MODULAR ARITHMETIC

KEITH CONRAD

1. Introduction

We will define the notion of congruent integers (with respect to a modulus) and develop some basic ideas of modular arithmetic. Applications of modular arithmetic are given to divisibility tests and to block ciphers in cryptography.

Modular arithmetic lets us carry out algebraic calculations on integers with a systematic disregard for terms divisible by a certain number (called the modulus). This kind of "reduced algebra" is essential background for the mathematics of computer science, coding theory, primality testing, and much more.

2. Integer congruences

The following definition was introduced by Gauss in his *Disquisitiones Arithmeticae* (Arithmetic Investigations) in 1801.

Definition 2.1 (Gauss)**.** Let m be an integer. For $a, b \in \mathbf{Z}$, we write

$$a \equiv b \bmod m$$

and say "a is congruent to b modulo m" if $m \mid (a - b)$.

Example 2.2. Check $18 \equiv 4 \bmod 7$, $-20 \equiv 13 \bmod 11$, and $19 \equiv 3 \bmod 2$.

Remark 2.3. The parameter m is called the modulus, *not* the modulo. The symbol $\equiv$ in LaTeX is written as `\equiv`, but it is always pronounced "congruent," **never** "equivalent". (The LaTeX command `\cong` is for the congruence symbol $\cong$ in elementary geometry.)

We always have $m \equiv 0 \bmod m$, and more generally $mk \equiv 0 \bmod m$ for any $k \in \mathbf{Z}$. In fact,

$$a \equiv 0 \bmod m \iff m \mid a,$$

so the congruence relation includes the divisibility relation as a special case: the multiples of m are exactly the numbers that "look like 0" modulo m. Because multiples of m are congruent to 0 modulo m, we will see that working with integers modulo m is tantamount to systematically ignoring additions and subtractions by multiples of m in algebraic calculations.

Since $a \equiv b \bmod m$ if and only if $b = a + mk$ for some $k \in \mathbf{Z}$, adjusting an integer modulo m is the same as adding (or subtracting) multiples of m to it. Thus, if we want to find a positive integer congruent to $-18 \bmod 5$, we can add a multiple of 5 to -18 until we go positive. Adding 20 does the trick: $-18 + 20 = 2$, so $-18 \equiv 2 \bmod 5$ (check!).

There is a useful analogy between integers modulo m and angle measurements (in radians, say). In both cases, the objects involved admit *different representations*, *e.g.*, the angles 0, 2π, and -4π are the same, just as

$$2 \equiv 12 \equiv -13 \bmod 5.$$

Every angle can be put in "standard" form as a real number in the interval $[0, 2\pi)$. There is a similar convention for the "standard" representation of an integer modulo m using *remainders*, as follows.

Theorem 2.4. *Let $m \in \mathbf{Z}$ be a nonzero integer. For each $a \in \mathbf{Z}$, there is a unique r with $a \equiv r \bmod m$ and $0 \leq r < |m|$.*

Proof. Using division with remainder in $\mathbf{Z}$, there are q and r in $\mathbf{Z}$ such that

$$a = mq + r, \qquad 0 \leq r < |m|.$$

Then $m \mid (a - r)$, so $a \equiv r \bmod m$.

To show r is the unique number in the range $\{0, 1, \ldots, |m| - 1\}$ that is congruent to a mod m, suppose two numbers in this range work:

$$a \equiv r \bmod m, \quad a \equiv r' \bmod m,$$

where $0 \leq r, r' < |m|$. Then we have

$$a = r + mk, \quad a = r' + m\ell$$

for some k and ℓ in $\mathbf{Z}$, so the remainders r and r' have difference

$$r - r' = m(\ell - k).$$

This is a multiple of m, and the bounds on r and r' tell us $|r - r'| < |m|$. A multiple of m has absolute value less than $|m|$ only if it is 0, so $r - r' = 0$, which means $r' = r$. □

Example 2.5. Taking $m = 2$, every integer is congruent modulo 2 to exactly one of 0 and 1. Saying $n \equiv 0 \bmod 2$ means $n = 2k$ for some integer k, so n is even, and saying $n \equiv 1 \bmod 2$ means $n = 2k + 1$ for some integer k, so n is odd. We have $a \equiv b \bmod 2$ precisely when a and b have the *same parity*: both are even or both are odd.

Example 2.6. Every integer is congruent mod 4 to exactly one of 0, 1, 2, or 3. Congruence mod 4 is a refinement of congruence mod 2: even numbers are congruent to 0 or 2 mod 4 and odd numbers are congruent to 1 or 3 mod 4. For instance, $10 \equiv 2 \bmod 4$ and $19 \equiv 3 \bmod 4$.

