}$ and $a \equiv b \pmod{n'}$, then $a \equiv b \pmod{nn'}$.

EXERCISE 2.4. Let $a, b, n \in \mathbb{Z}$ with $n > 0$ and $a \equiv b \pmod{n}$. Show that $\gcd(a, n) = \gcd(b, n)$.

EXERCISE 2.5. Consider again Example 2.2. Instead of using the "subsitution principle", prove that

$$(a_{k-1} \cdots a_1 a_0)_{10} \equiv a_{k-1} + \cdots + a_1 + a_0 \pmod{3}$$

directly, using Theorem 2.2 and a proof by induction on k.
Hint: use the fact that for $k > 1$, we have

$$(a_{k-1} \cdots a_1 a_0)_{10} = 10 \cdot (a_{k-1} \cdots a_1)_{10} + a_0.$$

Note: the "substitution principle" itself can be proved by a similar induction argument.

EXERCISE 2.6. Analogous to Example 2.2, formulate and justify a simple "rule of thumb" for testing divisibility by 11.

[1] Warning: an optimizing compiler may invalidate our assumption that arithmetic is computed mod modulo 2^n. However, such aggressively optimizing compilers have caused problems in legacy code, and so some compromises have been made; for example, the `-fwrapv` option of `gcc` ensures arithmetic is always carried out mod 2^n.

Exercise 2.7. Let n be a positive integer. For $x \in \{0, \ldots, 2^n - 1\}$, let $\tilde{x}$ denote the integer obtained by inverting the bits in the n-bit, binary representation of x (note that $\tilde{x} \in \{0, \ldots, 2^n - 1\}$). Show that $\tilde{x} + 1 \equiv -x \pmod{2^n}$. This justifies the usual rule for computing negatives in 2's complement arithmetic.
Hint: what is $x + \tilde{x}$?

Exercise 2.8. Show that the equation $7y^3 + 2 = z^3$ has no solutions $y, z \in \mathbb{Z}$.

Exercise 2.9. Show that there are 14 distinct, possible, yearly (Gregorian) calendars, and show that all 14 calendars actually occur.

2.2 Solving linear congruences

In this section, we consider the general problem of solving linear congruences. More precisely, for a given positive integer n, and arbitrary integers a and b, we wish to determine the set of integers z that satisfy the congruence

$$az \equiv b \pmod{n}. \tag{2.6}$$

2.2.1 Existence of solutions

We begin with a theorem that characterizes when (2.6) has any solutions at all.

Theorem 2.5. *Let $a, b, n \in \mathbb{Z}$ with $n > 0$, and let $d := \gcd(a, n)$. Then (2.6) has a solution $z \in \mathbb{Z}$ if and only if $d \mid b$.*

Proof. We have

$$\begin{aligned} & az \equiv b \pmod{n} \text{ for some } z \in \mathbb{Z} \\ \iff & az = b + ny \text{ for some } z, y \in \mathbb{Z} \text{ (by def'n of congruence)} \\ \iff & az - ny = b \text{ for some } z, y \in \mathbb{Z} \\ \iff & d \mid b \text{ (by Bezout's Lemma).} \quad \square \end{aligned}$$

2.2.2 Uniqueness of solutions: cancellation and modular inverses

Now suppose z satisfies the congruence (2.6). Clearly, all integers $z' \equiv z \pmod{n}$ also satisfy (2.6). The question we next address is: are there any other solutions? The next theorem says that the answer is "no", provided $\gcd(a, n) = 1$. In other words, if a and n are relatively prime, then (2.6) has a unique solution modulo n.

Theorem 2.6. *Let $a, b, n \in \mathbb{Z}$ with $n > 0$. Assume that $\gcd(a, n) = 1$. For any solutions z and z' to (2.6), we have $z \equiv z' \pmod{n}$.*

Proof. We have

$$\begin{aligned} & az \equiv b \pmod{n} \text{ and } az' \equiv b \pmod{n} \\ & \quad \implies az \equiv az' \pmod{n} \\ & \quad \implies a(z - z') \equiv 0 \pmod{n} \\ & \quad \implies n \mid a(z - z') \\ & \quad \implies n \mid (z - z') \text{ (by Theorem 1.8)} \\ & \quad \implies z \equiv z' \pmod{n}. \quad \square \end{aligned}$$

Combining Theorems 2.5 and 2.6, we have:

Theorem 2.7. *Let $a, b, n \in \mathbb{Z}$ with $n > 0$. Assume that $\gcd(a, n) = 1$. The congruence (2.6) has a unique solution modulo n; that is, there is a unique $z \in \{0, \ldots, n-1\}$ such that $az \equiv b \pmod{n}$.*

A cancellation law. Let $a, n \in \mathbb{Z}$ with $n > 0$. The proof of Theorem 2.6 gives us a **cancellation law** for congruences:

$$\text{if } \gcd(a, n) = 1 \text{ and } az \equiv az' \pmod{n}, \text{ then } z \equiv z' \pmod{n}.$$

Modular inverses. Again, let $a, n \in \mathbb{Z}$ with $n > 0$. We say that $z \in \mathbb{Z}$ is a **multiplicative inverse of a modulo n** if $az \equiv 1 \pmod{n}$. Theorem 2.5 says that a has a multiplicative inverse modulo n if and only if $\gcd(a, n) = 1$. Moreover, Theorem 2.6 says that the multiplicative inverse of a, if it exists, is uniquely determined modulo n; that is, if z and z' are multiplicative inverses of a modulo n, then $z \equiv z' \pmod{n}$.

Notation: We write $a^{-1} \bmod n$ to denote the unique multiplicative inverse of a modulo n that lies in the interval $\{0, \ldots, n-1\}$.

Example 2.4. As we saw in Example 2.3, we have $3^{-1} \bmod 7 = 5$, since $3 \cdot 5 = 15 \equiv 1 \pmod{7}$. □

Computing modular inverses. We can use the extended Euclidean algorithm, discussed in §1.3.1, to compute modular inverses, when they exist. Suppose we are given a and n, and want to compute $a^{-1} \bmod n$. Let use assume that $0 \leq a < n$—we can always replace a by $a \bmod n$, if this is not the case. Now compute

$$(d, s, t) \leftarrow \text{ExtEuclid}(n, a),$$

so that $d = \gcd(n, a)$ and s and t satisfy $ns + at = d$. If $d \neq 1$, the $a^{-1} \bmod n$ does not exist; otherwise, $a^{-1} \bmod n = t \bmod n$.

2.2.3 Determining all solutions via modular inverses

We now describe the complete set of solutions $z \in \mathbb{Z}$ to the congruence (2.6) by means of modular inverses.

If $\gcd(a, n) = 1$, then setting $t := a^{-1} \bmod n$, and $z := tb \bmod n$, we see that this z is the unique solution to the congruence (2.6) that lies in the interval $\{0, \ldots, n-1\}$. The set of all solutions to the congruence (2.6) over the integers is $\{z + ny : y \in \mathbb{Z}\}$, but z is the only one of these in the interval $\{0, \ldots, n-1\}$.

More generally, let $d := \gcd(a, n)$. If $d \nmid b$, then we know that (2.6) does not have any solutions. So suppose that $d \mid b$, and set $a' := a/d$, $b' := b/d$, and $n' := n/d$.

For each $z \in \mathbb{Z}$, we have $az \equiv b \pmod{n}$ if and only if $a'z \equiv b' \pmod{n'}$. This is because

$$\begin{aligned} az \equiv b \pmod{n} &\iff az = b + ny \text{ for some } y \in \mathbb{Z} \\ &\iff (a/d) = (b/d) + (n/d)y \text{ for some } y \in \mathbb{Z} \\ &\iff a'z \equiv b' \pmod{n'}. \end{aligned}$$

Also, we have $\gcd(a', n') = 1$ (see Exercise 1.11). So, if we set $t := (a')^{-1} \bmod n'$ and $z' := tb' \bmod n'$, then z' is the unique solution to $a'z \equiv b' \pmod{n'}$ in the interval $\{0, \ldots, n'-1\}$. It follows that the solutions to the congruence (2.6) that lie in the interval $\{0, \ldots, n-1\}$ are the d integers $z', z'+n', \ldots, z'+(d-1)n'$. The set of all solutions to the congruence (2.6) over the integers is $\{z' + n'y : y \in \mathbb{Z}\}$, but these d integers are the only ones that lie in the interval $\{0, \ldots, n-1\}$.

We can summarize the above observations:

Theorem 2.8. *Let $a, b, n \in \mathbb{Z}$ with $n > 0$. Let $d := \gcd(a, n)$. If $d \nmid b$, then the congruence (2.6) has no solutions; otherwise, it has d solutions z in the interval $\{0, \ldots, n-1\}$, namely,*

$$z', z' + (n/d), \ldots, z' + (d-1)(n/d),$$

where z' is the unique integer in the interval $\{0, \ldots, (n/d) - 1\}$ such that

$$(a/d)z' \equiv (b/d) \pmod{n/d}.$$

Specifically, $z' = t(b/d) \bmod (n/d)$, where $t := (a/d)^{-1} \bmod (n/d)$.

Example 2.5. As a concrete illustration, suppose we want to find the solutions to the congruence $6z \equiv 22 \pmod{100}$. Observe that $\gcd(6, 100) = 2$ and 2 divides 22. It follows that z is a solution to $6z \equiv 22 \pmod{100}$ if and only if $3z \equiv 11 \pmod{50}$. Here, we have simply divided each of the numbers appearing in the first congruence by the greatest common divisor to obtain the second congruence. Since $3 \cdot 17 = 51 \equiv 1 \pmod{50}$, we see that $3^{-1} \bmod 50 = 17$. So we multiply both sides of $3z \equiv 11 \pmod{50}$ by 17 to obtain $z \equiv 187 \equiv 37 \pmod{50}$. It follows that the original conguence $6z \equiv 22 \pmod{100}$ has exactly two solutions in the interval $\{0, \ldots, 99\}$, namely, $z = 37$ and $z = 37 + 50 = 87$. □

EXERCISE 2.10. Let $a, n \in \mathbb{Z}$ with $n > 0$. Show that $z \in \mathbb{Z}$ satisfies the congruence $az \equiv 0 \pmod{n}$ iff $z \equiv 0 \pmod{n/d}$, where $d := \gcd(a, n)$.