Congruence mod 4 is related to Master Locks. Every combination on a Master Lock is a triple of numbers (a, b, c) where a, b, and c vary from 0 to 39. Each number has 40 choices, with $b \neq a$ and $c \neq b$ (perhaps $c = a$). This means the number of combinations could be up to $40 \cdot 39^2 = 60840$, but in fact the true number of combinations is a lot smaller: every combination has $c \equiv a \bmod 4$ and $b \equiv a + 2 \bmod 4$, so once some number in a combination is known the other two numbers are each limited to 10 choices (among the 40 numbers in $\{0, 1, \ldots, 39\}$ exactly 10 will be congruent to a particular value mod 4). Thus the real number of Master Lock combinations is $40 \cdot 10^2 = 4000$.

Example 2.7. Every integer is congruent modulo 7 to exactly one of $0, 1, 2, \ldots, 6$. The choice is the remainder when the integer is divided by 7. For instance, $20 \equiv 6 \bmod 7$ and $-32 \equiv 3 \bmod 7$.

Definition 2.8. We call $\{0, 1, 2, \ldots, |m| - 1\}$ the *standard* representatives for integers modulo m.

In practice $m > 0$, so the standard representatives modulo m are $\{0, 1, 2, \ldots, m - 1\}$. In fact, congruence modulo m and modulo $-m$ are the same relation (just look back at the definition), so usually we never talk about negative moduli. Nevertheless, Theorem 2.4 is stated for any modulus $m \neq 0$ for completeness.

By Theorem 2.4, there are $|m|$ incongruent integers modulo m. We can represent each integer modulo m by one of the standard representatives, just like we can write any fraction in a reduced form. There are many other representatives which could be used, however, and this will be important in the next section.

3. Modular Arithmetic

When we add and multiply fractions, we can change the representation (that is, use a different numerator and denominator) and the results are the same. We will meet a similar idea with addition and multiplication modulo m.

Theorem 3.1. *If $a \equiv b \bmod m$ and $b \equiv c \bmod m$ then $a \equiv c \bmod m$.*

Proof. By hypothesis, $a - b = mk$ and $b - c = m\ell$ for some integers k and ℓ. Adding the equations, $a - c = m(k + \ell)$ and $k + \ell \in \mathbf{Z}$, so $a \equiv c \bmod m$. □

This result is called transitivity of congruences. We will usually use it without comment in our calculations below.

The following theorem is the *key* algebraic feature of congruences in $\mathbf{Z}$: they behave well under addition and multiplication.

Theorem 3.2. *If $a \equiv b \bmod m$ and $c \equiv d \bmod m$, then*

$$a + c \equiv b + d \bmod m \text{ and } ac \equiv bd \bmod m.$$

Proof. We want to show $(a + c) - (b + d)$ and $ac - bd$ are multiples of m. Write $a = b + mk$ and $c = d + m\ell$ for k and ℓ in $\mathbf{Z}$. Then

$$(a + c) - (b + d) = a - b + c - d = m(k + \ell),$$

so $a + c \equiv b + d \bmod m$. For multiplication,

$$\begin{aligned} ac - bd &= (b + mk)(d + m\ell) - bd \\ &= m(kd + b\ell + mk\ell), \end{aligned}$$

so $ac \equiv bd \bmod m$. □

Example 3.3. Check $11 \equiv 5 \bmod 6$, $-2 \equiv 4 \bmod 6$, and $11 \cdot (-2) \equiv 5 \cdot 4 \bmod 6$.

Example 3.4. What is the standard representative for $17^2 \bmod 19$? You could compute $17^2 = 289$ and then divide 289 by 19 to find a remainder of 4, so $17^2 \equiv 4 \bmod 19$. Another way is to notice $17 \equiv -2 \bmod 19$, so $17^2 \equiv (-2)^2 \equiv 4 \bmod 19$. That was quicker, and it illustrates the meaning of multiplication being independent of the choice of representative. Sometimes one representative can be more convenient than another.

Example 3.5. If we want to compute $10^4 \bmod 19$, compute successive powers of 10 but reduce modulo 19 *each time* the answer exceeds 19: using the formula $10^k = 10 \cdot 10^{k-1}$ and writing $\equiv$ for congruence modulo 19,

$$10^1 = 10,\ 10^2 = 100 \equiv 5,\ 10^3 \equiv 10 \cdot 5 = 50 \equiv 12,\ 10^4 \equiv 10 \cdot 12 = 120 \equiv 6.$$

Thus $10^4 \equiv 6 \bmod 19$. Theorem 3.2 says this kind of procedure leads to the right answer, since multiplication modulo 19 is independent of the choice of representatives, so we can always replace a larger integer with a smaller representative of it modulo 19 without affecting the results of (further) algebraic operations modulo 19.