EXERCISE 2.11. For each of the following congruences, determine all the integer solutions $z \in \{0, \ldots, 999\}$. To so this, you should first put the congruence in the form $az = b \pmod{1000}$, as we did in going from (2.2) to (2.3) in Example 2.3. Then follow the steps outlined in §2.2.3. Show all your steps (but you may use a calculator to help with the multiplications). Use the extended Euclidean algorithm to compute modular inverses, where necessary.

(a) $100z + 200 \equiv 93z + 171 \pmod{1000}$

(b) $115z + 130 \equiv 100z + 165 \pmod{1000}$

(c) $115z + 132 \equiv 100z + 140 \pmod{1000}$

(d) $119z + 132 \equiv 113z + 140 \pmod{1000}$

EXERCISE 2.12. Let $a_1, \ldots, a_k, b, n$ be integers with $n > 0$. Show that the congruence

$$a_1 z_1 + \cdots + a_k z_k \equiv b \pmod{n}$$

has a solution $z_1, \ldots, z_k \in \mathbb{Z}$ if and only if $d \mid b$, where $d := \gcd(a_1, \ldots, a_k, n)$.

EXERCISE 2.13. Let p be a prime, and let a, b, c, e be integers, such that $e > 0$, $a \not\equiv 0 \pmod{p^{e+1}}$, and $0 \leq c < p^e$. Define N to be the number of integers $z \in \{0, \ldots, p^{2e} - 1\}$ such that

$$\left\lfloor \left((az + b) \bmod p^{2e} \right) / p^e \right\rfloor = c.$$

Show that $N = p^e$.

2.3 The Chinese remainder theorem

Next, we consider systems of linear congruences with respect to moduli that are relatively prime in pairs. The result we state here is known as the Chinese remainder theorem, and is extremely useful in a number of contexts.

Theorem 2.9 (Chinese remainder theorem). *Let $\{n_i\}_{i=1}^k$ be a pairwise relatively prime family of positive integers, and let $a_1, \ldots, a_k$ be arbitrary integers. Then there exists a solution $a \in \mathbb{Z}$ to the system of congruences*

$$a \equiv a_i \pmod{n_i} \quad (i = 1, \ldots, k).$$

Moreover, any $a' \in \mathbb{Z}$ is a solution to this system of congruences if and only if $a \equiv a' \pmod{n}$, where $n := \prod_{i=1}^k n_i$.

Proof. To prove the existence of a solution a to the system of congruences, we first show how to construct integers $e_1, \ldots, e_k$ such that for $i, j = 1, \ldots, k$, we have

$$e_j \equiv \begin{cases} 1 \pmod{n_i} & \text{if } j = i, \\ 0 \pmod{n_i} & \text{if } j \neq i. \end{cases} \tag{2.7}$$

If we do this, then setting

$$a := \sum_{i=1}^k a_i e_i,$$

one sees that for $j = 1, \ldots, k$, we have

$$a \equiv \sum_{i=1}^k a_i e_i \equiv a_j \pmod{n_j},$$

since all the terms in this sum are zero modulo n_j, except for the term $i = j$, which is congruent to a_j modulo n_j.

To construct $e_1, \ldots, e_k$ satisfying (2.7), let $n := \prod_{i=1}^k n_i$ as in the statement of the theorem, and for $i = 1, \ldots, k$, let $n_i^* := n/n_i$; that is, n_i^* is the product of all the moduli n_j with $j \neq i$. From the fact that $\{n_i\}_{i=1}^k$ is pairwise relatively prime, it follows that for $i = 1, \ldots, k$, we have $\gcd(n_i, n_i^*) = 1$, and so we may define $t_i := (n_i^*)^{-1} \bmod n_i$ and $e_i := n_i^* t_i$. One sees that $e_i \equiv 1 \pmod{n_i}$, while for $j \neq i$, we have $n_i \mid n_j^*$, and so $e_j \equiv 0 \pmod{n_i}$. Thus, (2.7) is satisfied.

That proves the existence of a solution a to the given system of congruences. If $a \equiv a' \pmod{n}$, then since $n_i \mid n$ for $i = 1, \ldots, k$, we see that $a' \equiv a \equiv a_i \pmod{n_i}$ for $i = 1, \ldots, k$, and so a' also solves the system of congruences.

Finally, if a' is a solution to the given system of congruences, then $a \equiv a_i \equiv a' \pmod{n_i}$ for $i = 1, \ldots, k$. Thus, $n_i \mid (a - a')$ for $i = 1, \ldots, k$. Since $\{n_i\}_{i=1}^k$ is pairwise relatively prime, this implies $n \mid (a - a')$, or equivalently, $a \equiv a' \pmod{n}$. □

We can restate Theorem 2.9 in more concrete terms, as follows. For each positive integer m, let I_m denote $\{0, \ldots, m-1\}$. Suppose $\{n_i\}_{i=1}^k$ is a pairwise relatively prime family of positive integers, and set $n := n_1 \cdots n_k$. Then the map

$$\begin{aligned} \tau : \quad & I_n \to I_{n_1} \times \cdots \times I_{n_k} \\ & a \mapsto (a \bmod n_1, \ldots, a \bmod n_k) \end{aligned}$$

is a bijection.

Example 2.6. The following table illustrates what Theorem 2.9 says for $n_1 = 3$ and $n_2 = 5$.

a	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$a \bmod 3$	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
$a \bmod 5$	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

We see that as a ranges from 0 to 14, the pairs $(a \bmod 3, a \bmod 5)$ range over all pairs (a_1, a_2) with $a_1 \in \{0, 1, 2\}$ and $a_2 \in \{0, \ldots, 4\}$, with every pair being hit exactly once. □

Exercise 2.14. Compute the values e_1, e_2, e_3 in the proof of Theorem 2.9 in the case where $k = 3$, $n_1 = 3$, $n_2 = 5$, and $n_3 = 7$. Also, find an integer a such that $a \equiv 1 \pmod 3$, $a \equiv -1 \pmod 5$, and $a \equiv 5 \pmod 7$.

Exercise 2.15. If you want to show that you are a real nerd, here is an age-guessing game you might play at a party. You ask a fellow party-goer to divide his age by each of the numbers 3, 4, and 5, and tell you the remainders. Show how to use this information to determine his age.

Exercise 2.16. Let $\{n_i\}_{i=1}^k$ be a pairwise relatively prime family of positive integers. Let $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ be integers, and set $d_i := \gcd(a_i, n_i)$ for $i = 1, \ldots, k$. Show that there exists an integer z such that $a_i z \equiv b_i \pmod{n_i}$ for $i = 1, \ldots, k$ if and only if $d_i \mid b_i$ for $i = 1, \ldots, k$.

Exercise 2.17. For each prime p, let $\nu_p(\cdot)$ be defined as in §1.5. Let $p_1, \ldots, p_r$ be distinct primes, $a_1, \ldots, a_r$ be arbitrary integers, and $e_1, \ldots, e_r$ be arbitrary non-negative integers. Show that there exists an integer a such that $\nu_{p_i}(a - a_i) = e_i$ for $i = 1, \ldots, r$.

Exercise 2.18. Suppose n_1 and n_2 are positive integers, and let $d := \gcd(n_1, n_2)$. Let a_1 and a_2 be arbitrary integers. Show that there exists an integer a such that $a \equiv a_1 \pmod{n_1}$ and $a \equiv a_2 \pmod{n_2}$ if and only if $a_1 \equiv a_2 \pmod d$.

2.4 Residue classes

Let n be a positive integer. We define the set $\mathbb{Z}_n$ to be the set of n abstract objects

$$[0]_n, [1]_n, \ldots, [n-1]_n.$$

These objects are called the **residue classes modulo** n. We can extend this notation, where for arbitrary $a \in \mathbb{Z}$, we define

$$[a]_n := [a \bmod n]_n.$$

We also define addition and multiplication on residue classes as follows. For $a, b \in \{0, \ldots, n-1\}$, we define

$$[a]_n + [b]_n := [a + b]_n \qquad \text{and} \qquad [a]_n \cdot [b]_n := [a \cdot b]_n,$$

Note that in this definition, we are using the extended notation introduced above, so that $[a+b]_n = [(a + b) \bmod n]_n$ and $[a \cdot b]_n = [(a \cdot b) \bmod n]_n$.

Note that because of Theorem 2.2, the equations defining addition and multiplication of residue classes hold for all $a, b \in \mathbb{Z}$. That is, for all $a, b \in \mathbb{Z}$, we have

$$[a]_n + [b]_n = [a + b]_n \qquad \text{and} \qquad [a]_n \cdot [b]_n = [a \cdot b]_n.$$

To see why, if $a' := a \bmod n$ and $b' := b \bmod n$, then

$$[a]_n = [a']_n, \quad [b]_n = [b']_n, \quad [a+b]_n = [a'+b']_n, \quad \text{and} \quad [a \cdot b]_n = [a' \cdot b']_n,$$

and it follows that

$$[a]_n + [b]_n = [a']_n + [b']_n = [a'+b']_n = [a+b]_n$$

and

$$[a]_n \cdot [b]_n = [a']_n \cdot [b']_n = [a' \cdot b']_n = [a \cdot b]_n.$$

More generally, for all $a, b, c \in \mathbb{Z}$, we have

$$[a]_n + [b]_n = [c]_n \iff a + b \equiv c \pmod{n},$$

and

$$[a]_n \cdot [b]_n = [c]_n \iff a \cdot b \equiv c \pmod{n}.$$

When n is clear from context, we often write $[a]$ instead of $[a]_n$.