Corollary 3.6. *If $a \equiv b \bmod m$ and $k \in \mathbf{Z}^+$ then $a^k \equiv b^k \bmod m$.*

Proof. This follows from the second congruence in Theorem 3.2 using induction on k. Details are left to the reader. □

Remark 3.7. Do **not** try to extend Corollary 3.6 to fractional exponents until you have much more experience in modular arithmetic. Extracting roots is *not* repeated multiplication, and extracting roots in modular arithmetic could be undefined or have unexpected behavior compared with your experience extracting roots in $\mathbf{R}$.

For example, in $\mathbf{R}$ all positive numbers are perfect squares, but $x^2 \equiv 2 \bmod 5$ has no solution (that is, no integer x satisfies that congruence). Positivity is really a meaningless concept mod m: every integer is congruent to a positive integer and a negative integer mod m (just add and subtract a suitably large multiple of m to the integer). There is no good notion of ordering mod m.

In $\mathbf{R}$ you know $x^2 = y^2 \Rightarrow x = \pm y$, but it is *false* that $a^2 \equiv b^2 \bmod m \Rightarrow a \equiv \pm b \bmod m$ in general. Consider $4^2 \equiv 1^2 \bmod 15$ with $4 \not\equiv \pm 1 \bmod 15$.

In $\mathbf{R}$ you are used to $x^3 = y^3 \Rightarrow x = y$. But $2^3 \equiv 1^3 \bmod 7$ and $2 \not\equiv 1 \bmod 7$.

When we add and multiply modulo m, we are carrying out *modular arithmetic*.

Theorem 3.8. *If $ax \equiv ay \bmod m$ and $(a, m) = 1$ then $x \equiv y \bmod m$.*

Proof. We are told that $m \mid (ax - ay)$, so $m \mid a(x - y)$. Since also $(a, m) = 1$, by a consequence of Bezout's identity $m \mid (x - y)$, so $x \equiv y \bmod m$. □

Example 3.9. The following interpretation of Theorem 3.8 was pointed out to me by Nathaniel Harris. The table of integers below has rows listing 14 consecutive numbers at a time. In *each column*, the multiples of 3 (in red) are 3 rows apart, the multiples of 5 (in orange) are 5 rows apart, and the multiples of 9 (in blue) are 9 rows apart. Why is this?

1	2	3	4	5	6	7	8	9	10	11	12	13	14
15	16	17	18	19	20	21	22	23	24	25	26	27	28
29	30	31	32	33	34	35	36	37	38	39	40	41	42
43	44	45	46	47	48	49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80	81	82	83	84
85	86	87	88	89	90	91	92	93	94	95	96	97	98
99	100	101	102	103	104	105	106	107	108	109	110	111	112
113	114	115	116	117	118	119	120	121	122	123	124	125	126
127	128	129	130	131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	150	151	152	153	154
155	156	157	158	159	160	161	162	163	164	165	166	167	168

The pattern is due to 3, 5, and 9 being relatively prime to 14: if $(a, 14) = 1$ and ax and ay are in the same column, then $ax \equiv ay \bmod 14$, so $x \equiv y \bmod 14$ by Theorem 3.8. Thus x and y differ by a multiple of 14, so the *first* multiple of a after ax in the same column as ax is $a(x + 14) = ax + 14a$: that's a rows after ax. For $a = 2$, $a = 4$, and $a = 7$, all not relatively prime to 14, the nearest multiples of a in a column are less than a rows apart.

That addition and multiplication can be carried out on integers modulo m without having the answer change (modulo m) if we replace an integer by a congruent integer is similar to other computations in mathematics. In elementary school, you learned that addition and multiplication of fractions is independent of the particular choice of numerator and denominator for the fractions, *e.g.*, $1/2 + 3/5 = 11/10$ and $2/4 + 9/15 = 66/60 = 11/10$.

In calculus, the formula $\int_a^b f(x)\,dx = F(b) - F(a)$ for definite integrals is independent of the choice of antiderivative $F(x)$ for $f(x)$: any two antiderivatives for $f(x)$ on $[a, b]$ differ by a constant, so the difference $F(b) - F(a)$ doesn't change if we change $F(x)$ to another antiderivative of $f(x)$: $(F(b) + C) - (F(a) + C) = F(b) - F(a)$.