Example 2.7. Consider again the residue classes modulo 6. These are

$$[0], [1], [2], [3], [4], [5].$$

Using the extended notation, we see, for example, that $[-1]$ is the same thing as $[5]$.

Let us write down the addition and multiplication tables for $\mathbb{Z}_6$. The addition table looks like this:

$+$	$[0]$	$[1]$	$[2]$	$[3]$	$[4]$	$[5]$
$[0]$	$[0]$	$[1]$	$[2]$	$[3]$	$[4]$	$[5]$
$[1]$	$[1]$	$[2]$	$[3]$	$[4]$	$[5]$	$[0]$
$[2]$	$[2]$	$[3]$	$[4]$	$[5]$	$[0]$	$[1]$
$[3]$	$[3]$	$[4]$	$[5]$	$[0]$	$[1]$	$[2]$
$[4]$	$[4]$	$[5]$	$[0]$	$[1]$	$[2]$	$[3]$
$[5]$	$[5]$	$[0]$	$[1]$	$[2]$	$[3]$	$[4]$.

The multiplication table looks like this:

$\cdot$	$[0]$	$[1]$	$[2]$	$[3]$	$[4]$	$[5]$
$[0]$	$[0]$	$[0]$	$[0]$	$[0]$	$[0]$	$[0]$
$[1]$	$[0]$	$[1]$	$[2]$	$[3]$	$[4]$	$[5]$
$[2]$	$[0]$	$[2]$	$[4]$	$[0]$	$[2]$	$[4]$
$[3]$	$[0]$	$[3]$	$[0]$	$[3]$	$[0]$	$[3]$
$[4]$	$[0]$	$[4]$	$[2]$	$[0]$	$[4]$	$[2]$
$[5]$	$[0]$	$[5]$	$[4]$	$[3]$	$[2]$	$[1]$.

□

The addition and multiplication operations on $\mathbb{Z}_n$ yield a very natural algebraic structure. For example, addition and multiplication are commutative and associative; that is, for all $\alpha, \beta, \gamma \in \mathbb{Z}_n$, we have

$$\alpha + \beta = \beta + \alpha, \quad (\alpha + \beta) + \gamma = \alpha + (\beta + \gamma),$$
$$\alpha\beta = \beta\alpha, \quad (\alpha\beta)\gamma = \alpha(\beta\gamma).$$

Note that we have adopted here the usual convention of writing $\alpha\beta$ in place of $\alpha \cdot \beta$. Furthermore, multiplication distributes over addition; that is, for all $\alpha, \beta, \gamma \in \mathbb{Z}_n$, we have

$$\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma.$$

All of these properties follow from the definitions, and the corresponding properties for $\mathbb{Z}$. For example, for the distributive law, suppose $\alpha = [a]$, $\beta = [b]$, and $\gamma = [c]$. Then we have

$$\alpha(\beta + \gamma) = [a(b + c)] = [ab + ac] = \alpha\beta + \alpha\gamma.$$

Because addition and multiplication in $\mathbb{Z}_n$ are associative, for $\alpha_1, \ldots, \alpha_k \in \mathbb{Z}_n$, we may write the sum $\alpha_1 + \cdots + \alpha_k$ and the product $\alpha_1 \cdots \alpha_k$ without any parentheses, and there is no ambiguity; moreover, since both addition and multiplication are commutative, we may rearrange the terms in such sums and products without changing their values.

The residue class $[0]$ acts as an **additive identity**; that is, for all $\alpha \in \mathbb{Z}_n$, we have $\alpha + [0] = \alpha$; indeed, if $\alpha = [a]$, then $a + 0 \equiv a \pmod{n}$. Moreover, $[0]$ is the only element of $\mathbb{Z}_n$ that acts as an additive identity; indeed, if $a + z \equiv a \pmod{n}$ holds for all integers a, then it holds in particular for $a = 0$, which implies $z \equiv 0 \pmod{n}$. The residue class $[0]$ also has the property that $\alpha \cdot [0] = [0]$ for all $\alpha \in \mathbb{Z}_n$.

Every $\alpha \in \mathbb{Z}_n$ has an **additive inverse**, that is, an element $\beta \in \mathbb{Z}_n$ such that $\alpha + \beta = [0]$; indeed, if $\alpha = [a]$, then clearly $\beta := [-a]$ does the job, since $a + (-a) \equiv 0 \pmod{n}$. Moreover, α has a unique additive inverse; indeed, if $a + z \equiv 0 \pmod{n}$, then subtracting a from both sides of this congruence yields $z \equiv -a \pmod{n}$. We naturally denote the additive inverse of α by $-\alpha$. Observe that the additive inverse of $-\alpha$ is α; that is $-(-\alpha) = \alpha$. Also, we have the identities

$$-(\alpha + \beta) = (-\alpha) + (-\beta),\ (-\alpha)\beta = -(\alpha\beta) = \alpha(-\beta),\ (-\alpha)(-\beta) = \alpha\beta.$$

For $\alpha, \beta \in \mathbb{Z}_n$, we naturally write $\alpha - \beta$ for $\alpha + (-\beta)$.

The residue class $[1]$ acts as a **multiplicative identity**; that is, for all $\alpha \in \mathbb{Z}_n$, we have $\alpha \cdot [1] = \alpha$; indeed, if $\alpha = [a]$, then $a \cdot 1 \equiv a \pmod{n}$. Moreover, $[1]$ is the only element of $\mathbb{Z}_n$ that acts as a multiplicative identity; indeed, if $a \cdot z \equiv a \pmod{n}$ holds for all integers a, then in particular, it holds for $a = 1$, which implies $z \equiv 1 \pmod{n}$.

For $\alpha \in \mathbb{Z}_n$, we call $\beta \in \mathbb{Z}_n$ a **multiplicative inverse** of α if $\alpha\beta = [1]$. Not all $\alpha \in \mathbb{Z}_n$ have multiplicative inverses. If $\alpha = [a]$ and $\beta = [b]$, then β is a multiplicative inverse of α if and only if $ab \equiv 1 \pmod{n}$. The results in §2.2 imply that α has a multiplicative inverse if and only if $\gcd(a, n) = 1$, and that if it exists, it is unique. When it exists, we denote the multiplicative inverse of α by α^{-1}. We saw at the end of §2.2 how to compute modular inverses using the extended Euclidean algorithm.

We define $\mathbb{Z}_n^*$ to be the set of elements of $\mathbb{Z}_n$ that have a multiplicative inverse. By the above discussion, we have

$$\mathbb{Z}_n^* = \{[a] : a = 0, \ldots, n-1,\ \gcd(a, n) = 1\}.$$

If n is prime, then $\gcd(a, n) = 1$ for $a = 1, \ldots, n-1$, and we see that $\mathbb{Z}_n^* = \mathbb{Z}_n \setminus \{[0]\}$. If n is composite, then $\mathbb{Z}_n^* \subsetneq \mathbb{Z}_n \setminus \{[0]\}$; for example, if $d \mid n$ with $1 < d < n$, we see that $[d]$ is not zero, nor does it belong to $\mathbb{Z}_n^*$. Observe that if $\alpha, \beta \in \mathbb{Z}_n^*$, then so are α^{-1} and $\alpha\beta$; indeed,

$$(\alpha^{-1})^{-1} = \alpha \quad \text{and} \quad (\alpha\beta)^{-1} = \alpha^{-1}\beta^{-1}.$$

For $\alpha \in \mathbb{Z}_n$ and $\beta \in \mathbb{Z}_n^*$, we naturally write α/β for $\alpha\beta^{-1}$.

Example 2.8. We list the elements of $\mathbb{Z}_{15}^*$, and for each $\alpha \in \mathbb{Z}_{15}^*$, we also give α^{-1}:

α	[1]	[2]	[4]	[7]	[8]	[11]	[13]	[14]
α^{-1}	[1]	[8]	[4]	[13]	[2]	[11]	[7]	[14]

. □

Suppose α, β, γ are elements of $\mathbb{Z}_n$ that satisfy the equation

$$\alpha\beta = \alpha\gamma.$$

If $\alpha \in \mathbb{Z}_n^*$, we may multiply both sides of this equation by α^{-1} to infer that

$$\beta = \gamma.$$

This is the **cancellation law** for $\mathbb{Z}_n$. In particular, if n is prime, then this cancellation law holds for all $\alpha \neq [0]$.

We stress that for arbitrary n, we require that $\alpha \in \mathbb{Z}_n^*$, and not just $\alpha \neq [0]$. Indeed, suppose that n is composite, so we can factor it as $n = ab$, where $1 < a < n$ and $1 < b < n$, and set $\alpha := [a]$, $\beta := [b]$, and $\gamma := [0]$. Then we have $\alpha\beta = [0] = \alpha\gamma$; however, $\beta \neq \gamma$.

Analogous to Theorems 2.5 and 2.6, we have the following result on the **existence and uniqueness to solutions of linear equations in** $\mathbb{Z}_n$. Specifically, let $\alpha \in \mathbb{Z}_n^*$ and $\beta \in \mathbb{Z}_n$. Then the equation

$$\alpha\zeta = \beta$$

has a unique solution ζ, namely, $\zeta := \alpha^{-1}\beta$. In particular, if n is prime, then this equation has a unique solution ζ provided $\alpha \neq [0]$. Again, we stress that for arbitrary n, we require that $\alpha \in \mathbb{Z}_n^*$, and not just $\alpha \neq [0]$.

More notational conventions. For $\alpha_1, \ldots, \alpha_k \in \mathbb{Z}_n$, we may naturally write their sum as $\sum_{i=1}^k \alpha_i$. By convention, this sum is [0] when $k = 0$. It is easy to see that $-\sum_{i=1}^k \alpha_i = \sum_{i=1}^k (-\alpha_i)$; that is, the additive inverse of the sum is the sum of the additive inverses. In the special case where all the α_i's have the same value α, we define $k \cdot \alpha := \sum_{i=1}^k \alpha$; thus, $0 \cdot \alpha = [0]$, $1 \cdot \alpha = \alpha$, $2 \cdot \alpha = \alpha + \alpha$, $3 \cdot \alpha = \alpha + \alpha + \alpha$, and so on. The additive inverse of $k \cdot \alpha$ is $k \cdot (-\alpha)$, which we may also write as $(-k) \cdot \alpha$; thus, $(-1) \cdot \alpha = -\alpha$, $(-2) \cdot \alpha = (-\alpha) + (-\alpha) = -(\alpha + \alpha)$, and so on. Therefore, the notation $k \cdot \alpha$, or more simply, $k\alpha$, is defined for all integers k. Note that for all integers k and a, we have $k[a] = [ka] = [k][a]$.

Analogously, for $\alpha_1, \ldots, \alpha_k \in \mathbb{Z}_n$, we may write their product as $\prod_{i=1}^k \alpha_i$. By convention, this product is [1] when $k = 0$. It is easy to see that if all of the α_i's belong to $\mathbb{Z}_n^*$, then so does their product, and in particular, $(\prod_{i=1}^k \alpha_i)^{-1} = \prod_{i=1}^k \alpha_i^{-1}$; that is, the multiplicative inverse of the product is the product of the multiplicative inverses. In the special case where all the α_i's have the same value α, we define $\alpha^k := \prod_{i=1}^k \alpha$; thus, $\alpha^0 = [1]$, $\alpha^1 = \alpha$, $\alpha^2 = \alpha\alpha$, $\alpha^3 = \alpha\alpha\alpha$, and so on. If $\alpha \in \mathbb{Z}_n^*$, then the multiplicative inverse of α^k is $(\alpha^{-1})^k$, which we may also write as α^{-k}; for example, $\alpha^{-2} = \alpha^{-1}\alpha^{-1} = (\alpha\alpha)^{-1}$. Therefore, when $\alpha \in \mathbb{Z}_n^*$, the notation α^k is defined for all integers k.

One last notational convention. As already mentioned, when the modulus n is clear from context, we usually write $[a]$ instead of $[a]_n$. Although we want to maintain a clear distinction between integers and their residue classes, occasionally even the notation $[a]$ is not only redundant, but distracting; in such situations, we may simply write a instead of $[a]$. For example, for every $\alpha \in \mathbb{Z}_n$, we have the identity $(\alpha + [1]_n)(\alpha - [1]_n) = \alpha^2 - [1]_n$, which we may write more simply as $(\alpha+[1])(\alpha-[1]) = \alpha^2-[1]$, or even more simply, and hopefully more clearly, as $(\alpha+1)(\alpha-1) = \alpha^2-1$. Here, the only reasonable interpretation of the symbol "1" is [1], and so there can be no confusion.

In summary, algebraic expressions involving residue classes may be manipulated in much the same way as expressions involving ordinary numbers. Extra complications arise only because when n is composite, some non-zero elements of $\mathbb{Z}_n$ do not have multiplicative inverses, and the usual cancellation law does not apply for such elements.

In general, one has a choice between working with congruences modulo n, or with the algebraic structure $\mathbb{Z}_n$; ultimately, the choice is one of taste and convenience, and it depends on what one prefers to treat as "first class objects": integers and congruence relations, or elements of $\mathbb{Z}_n$.

EXERCISE 2.19. Let p be an odd prime. Show that $\sum_{\beta \in \mathbb{Z}_p^*} \beta^{-1} = \sum_{\beta \in \mathbb{Z}_p^*} \beta = 0$.

EXERCISE 2.20. Let p be an odd prime. Show that the numerator of $\sum_{i=1}^{p-1} 1/i$ is divisible by p. Hint: use the previous exercise.

2.5 Fermat's little theorem

Let n be a positive integer, and let $\alpha \in \mathbb{Z}_n^*$.

Consider the sequence of powers of α:

$$1 = \alpha^0, \alpha^1, \alpha^2, \ldots.$$

Since each such power is an element of $\mathbb{Z}_n^*$, and since $\mathbb{Z}_n^*$ is a finite set, this sequence of powers must start to repeat at some point; that is, there must be a positive integer k such that $\alpha^k = \alpha^i$ for some $i = 0, \ldots, k-1$. Let us assume that k is chosen to be the smallest such positive integer. This value k is called the **multiplicative order of** α.

We claim that $\alpha^k = 1$. To see this, suppose by way of contradiction that $\alpha^k = \alpha^i$, for some $i = 1, \ldots, k-1$; we could then cancel α from both sides of the equation $\alpha^k = \alpha^i$, obtaining $\alpha^{k-1} = \alpha^{i-1}$, which would contradict the minimality of k.

Thus, we can characterize the multiplicative order of α as the smallest positive integer k such that

$$\alpha^k = 1.$$

If $\alpha = [a]$ with $a \in \mathbb{Z}$ (and $\gcd(a, n) = 1$, since $\alpha \in \mathbb{Z}_n^*$), then k is also called the **multiplicative order of** a **modulo** n, and can be characterized as the smallest positive integer k such that

$$a^k \equiv 1 \pmod{n}.$$

From the above discussion, we see that the first k powers of α, that is, $\alpha^0, \alpha^1, \ldots, \alpha^{k-1}$, are distinct. Moreover, other powers of α simply repeat this pattern. The following is an immediate consequence of this observation.

Theorem 2.10. *Let n be a positive integer, and let α be an element of $\mathbb{Z}_n^*$ of multiplicative order k. Then for every $i \in \mathbb{Z}$, we have $\alpha^i = 1$ if and only if k divides i. More generally, for all $i, j \in \mathbb{Z}$, we have $\alpha^i = \alpha^j$ if and only if $i \equiv j \pmod{k}$.*

Example 2.9. Let $n = 7$. For each value $a = 1, \ldots, 6$, we can compute successive powers of a modulo n to find its multiplicative order modulo n.

i	1	2	3	4	5	6
$1^i \bmod 7$	1	1	1	1	1	1
$2^i \bmod 7$	2	4	1	2	4	1
$3^i \bmod 7$	3	2	6	4	5	1
$4^i \bmod 7$	4	2	1	4	2	1
$5^i \bmod 7$	5	4	6	2	3	1
$6^i \bmod 7$	6	1	6	1	6	1

So we conclude that modulo 7: 1 has order 1; 6 has order 2; 2 and 4 have order 3; and 3 and 5 have order 6. $\square$

In the above example, we see that every element of $\mathbb{Z}_7^*$ has multiplicative order either 1, 2, 3, or 6. In particular, $\alpha^6 = 1$ for all $\alpha \in \mathbb{Z}_7^*$. This is a special case of Fermat's little theorem:

Theorem 2.11 (Fermat's little theorem). *Let p be a prime and let $\alpha \in \mathbb{Z}_p^*$. Then $\alpha^{p-1} = 1$.*

Proof. Since $\alpha \in \mathbb{Z}_p^*$, for every $\beta \in \mathbb{Z}_p^*$ we have $\alpha\beta \in \mathbb{Z}_p^*$, and so we may define the "multiplication by α" map

$$\begin{aligned} \tau_\alpha : \quad & \mathbb{Z}_p^* \to \mathbb{Z}_p^* \\ & \beta \mapsto \alpha\beta. \end{aligned}$$

It is easy to see that τ_α is a bijection:

Injectivity: If $\alpha\beta = \alpha\beta'$, then cancel α to obtain $\beta = \beta'$.

Surjectivity: For every $\gamma \in \mathbb{Z}_p^*$, $\alpha^{-1}\gamma$ is a pre-image of γ under τ_α.

Thus, as β ranges over the set $\mathbb{Z}_p^*$, so does $\alpha\beta$, and we have

$$\prod_{\beta \in \mathbb{Z}_p^*} \beta = \prod_{\beta \in \mathbb{Z}_p^*} (\alpha\beta) = \alpha^{p-1} \Big(\prod_{\beta \in \mathbb{Z}_p^*} \beta \Big). \tag{2.8}$$

Canceling the common factor $\prod_{\beta \in \mathbb{Z}_p^*} \beta \in \mathbb{Z}_p^*$ from the left- and right-hand side of (2.8), we obtain

$$1 = \alpha^{p-1}. \ \square$$

As a consequence of Fermat's little theorem and Theorem 2.10, we see that for a prime p, the multiplicative order of α divides $p-1$ for every $\alpha \in \mathbb{Z}_p^*$. It turns out that for every prime p, there exists an element $\alpha \in \mathbb{Z}_p^*$ of whose multiplicative order is equal to $p-1$. Such an α is called a **primitive root mod** p. For instance, in Example 2.9, we saw that $[3]_7$ and $[5]_7$ are primitive roots mod 7. We shall prove this fact later (see Theorem 3.11), after developing some other tools. Observe that if $\alpha \in \mathbb{Z}_p^*$ is a primitive root, then every $\beta \in \mathbb{Z}_p^*$ can be expressed as a power of α.

Fermat's little theorem is sometimes stated in the following form: for every prime p and every $\alpha \in \mathbb{Z}_p$, we have

$$\alpha^p = \alpha. \tag{2.9}$$

Observe that for $\alpha = 0$, the equation (2.9) obviously holds. Otherwise, for $\alpha \neq 0$, Theorem 2.11 says that $\alpha^{p-1} = 1$, and multiplying both sides of this equation by α yields (2.9).

Finally, in terms of congruences, Fermat's little theorem can be stated as follows: for every prime p and every integer a, we have

$$a^p \equiv a \pmod{p}. \tag{2.10}$$

2.5.1 Application: primality testing

Fermat's little theorem forms the basis for several primality testing algorithms.

Consider the following computational problem: given a large number n, decide whether n is prime or not. Here, we are thinking of n has being *very* large, perhaps several hundred decimal digits in length—these are the size of prime numbers needed in a number of cryptographic applications.

A naive approach to testing if n is prime is trial division: simply test if n is divisible by 2, 3, 5, etc., testing divisibility by all primes p up to $n^{1/2}$ (see Exercise 1.14). Unfortunately, for the large values of n we are considering here, this would take an enormous amount of time, and is completely impractical.