Definition 3.10. The integers mod m under addition and multiplication is denoted $\mathbf{Z}/(m)$.

Other notations you may meet for $\mathbf{Z}/(m)$ are $\mathbf{Z}_m$ and $\mathbf{Z}/m\mathbf{Z}$.

Example 3.11. Here are the elements of $\mathbf{Z}/(5)$:

$$\{\dots, -15, -10, -5, \mathbf{0}, 5, 10, 15, \dots\},$$
$$\{\dots, -14, -9, -4, \mathbf{1}, 6, 11, 16, \dots\},$$
$$\{\dots, -13, -8, -3, \mathbf{2}, 7, 12, 17, \dots\},$$
$$\{\dots, -12, -7, -2, \mathbf{3}, 8, 13, 18, \dots\},$$
$$\{\dots, -11, -6, -1, \mathbf{4}, 9, 14, 19, \dots\}.$$

These five sets consist of integers congruent to each other mod 5 and are called *congruence classes* mod 5. A representative from each congruence class usually stands for the whole congruence class. In bold type is a set of representatives for the congruence classes mod 5: 0, 1, 2, 3, and 4. That is the standard choice. Another set of representatives is 5, 6, -3, 18, and -6 (since $5 \equiv 0 \bmod 5$, $6 \equiv 1 \bmod 5$, $-3 \equiv 2 \bmod 5$, and so on).

Integers in the same congruence class are like real numbers representing the same angle, such as π and 3π. They're different ways of representing the "same thing". Just as π and 3π are not the same in $\mathbf{R}$ but become "the same" in as angles, think that way about congruence classes: 1 and 6 are not the same in $\mathbf{Z}$ but they are the same modulo 5.

4. First application: Divisibility tests

A basic application of modular arithmetic is the explanation of divisibility tests in terms of digits. We will look at three such tests. They describe when a positive integer n is divisible by certain numbers in terms of its base 10 expansion

$$n = c_k 10^k + c_{k-1}10^{k-1} + \cdots + c_1 \cdot 10 + c_0, \tag{4.1}$$

where $0 \le c_i \le 9$. For example, $n = 7405$ has $c_3 = 7$, $c_2 = 4$, $c_1 = 0$, and $c_0 = 5$.

Theorem 4.1. *The number n is divisible by 3 if and only if the sum of its digits $\sum_{i=0}^k c_i$ is divisible by 3.*

Proof. To say n is divisible by 3 is the same as saying $n \equiv 0 \bmod 3$. Thus, we will show $n \equiv 0 \bmod 3$ if and only if $\sum_{i=0}^k c_i \equiv 0 \bmod 3$.

The key point is this: $10 \equiv 1 \bmod 3$. Multiplying both sides of this congruence by themselves, $10^2 \equiv 1^2 \bmod 3$, so $10^2 \equiv 1 \bmod 3$. In the same way, $10^i \equiv 1 \bmod 3$ for all $i \ge 0$ by induction. Then

$$c_k 10^k + c_{k-1}10^{k-1} + \cdots + c_1 \cdot 10 + c_0 \equiv c_k + c_{k-1} + \cdots + c_1 + c_0 \bmod 3$$

since each 10^i can be replaced with 1 when we work mod 3. Since $3 \mid n$ if and only if $n \equiv 0 \bmod 3$, which is equivalent to $\sum_{i=0}^k c_i \equiv 0 \bmod 3$, n being divisible by 3 is equivalent to the sum of the digits of n being divisible by 3. □

Example 4.2. Let $n = 84435$. The sum of the digits is $8 + 4 + 4 + 3 + 5 = 24$, which is divisible by 3, so 84435 is divisible by 3. Indeed, $84435 = 3 \cdot 28145$.

Theorem 4.3. *The number n is divisible by 9 if and only if the sum of its digits $\sum_{i=0}^{k} c_i$ is divisible by 9.*

Proof. The *only* important property of 3 in the proof of Theorem 4.1 was the condition $10 \equiv 1 \bmod 3$. Since $10 \equiv 1 \bmod 9$ too, we get the same test for divisibility by 9 as we did for divisibility by 3, with the same proof. □

Example 4.4. Again taking $n = 84435$, the sum of the digits is 24, which is not divisible by 9, so n is not divisible by 9 either. In fact, 84435 leaves a remainder of 6 when divided by 9: $84435 = 9 \cdot 9381 + 6$.