Fermat's little theorem suggests the following primality test: simply select a non-zero $\alpha \in \mathbb{Z}_n$, compute $\beta := \alpha^{n-1} \in \mathbb{Z}_n$, and check if $\beta = 1$. On the one hand, if $\beta \neq 1$, Fermat's little theorem tells us that n cannot be prime. On the other hand, if $\beta = 1$, the test is inconclusive: n may or may not prime.

We can repeat the above test several times. If any of the β's are not 1, we know n is not prime; otherwise, the test is still inconclusive.

While the Fermat primality test is not perfect, it is not hard to modify it slightly so that it becomes much more effective—the most practical primality tests used today are all minor variations on the Fermat primality test. We shall not go into the details of these variations. Rather, we shall show how to efficiently implement the Fermat primality test. The techniques we present apply to the more effective primality tests, and have many other applications.

The first issue to be addressed is how to represent elements of $\mathbb{Z}_n$. It is natural and convenient to work with the set of representatives $\{0, \ldots, n-1\}$. So to multiply two elements in $\mathbb{Z}_n$, we multiply their representatives, and then reduce the product mod n. Just as in §1.3.1, we shall assume that we have efficient algorithms to implement these basic arithmetic operations. That still leaves the issue of how to efficiently perform the exponentiation α^{n-1}.

A simple algorithm to compute $\beta := \alpha^{n-1}$ is the following:

$\beta \leftarrow [1]_n \in \mathbb{Z}_n$
repeat $n-1$ times
    $\beta \leftarrow \beta \cdot \alpha$

This algorithm computes the value β correctly. Moreover, the numbers that arise in the computation never get too large, since in every multiplication in $\mathbb{Z}_n$, the result gets reduced mod n. Unfortunately, the number of loop iterations is $n-1$, and so this algorithm is actually slower than the trial division primality test that we started out with.

Fortunately, there is a much faster exponentiation algorithm, called **repeated squaring**. Consider the following more general problem: given $\alpha \in \mathbb{Z}_n$ and a non-negative integer e, compute α^e. Using the repeated squaring algorithm, we can compute α^e using $O(\log e)$ multiplications in $\mathbb{Z}_n$. Setting $e := n-1$, we can therefore implement the Fermat primality test using $O(\log n)$ multiplications in $\mathbb{Z}_n$, which is much more practical.

As a warmup, suppose e is a power of 2, say $e = 2^k$. Then the following simple algorithm computes $\beta := \alpha^e$:

$\beta \leftarrow \alpha$
repeat k times
    $\beta \leftarrow \beta^2$

This algorithm requires just $k = \log_2 e$ multiplications in $\mathbb{Z}_n$.

For arbitrary e, we can use the following strategy. Suppose that the binary representation of e is $e = (b_1 \cdots b_k)_2$, where b_1 is the high-order bit of e and b_k is the low-order bit. Then the following iterative algorithm computes $\beta := \alpha^e$:

$\beta \leftarrow [1]_n \in \mathbb{Z}_n$
for i in $[1 \,..\, k]$ do
    $\beta \leftarrow \beta^2$
    if $b_i = 1$ then $\beta \leftarrow \beta \cdot \alpha$

One can easily verify that after the ith loop iteration, we have $\beta = \alpha^{e_i}$, where $e_i = (b_1 \cdots b_i)_2$. Indeed, observe that $e_i = 2e_{i-1} + b_i$, and therefore,

$$\alpha^{e_i} = \alpha^{2e_{i-1}+b_i} = (\alpha^{e_{i-1}})^2 \cdot \alpha^{b_i}.$$

Example 2.10. Suppose $e = 37 = (100101)_2$. The above algorithm performs the following operations in this case:

	// computed exponent (in binary)
$\beta \leftarrow [1]$	*// 0*
$\beta \leftarrow \beta^2$, $\beta \leftarrow \beta \cdot \alpha$	*// 1*
$\beta \leftarrow \beta^2$	*// 10*
$\beta \leftarrow \beta^2$	*// 100*
$\beta \leftarrow \beta^2$, $\beta \leftarrow \beta \cdot \alpha$	*// 1001*
$\beta \leftarrow \beta^2$	*// 10010*
$\beta \leftarrow \beta^2$, $\beta \leftarrow \beta \cdot \alpha$	*// 100101* . $\square$

EXERCISE 2.21. Suppose $\alpha \in \mathbb{Z}_n^*$ has multiplicative order k. Let m be any integer. Show that α^m has multiplicative order $k/\gcd(m, k)$. Hint: use Theorem 2.10 and Exercise 2.10.

EXERCISE 2.22. Suppose $\alpha \in \mathbb{Z}_n^*$ has multiplicative order k and $\beta \in \mathbb{Z}_n^*$ has multiplicative order ℓ, where $\gcd(k, \ell) = 1$. Show that $\alpha\beta$ has multiplicative order $k\ell$. Hint: suppose $(\alpha\beta)^m = 1$, and deduce that both sides of the equation $\alpha^m = \beta^{-m}$ must have multiplicative order 1 (the previous exercise may be helpful); from this, deduce that both k and ℓ divide m and then apply the result of Exercise 1.12.

EXERCISE 2.23. Let $\alpha \in \mathbb{Z}_n^*$. Suppose that for a prime q and positive integer e, we have

$$\alpha^{q^e} = 1 \text{ and } \alpha^{q^{e-1}} \neq 1.$$

Show that α has multiplicative order q^e.

EXERCISE 2.24. Find all primitive roots mod 19. Show your work. To simplify the calculations, you may make make use of Fermat's little theorem, as well as the result of Exercise 2.21.

EXERCISE 2.25. Calculate the order of 9 mod 100. Show your work by building a table of powers 9^i mod 100. In addition, from this table, identify the multiplicative inverse of 9 mod 100.

EXERCISE 2.26. Let $n \in \mathbb{Z}$ with $n > 1$. Show that n is prime if and only if $\alpha^{n-1} = 1$ for every non-zero $\alpha \in \mathbb{Z}_n$.

EXERCISE 2.27. Let p be any prime other than 2 or 5. Show that p divides infinitely many of the numbers 9, 99, 999, etc.

Exercise 2.28. Let n be an integer greater than 1. Show that n does not divide $2^n - 1$. Hint: assume that n divides $2^n - 1$; now consider a prime p which divides n, so that p also divides $2^n - 1$, and derive a contradiction.

Exercise 2.29. This exercise develops an alternative proof of Fermat's little theorem.

(a) Using Exercise 1.16, show that for all primes p and integers a, we have $(a+1)^p \equiv a^p + 1 \pmod{p}$.

(b) Now derive Fermat's little theorem from part (a).

Exercise 2.30. Compute $3^{99} \bmod 100$ using the repeated squaring algorithm. Show your work.

Chapter 3

Rings and polynomials

This chapter introduces the notion of polynomials whose coefficients are in a general algebraic structure called a "ring". So to begin with, we introduce the mathematical notion of a ring. While there is a lot of terminology associated with rings, the basic ideas are fairly simple. Intuitively speaking, a ring is just a structure with addition and multiplication operations that behave exactly as one would expect—there are very few surprises.

3.1 Rings: Definitions, examples, and basic properties

Definition 3.1 (Ring). *A* **ring** *is a set R together with addition and multiplication operations on R, such that:*

(i) addition is commutative and association;

(ii) there exists an additive identity (or zero element), denoted 0_R, where $a+0_R = a$ for all $a \in R$;

(iii) every $a \in R$ has an additive inverse, denoted $-a$, where $a + (-a) = 0_R$;

(iv) multiplication is commutative and associative;

(v) there exists a multiplicative identity, denoted 1_R, where $a \cdot 1_R = a$ for all $a \in R$;

(vi) multiplication distributes over addition; that is, for all $a, b, c \in R$, we have $a(b+c) = ab+ac$.

Note that in this definition, the only property that connects addition and multiplication is the distributive law (property (vi)). If the ring R is clear from context, we may just write 0 and 1 instead of 0_R and 1_R.

Example 3.1. The set $\mathbb{Z}$ under the usual rules of multiplication and addition forms a ring. □

Example 3.2. For $n \geq 1$, the set $\mathbb{Z}_n$ under the rules of multiplication and addition defined in §2.4 forms a ring. □

Example 3.3. The set $\mathbb{Q}$ of rational numbers under the usual rules of multiplication and addition forms a ring. □

Example 3.4. The set $\mathbb{R}$ of real numbers under the usual rules of multiplication and addition forms a ring. □

Example 3.5. The set $\mathbb{C}$ of complex numbers under the usual rules of multiplication and addition forms a ring. Every $\alpha \in \mathbb{C}$ can be written (uniquely) as $\alpha = a + bi$, where $a, b \in \mathbb{R}$ and $i = \sqrt{-1}$. If $\alpha' = a' + b'i$ is another complex number, with $a', b' \in \mathbb{R}$, then

$$\alpha + \alpha' = (a + a') + (b + b')i \quad \text{and} \quad \alpha\alpha' = (aa' - bb') + (ab' + a'b)i. \ \square$$

Example 3.6. The **trivial ring** consists of a single element. It is not a very interesting ring. $\square$

We state some simple facts:

Theorem 3.2. *Let R be a ring. Then:*

(i) the additive and multiplicative identities are unique;

(ii) for all $a, b, c \in R$, if $a + b = a + c$, then $b = c$;

(iii) $-(a + b) = (-a) + (-b)$ for all $a, b \in R$;

(iv) $-(-a) = a$ for all $a \in R$;

(v) $0_R \cdot a = 0_R$ for all $a \in R$;

(vi) $(-a)b = -(ab) = a(-b)$ for all $a, b \in R$;

(vii) $(-a)(-b) = ab$ for all $a, b \in R$.

While there are many parts to this theorem, everything in it just states familiar properties that are satisfied by ordinary numbers. What is interesting about this theorem is that all of these properties follow directly from Definition 3.1. We shall not prove these properties here, but invite the reader to do so as a straightforward exercise.

Because addition and multiplication in a ring R are associative, for $a_1, \ldots, a_k \in \mathbb{Z}_n$, we may write the sum $a_1 + \cdots + a_k$ and the product $a_1 \cdots a_k$ without any parentheses, and there is no ambiguity; moreover, since both addition and multiplication are commutative, we may rearrange the terms in such sums and products without changing their values. We can write the sum as $\sum_{i=1}^{k} a_i$ and the product as $\prod_{i=1}^{k} a_i$. By convention, if $k = 0$, the sum is 0_R and the product is 1_R. If all of the a_i's are equal to a, we can write the sum as ka and the product as a^k. Note that the additive inverse of $\sum_{i=1}^{k} a_i$ is $\sum_{i=1}^{k}(-a_i)$, and the additive inverse of ka is $k(-a)$, which we can also write as $(-k)a$.

One other matter of notation: for $a, b \in R$, we can write $a - b$ instead of $a + (-b)$.

3.1.1 Multiplicative inverses and fields

While the definition of a ring requires that every element has an additive inverse, it does not require that any element has a multiplicative inverse.

Definition 3.3 (Multiplicative inverses in a ring). *Let R be a ring. For $a \in R$, we say that $b \in R$ is a* **multiplicative inverse** *of a if $ab = 1_R$. We define R^* to be the set of all elements of R that have a multiplicative inverse.*

Example 3.7. In the ring $\mathbb{Q}$, every non-zero element has a multiplicative inverse. The same holds for the rings $\mathbb{R}$ and $\mathbb{C}$. $\square$

Example 3.8. In the ring $\mathbb{Z}$, the only elements with multiplicative inverse are ± 1. That is, $\mathbb{Z}^* = \{\pm 1\}$. While the integer 2 has a multiplicative inverse $1/2 \in \mathbb{Q}$, it does not have a multiplicative inverse in $\mathbb{Z}$. $\square$

Example 3.9. As we saw in §2.4, $\mathbb{Z}_n^* = \{[a]_n : \gcd(a, n) = 1\}$. $\square$

Example 3.10. If R is a ring, then $1_R \in R^*$ (this follows from the fact that $1_R \cdot 1_R = 1_R$, by definition). Moreover, if R is non-trivial ring, then we have $0_R \neq 1_R$, and moreover, $0_R \notin R^*$ (these observations follow from part (v) of Theorem 3.2). $\square$

Let R be a ring. If $a \in R^*$, then it is not hard to prove that the multiplicative inverse must be unique, and it is denoted a^{-1}. It is not hard to see that if $a, b \in R^*$, then so are a^{-1} and ab; indeed, one can easily verify that we must have

$$(a^{-1})^{-1} = a \quad \text{and} \quad (ab)^{-1} = a^{-1}b^{-1}.$$

If a has a multiplicative inverse a^{-1}, and k is a non-negative integer, then the multiplicative inverse of a^k is $(a^{-1})^k$, which we may write as a^{-k}. Naturally, for $a \in R$ and $b \in R^*$, we may write a/b instead of ab^{-1}.

As the above examples demonstrate, it need not be the case that every non-zero element in a ring has a multiplicative inverse. However, rings with this property are especially nice and deserve to be singled out:

Definition 3.4 (Field). *A non-trivial ring where every non-zero element has a multiplicative inverse is called a* **field**.

Example 3.11. The rings $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields. $\square$

Example 3.12. The ring $\mathbb{Z}$ is not a field. $\square$

Example 3.13. The ring $\mathbb{Z}_n$ is a field of and only if n is prime. $\square$

Suppose a, b, c are elements of R that satisfy the equation

$$ab = ac.$$

If $a \in R^*$, we may multiply both sides of this equation by a^{-1} to infer that

$$b = c.$$

This is the **cancellation law** for R. In particular, if R is a field, then this cancellation law holds provided $a \neq 0$.

We also have the following result on the **existence and uniqueness to solutions of linear equations in** R. Let $a \in R^*$ and $b \in R$. Then the equation

$$az = b$$

has a unique solution z, namely, $z := a^{-1}b$. In particular, if R is a field, then this equation has a unique solution z provided $a \neq 0$.

EXERCISE 3.1. Show that the product of two non-zero elements in a field is also non-zero.

EXERCISE 3.2. Give an example of a ring R and two non-zero elements $a, b \in R$ such $ab = 0_R$.

3.2 Polynomial rings

If R is a ring, then we can form the **ring of polynomials** $R[\mathtt{X}]$, consisting of all polynomials $g = a_0 + a_1\mathtt{X} + \cdots + a_k\mathtt{X}^k$ in the **indeterminate**, or "formal" variable, $\mathtt{X}$, with coefficients a_i in R, and with addition and multiplication defined in the usual way.

Example 3.14. Let us define a couple of polynomials over the ring $\mathbb{Z}_5$:

$$g := [3]\mathtt{X} + \mathtt{X}^2, \quad h := [1] + [2]\mathtt{X} + [4]\mathtt{X}^3$$

We have:

$$g + h = [1] + \mathtt{X}^2 + [4]\mathtt{X}^3, \quad g \cdot h = [3]\mathtt{X} + [2]\mathtt{X}^2 + [2]\mathtt{X}^3 + [2]\mathtt{X}^4 + [4]\mathtt{X}^5. \ \square$$

Elements of R are also considered to be polynomials. Such polynomials are called **constant polynomials**. In particular, 0_R is the additive identity in $R[\mathtt{X}]$ and 1_R is the multiplicative identity in $R[\mathtt{X}]$. Note that if R is the trivial ring, then so is $R[\mathtt{X}]$.

So as to keep the distinction between ring elements and indeterminates clear, we shall use the symbol "$\mathtt{X}$" only to denote the latter. Also, for a polynomial $g \in R[\mathtt{X}]$, we shall in general write this simply as "g," and not as "$g(\mathtt{X})$." Of course, the choice of the symbol "$\mathtt{X}$" is arbitrary.

3.2.1 Formalities

For completeness, we present a more formal definition of the ring $R[\mathtt{X}]$. The reader should bear in mind that this formalism is rather tedious, and may be more distracting than it is enlightening. Formally, a polynomial $g \in R[\mathtt{X}]$ is an infinite sequence $(a_0, a_1, a_2, \ldots)$, where each $a_i \in R$, but only finitely many of the a_i's are non-zero (intuitively, a_i represents the coefficient of $\mathtt{X}^i$).

For

$$g = (a_0, a_1, a_2, \ldots) \in R[\mathtt{X}] \ \text{ and } \ h = (b_0, b_1, b_2, \ldots) \in R[\mathtt{X}],$$

we define

$$g + h := (s_0, s_1, s_2, \ldots) \ \text{ and } \ gh := (p_0, p_1, p_2, \ldots),$$

where for $i = 0, 1, 2, \ldots$,

$$s_i := a_i + b_i \tag{3.1}$$

and

$$p_i := \sum_{i=j+k} a_j b_k, \tag{3.2}$$

the sum being over all pairs (j, k) of non-negative integers such that $i = j+k$ (which is a finite sum). We leave it to the reader to verify that $g + h$ and gh are polynomials (i.e., only finitely many of the s_i's and p_i's are non-zero). The reader may also verify that all the requirements of Definition 3.1 are satisfied: the additive identity is the all-zero sequence $(0_R, 0_R, 0_R, \ldots)$; the multiplicative identity is the sequence $(1_R, 0_R, 0_R, \ldots)$, that is, the sequence consists of 1_R followed by all zeros.

For $c \in R$, we can identify c with the corresponding "constant" polynomial $(c, 0_R, 0_R, \ldots)$. If we define the polynomial

$$\mathtt{X} := (0_R, 1_R, 0_R, 0_R, \ldots),$$

then for any polynomial $g = (a_0, a_1, a_2, \ldots)$, if $a_i = 0_R$ for all i exceeding some value k, then we have $g = \sum_{i=0}^{k} a_i \mathtt{X}^i$, and so we can return to the standard practice of writing polynomials as we did in Example 3.14, without any loss of precision.

3.2.2 Basic properties of polynomials

Let R be a ring. For non-zero $g \in R[\mathtt{X}]$, if $g = \sum_{i=0}^{k} a_i \mathtt{X}^i$ with $a_k \neq 0$, then we call k the **degree** of g, denoted $\deg(g)$, we call a_k the **leading coefficient** of g, denoted $\mathrm{lc}(g)$, and we call a_0 the **constant term** of g. If $\mathrm{lc}(g) = 1$, then g is called **monic**.

Suppose $g = \sum_{i=0}^{k} a_i \mathtt{X}^i$ and $h = \sum_{i=0}^{\ell} b_i \mathtt{X}^i$ are polynomials such that $a_k \neq 0$ and $b_\ell \neq 0$, so that $\deg(g) = k$ and $\mathrm{lc}(g) = a_k$, and $\deg(h) = \ell$ and $\mathrm{lc}(h) = b_\ell$. When we multiply these two polynomials, we get

$$gh = a_0 b_0 + (a_0 b_1 + a_1 b_0)\mathtt{X} + \cdots + a_k b_\ell \mathtt{X}^{k+\ell}.$$

In particular, if $gh \neq 0$, then $\deg(gh) \leq \deg(g) + \deg(h)$. Note that if R is a field, we must have $a_k b_\ell \neq 0$, and so $\deg(gh) = \deg(g) + \deg(h)$ in this case.

For the zero polynomial, we establish the following conventions: its leading coefficient and constant term are defined to be 0_R, and its degree is defined to be $-\infty$. With these conventions, we may succinctly state that

> *for all $g, h \in R[\mathtt{X}]$, we have $\deg(gh) \leq \deg(g) + \deg(h)$, and equality always holds if R is a field.*

3.2.3 Polynomial evaluation

A polynomial $g = \sum_{i=0}^{k} a_i \mathtt{X}^i \in R[\mathtt{X}]$ naturally defines a polynomial function on R that sends $u \in R$ to $\sum_{i=0}^{k} a_i u^i \in R$, and we denote the value of this function as $g(u)$ (note that "$\mathtt{X}$" denotes an indeterminate, while "u" denotes an element of R). As usual, we define $u \in R$ to be a **root** of g if $g(u) = 0$.

It is important to regard polynomials over R as formal expressions, and *not* to identify them with their corresponding functions. In particular, two polynomials are equal if and only if their coefficients are equal, while two functions are equal if and only if their values agree at all inputs in R. This distinction is important, since there are rings R over which two different polynomials define the same function. One can of course define the ring of polynomial functions on R, but in general, that ring has a different structure from the ring of polynomials over R.

Example 3.15. In the ring $\mathbb{Z}_p$, for prime p, by Fermat's little theorem, we have $u^p = u$ for all $u \in \mathbb{Z}_p$. However, the polynomials $\mathtt{X}^p$ and $\mathtt{X}$ are not the same polynomials (in particular, the former has degree p, while the latter has degree 1). □

An obvious, yet important, fact is the following:

Theorem 3.5. *Let R be a ring. For all $g, h \in R[\mathtt{X}]$ and $u \in R$, if $s := g + h \in R[\mathtt{X}]$ and $p := gh \in R[\mathtt{X}]$, then we have*

$$s(u) = g(u) + h(u) \quad \text{and} \quad p(u) = g(u)h(u).$$

Proof. The proof is really just symbol pushing. Indeed, suppose $g = \sum_i a_i \mathtt{X}^i$ and $h = \sum_i b_i \mathtt{X}^i$. Then $s = \sum_i (a_i + b_i)\mathtt{X}^i$, and so

$$s(u) = \sum_i (a_i + b_i)u^i = \sum_i a_i u^i + \sum_i b_i u^i = g(u) + h(u).$$

Also, we have

$$p = \left(\sum_i a_i \mathtt{X}^i\right)\left(\sum_j b_j \mathtt{X}^j\right) = \sum_{i,j} a_i b_j \mathtt{X}^{i+j},$$

and employing the result for evaluating sums of polynomials, we have

$$p(u) = \sum_{i,j} a_i b_j u^{i+j} = \left(\sum_i a_i u^i\right)\left(\sum_j b_j u^j\right) = g(u)h(u). \quad \square$$

3.2.4 Polynomial interpolation

The reader is surely familiar with the fact that two points determine a line, in the context of real numbers. The reader is perhaps also familiar with the fact that over the real (or complex) numbers, every polynomial of degree k has at most k distinct roots, and the fact that every set of k points can be interpolated by a unique polynomial of degree less than k. As we will now see, these results extend to arbitrary fields.[1]

Let F be a field, and consider the ring of polynomials $F[\mathtt{X}]$. Analogous to integers, we can define a notion of divisibility for $F[\mathtt{X}]$: for polynomials $g, h \in F[\mathtt{X}]$ we say that g **divides** h, which we may write as $g \mid h$, if $gz = h$ for some $z \in F[\mathtt{X}]$.

Just like the integers, there is a corresponding division with remainder property for polynomials:

Theorem 3.6 (Division with remainder property). *Let F be a field. For all $g, h \in F[\mathtt{X}]$ with $h \neq 0$, there exist unique $q, r \in F[\mathtt{X}]$ such that $g = hq + r$ and $\deg(r) < \deg(h)$.*

Proof. Consider the set $S := \{g - ht : t \in F[\mathtt{X}]\}$. Let $r = g - hq$ be an element of S of minimum degree. We must have $\deg(r) < \deg(h)$, since otherwise, we could subtract an appropriate multiple of h from r so as to eliminate the leading coefficient of r, obtaining

$$r' := r - h \cdot (\mathrm{lc}(r)\,\mathrm{lc}(h)^{-1}\mathtt{X}^{\deg(r)-\deg(h)}) \in S,$$

where $\deg(r') < \deg(r)$, contradicting the minimality of $\deg(r)$.

That proves the existence of r and q. For uniqueness, suppose that $g = hq + r$ and $g = hq' + r'$, where $\deg(r) < \deg(h)$ and $\deg(r') < \deg(h)$. This implies $r' - r = h \cdot (q - q')$. However, if $q \neq q'$, then

$$\deg(h) > \deg(r' - r) = \deg(h \cdot (q - q')) = \deg(h) + \deg(q - q') \geq \deg(h),$$

which is impossible. Therefore, we must have $q = q'$, and hence $r = r'$. $\square$

If $g = hq + r$ as in the above theorem, we define $g \bmod h := r$. Clearly, $h \mid g$ if and only if $g \bmod h = 0$. Moreover, note that if $\deg(g) < \deg(h)$, then $q = 0$ and $r = g$; otherwise, if $\deg(g) \geq \deg(h)$, then $q \neq 0$ and $\deg(g) = \deg(h) + \deg(q)$.

Theorem 3.7. *Let F be a field, $g \in F[\mathtt{X}]$, and $u \in F$ be a root of g. Then $(\mathtt{X} - u)$ divides g.*

Proof. Using the division with remainder property for polynomials, there exist $q, r \in F[\mathtt{X}]$ such that $g = (\mathtt{X} - u)q + r$, with $q, r \in F[\mathtt{X}]$ and $\deg(r) < 1$, which means that $r \in F$. Evaluating at u, we see that $g(u) = (u - u)q(u) + r = r$. Since u is a root of g, we must have $r = 0$, and therefore, $g = (\mathtt{X} - u)q$, and so $(\mathtt{X} - u)$ divides g. $\square$

Theorem 3.8. *Let F be a field, and let $u_1, \ldots, u_k$ be distinct elements of F. Then for every polynomial $g \in F[\mathtt{X}]$, the elements $u_1, \ldots, u_k$ are roots of g if and only if the polynomial $\prod_{i=1}^{k}(\mathtt{X} - u_i)$ divides g.*

[1] In fact, much of what we discuss here extends, with some modification, to more general coefficient rings.

Proof. One direction is trivial: if $\prod_{i=1}^{k}(\mathtt{X} - u_i)$ divides g, then it is clear that each u_i is a root of g. We prove the converse by induction on k. The base case $k = 1$ is just Theorem 3.7. So assume $k > 1$, and that the statement holds for $k - 1$. Let $g \in F[\mathtt{X}]$ and let $u_1, \ldots, u_k$ be distinct roots of g. Since u_k is a root of g, then by Theorem 3.7, there exists $q \in F[\mathtt{X}]$ such that $g = (\mathtt{X} - u_k)q$. Moreover, for each $i = 1, \ldots, k-1$, we have

$$0 = g(u_i) = (u_i - u_k)q(u_i),$$

and since $u_i - u_k \neq 0$ and F is a field, we must have $q(u_i) = 0$. Thus, q has roots $u_1, \ldots, u_{k-1}$, and by induction $\prod_{i=1}^{k-1}(\mathtt{X} - u_i)$ divides q, from which it then follows that $\prod_{i=1}^{k}(\mathtt{X} - u_i)$ divides g. □

As an immediate consequence of this theorem, we obtain:

Theorem 3.9. *Let F be a field, and suppose that $g \in F[\mathtt{X}]$, with $\deg(g) = k \geq 0$. Then g has at most k distinct roots.*

Proof. If g had $k + 1$ distinct roots $u_1, \ldots, u_{k+1}$, then by the previous theorem, the polynomial $\prod_{i=1}^{k+1}(\mathtt{X} - u_i)$, which has degree $k + 1$, would divide g, which has degree k—an impossibility. □

We now present the main result of this section. It says given k points $(u_1, v_1), \ldots, (u_k, v_k)$, were the u_i's are distinct elements of a field F and the v_i's are arbitrary elements of F, there is a unique polynomial $g \in F[\mathtt{X}]$ of degree less than k that "interpolates" or "passes through" these points, that is, $g(u_i) = v_i$ for $i = 1, \ldots, k$.

Theorem 3.10 (Lagrange interpolation). *Let F be a field, let $u_1, \ldots, u_k$ be distinct elements of F, and let $v_1, \ldots, v_k$ be arbitrary elements of F. Then there exists a unique polynomial $g \in F[\mathtt{X}]$ with $\deg(g) < k$ such that $g(u_i) = v_i$ for $i = 1, \ldots, k$, namely*

$$g := \sum_{i=1}^{k} v_i \frac{\prod_{j \neq i}(\mathtt{X} - u_j)}{\prod_{j \neq i}(u_i - u_j)}.$$

Proof. For the existence part of the theorem, one just has to verify that $g(u_i) = v_i$ for the given g, which clearly has degree less than k. This is easy to see: for $i = 1, \ldots, k$, evaluating the ith term in the sum defining g at u_i yields v_i, while evaluating any other term at u_i yields 0. The uniqueness part of the theorem follows almost immediately from Theorem 3.9: if g and h are polynomials of degree less than k such that $g(u_i) = v_i = h(u_i)$ for $i = 1, \ldots, k$, then $g - h$ is a polynomial of degree less than k with k distinct roots, which, by the previous theorem, is impossible. □

Example 3.16. Consider the field $F = \mathbb{Z}_5$. We use the Lagrange interpolation formula to compute the coefficients of the polynomial $g \in F[\mathtt{X}]$ of degree less than 3 such that

$$g([1]) = [1], \quad g([3]) = [4], \quad g([4]) = [2].$$

Applying Theorem 3.10 with

$$u_1 = [1],\ u_2 = [3],\ u_3 = [4], \quad \text{and} \quad v_1 = [1],\ v_2 = [4], \text{ and } v_3 = [2],$$

we have $g = L_1 + L_2 + L_3$, where

$$\begin{aligned} L_1 &= [1]\frac{(\mathtt{X}-[3])(\mathtt{X}-[4])}{([1]-[3])([1]-[4])} = [1]\frac{\mathtt{X}^2+[3]\mathtt{X}+[2]}{[1]} \\ &= \mathtt{X}^2+[3]\mathtt{X}+[2], \end{aligned}$$

$$\begin{aligned} L_2 &= [4]\frac{(\mathtt{X}-[1])(\mathtt{X}-[4])}{([3]-[1])([3]-[4])} = [4]\frac{\mathtt{X}^2+[4]}{[3]} \\ &= [4][2](\mathtt{X}^2+[4]) \quad \text{(using the fact that } [3]^{-1}=[2]) \\ &= [3]\mathtt{X}^2+[2], \end{aligned}$$

$$\begin{aligned} L_3 &= [2]\frac{(\mathtt{X}-[1])(\mathtt{X}-[3])}{([4]-[1])([4]-[3])} = [2]\frac{\mathtt{X}^2+\mathtt{X}+[3]}{[3]} \\ &= [2][2](\mathtt{X}^2+\mathtt{X}+[3]) \quad \text{(again, using the fact that } [3]^{-1}=[2]) \\ &= [4]\mathtt{X}^2+[4]\mathtt{X}+[2]. \end{aligned}$$

Putting it all together, we have $g = [3]\mathtt{X}^2+[2]\mathtt{X}+[1]$. $\square$

Given distinct $u_1, \ldots, u_k \in F$ and arbitrary $v_1, \ldots, v_k \in F$, as in Theorem 3.10, the coefficients $a_0, \ldots, a_{k-1}$ of the interpolating polynomial $g = \sum_i a_i \mathtt{X}^i \in F[\mathtt{X}]$ can be expressed as the unique solution to the following matrix equation:

$$\underbrace{\begin{pmatrix} 1 & u_1 & \cdots & u_1^{k-1} \\ 1 & u_2 & \cdots & u_2^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & u_k & \cdots & u_k^{k-1} \end{pmatrix}}_{V:=} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}$$

The matrix V is called a **Vandermonde matrix**. Note that one can derive the uniqueness part of Theorem 3.10 from the existence part by general facts from linear algebra. Indeed, the existence part implies that the column space of V has full rank k, which means that the null space of V has dimension 0.

Application: Reed-Solomon Codes

As an application of polynomial interpolation, we present some ideas from the theory of error correcting codes.

Error correcting codes are a powerful technology that are used to to correct errors that occur in data transmission (or storage). They are widely used and there is a vast literature on the subject. Here, we present some of the main ideas underlying one such code, called a Reed-Solomon code.

Suppose a message is represented as a tuple $(a_0, \ldots, a_{k-1})$, where each $a_i \in \mathbb{Z}_p$ and p is prime. For example, if each a_i represents an 8-bit byte, we can think of it as an element of $\mathbb{Z}_p$ where $p = 257$. Let $g := \sum_{i=0}^{k-1} a_i \mathtt{X}^i \in \mathbb{Z}_p[\mathtt{X}]$. Let $u_1, \ldots, u_n$ be arbitrary, fixed elements of $\mathbb{Z}_p$, where $n > k$. We form the "code word" $(v_1, \ldots, v_n)$, where $v_i := g(u_i) \in \mathbb{Z}_p$ for $i = 1, \ldots, n$.

Now suppose that code word $(v_1, \ldots, v_n)$ is transmitted, and that some errors may occur during transmission. Let $(v_1^*, \ldots, v_n^*)$ be the received code word. Suppose $v_i^* \neq v_i$ at t different positions. This means that t of the v_i's in the original code word were modified in transmission, and the remaining $n - t$ of the v_i's remain unmodified.

Let us call the points (u_i, v_i) for $i = 1, \ldots, n$ the "transmitted points", the points (u_i, v_i^*) for $i = 1, \ldots, n$ the "received points", and the points (u_i, v_i^*) where $v_i^* \neq v_i$ the "mangled points".

The polynomial g interpolates $n - t$ of the received points (namely, the non-mangled points). We will show that no other polynomial of degree less than k interpolates this many of the received points, as long as $t \leq (n-k)/2$. That is, provided the number of errors is not too large, then among all polynomial of degree less than k, the polynomial g is the unique "best fitting" polynomial, in the sense that it interpolates more of the received points than any other.

To see why, suppose $\tilde{g} \neq g$ is a polynomial of degree less than k. We claim that $\tilde{g}$ interpolates *fewer* than $t + k$ received points. To see why, suppose $\tilde{g}$ does interpolate $t + k$ received points. Of these $t + k$ receved points, at least k of them must be non-mangled points, so there is a set of k transmitted points that are interpolated by both $\tilde{g}$ and g. By the uniqueness of interpolation, we must have $\tilde{g} = g$, a contradiction.

That proves that $\tilde{g}$ interpolates fewer than $t + k$ received points. However, we want to show that $\tilde{g}$ interpolates fewer than $n - t$ reveived points. To show this, we just need to show that $t + k \leq n - 1$. But observer that

$$t + k \leq n - t \iff 2t \leq n - k \iff t \leq \frac{n-k}{2},$$

and we are assuming that $t \leq (n - k)/2$.

So, provided $t \leq (n - k)/2$, the original message $(a_0, \ldots, a_{k-1})$ is not lost in principle, since the corresponding polynomial $g = \sum_{i=0}^{k-1} a_i \mathtt{X}^i$ is the unique "best fitting" polynomial of degree less than k. But one problem remains: how to efficiently compute the polynomial g from the mangled points. We cannot use the Lagrange interpolation formula directly, because we do not know which of these points are "good" and which are "bad". However, there is an efficient algorithm (called the Berlekamp-Welch algorithm) that does solve this problem. Unfortunately, the details of that algorithm are out of the scope of these notes.

Application: Existence of primitive roots

As another application of polynomial interpolation, we give a simple proof that for every prime p, there exists a primitive root mod p (see §2.5).

Theorem 3.11 (Existence of a primitive root). *For every prime p, there exists a primitive root mod p.*

Proof. Let p be a prime. We know by Theorem 2.10 that every element of $\mathbb{Z}_p^*$ has multiplicative order dividing $p - 1$. We want to show that there exists an element $\alpha \in \mathbb{Z}_p^*$ of order *equal* to $p - 1$.

Suppose the factorization of $p - 1$ into primes is

$$p - 1 = q_1^{e_1} \cdots q_r^{e_r}.$$

Claim. For each $i = 1, \ldots, r$, there exists an element $\beta_i \in \mathbb{Z}_p^*$ of order $q_i^{e_i}$.

The theorem follows from the claim by setting

$$\alpha := \beta_1 \cdots \beta_r.$$

The fact that α has multiplicative order $p - 1$ follows from Exercise 2.22.

It remains to prove the claim. Fix $i = 1, \ldots, r$, and set $q := q_i$ and $e := e_i$. We want to exhibit an element $\beta \in \mathbb{Z}_p^*$ of order q^e. Consider the polynomial

$$\mathtt{X}^{(p-1)/q} - 1 \in \mathbb{Z}_p[\mathtt{X}].$$

By Theorem 3.9, this has at most $(p-1)/q$ distinct roots, so there must be some element of $\mathbb{Z}_p^*$ that is *not* a root of this polynomial. Fix such an element $\gamma \in \mathbb{Z}_p^*$. So we have

$$\gamma^{(p-1)/q} \neq 1.$$

We contend that

$$\beta := \gamma^{(p-1)/q^e}$$

does the job. On the one hand, we have

$$\beta^{q^e} = (\gamma^{(p-1)/q^e})^{q^e} = \gamma^{p-1} = 1.$$

On the other hand, we have

$$\beta^{q^{e-1}} = (\gamma^{(p-1)/q^e})^{q^{e-1}} = \gamma^{(p-1)/q} \neq 1.$$

The fact that β has multiplicative order q^e follows from Exercise 2.23. $\square$

Exercise 3.3. Consider the field $F = \mathbb{Z}_5$. Let $f = \mathtt{X}^3 + \mathtt{X} + [1] \in F[\mathtt{X}]$, $g = [2]\mathtt{X}^2 + [3]\mathtt{X} + [4] \in F[\mathtt{X}]$, and $h = [3]\mathtt{X}^2 + [2]\mathtt{X} + [1] \in F[\mathtt{X}]$. Compute $(gh) \bmod f$. Show your work. That is, you should compute the product polynomial $P = gh \in F[\mathtt{X}]$, and then use polynomial division with remainder to compute $P \bmod f$. Along the way, you the coefficients of any polynomials you compute should be reduced mod 5.

Exercise 3.4. Consider the field $F = \mathbb{Z}_5$. Use the Lagrange interpolation formula to compute the coefficients of the polynomial $g \in F[\mathtt{X}]$ of degree less than 3 such that

$$g([1]) = [3], \quad g([2]) = [4], \quad g([3]) = [1].$$

Show your work.

Exercise 3.5. Suppose F is a field and let $g, h \in F[\mathtt{X}]$. Show that if $g \mid h$ and $h \mid g$, then $g = ch$ for some $c \in F^*$.

Exercise 3.6. Let F be an infinite field, and let $g, h \in F[\mathtt{X}]$. Show that if $g(u) = h(u)$ for all $u \in F$, then $g = h$. Thus, for an infinite field F, there is a one-to-one correspondence between polynomials over F and polynomial functions on F.

Exercise 3.7. Let F be a field.

(a) Show that for all $b \in F$, we have $b^2 = 1$ if and only if $b = \pm 1$.

(b) Show that for all $a, b \in F$, we have $a^2 = b^2$ if and only if $a = \pm b$.

(c) Show that the familiar **quadratic formula** holds for F, assuming that $2_F := 1_F + 1_F \neq 0_F$. That is, for all $a, b, c \in F$ with $a \neq 0_F$, the polynomial $g := a\mathtt{X}^2 + b\mathtt{X} + c \in F[\mathtt{X}]$ has a root in F if and only if there exists $e \in F$ such that $e^2 = d$, where d is the **discriminant** of g, defined as $d := b^2 - 4ac$, and in this case the roots of g are $(-b \pm e)/2a$.