Theorem 4.5. *The number n is divisible by 11 if and only if the alternating sum of its digits $\sum_{i=0}^{k}(-1)^i c_i$ is divisible by 11.*

Proof. We know n is divisible by 11 if and only if $n \equiv 0 \bmod 11$. Write n as in (4.1). Since $10 \equiv -1 \bmod 11$, we can rewrite the base 10 expansion of n, viewed modulo 11, as

$$n \equiv c_k(-1)^k + c_{k-1}(-1)^{k-1} + \cdots + c_1(-1) + c_0 \bmod 11.$$

This is the alternating sum of the digits, so $n \equiv 0 \bmod 11$ if and only if the alternating sum of the digits is $\equiv 0 \bmod 11$, which means the alternating sum is divisible by 11. □

Example 4.6. If $n = 84d35$, which digit d permits $11 \mid n$? We need $8 - 4 + d - 3 + 5 \equiv 0 \bmod 11$, or equivalently $6 + d \equiv 0 \bmod 11$. Then $d \equiv 5 \bmod 11$, so we need $d = 5$ since d is a decimal digit. Thus 84535 is a multiple of 11. As an explicit check, $84535 = 11 \cdot 7685$.

5. Solving equations in $\mathbf{Z}/(m)$

In school you learned how to solve polynomial equations like $2x+3 = 8$ or $x^2-3x+1 = 0$ by the "rules of algebra": cancellation (if $a \neq 0$ and $ax = ay$ then $x = y$), equals plus equals are equal, and so on. When there is no simple formula for a solution, you might graph the equation and estimate solutions numerically. This is in the setting of the real numbers.

We can also try to solve polynomial equations in modular arithmetic, counting solutions as different if they are incongruent to each other. We will focus on the simplest case of a linear congruence $ax \equiv b \bmod m$. Already in this case we will meet some phenomena with no parallel in the case of a real linear equation (Examples 5.2 and 5.3 below).

Example 5.1. Let's try to solve $8x \equiv 1 \bmod 11$. If there is an answer, it can be represented by one of $0, 1, 2, \ldots, 10$, so let's run through the possibilities:

$x \bmod 11$	0	1	2	3	4	5	6	7	8	9	10
$8x \bmod 11$	0	8	5	2	10	7	4	1	9	6	3

The only solution is 7 mod 11: $8 \cdot 7 = 56 \equiv 1 \bmod 11$.

That problem concerned finding an inverse for 8 modulo 11. We can find multiplicative inverses for every nonzero element of $\mathbf{Z}/(11)$:

x	1	2	3	4	5	6	7	8	9	10
x^{-1}	1	6	4	3	9	2	8	7	5	10

Check in each case that the product of the numbers in each column is 1 in $\mathbf{Z}/(11)$.

Example 5.2. Find a solution to $8x \equiv 1 \bmod 10$. We run through the standard representatives for $\mathbf{Z}/(10)$, and find *no* answer:

x	0	1	2	3	4	5	6	7	8	9
$8x$	0	8	6	4	2	0	8	6	4	2

In retrospect, we can see *a priori* why there shouldn't be an answer. If $8x \equiv 1 \bmod 10$ for some integer x, then we can lift the congruence up to $\mathbf{Z}$ in the form

$$8x + 10y = 1$$

for some $y \in \mathbf{Z}$. But this is absurd: $8x$ and $10y$ are even, so the left side is a multiple of 2 but the right side is not.

Example 5.3. The linear congruence $6x + 1 \equiv 4 \bmod 15$ has three solutions! In the following table we can see the solutions are 3, 8, and 13:

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$6x+1$	1	7	13	4	10	1	7	13	4	10	1	7	13	4	10

These examples show us that a linear congruence $ax \equiv b \bmod m$ doesn't have to behave like real linear equations: there may be no solutions or more than one solution. In particular, taking $b = 1$, we can't always find a multiplicative inverse for each nonzero element of $\mathbf{Z}/(m)$.

The obstruction to inverting 8 in $\mathbf{Z}/(10)$ can be extended to other moduli as follows.

Theorem 5.4. *For integers a and m, the following three conditions are equivalent:*

- *there is a solution x in $\mathbf{Z}$ to $ax \equiv 1 \bmod m$,*
- *there are solutions x and y in $\mathbf{Z}$ to $ax + my = 1$,*
- *a and m are relatively prime.*

Proof. Suppose $ax \equiv 1 \bmod m$ for some $x \in \mathbf{Z}$. Then $m \mid (1 - ax)$, so there is some $y \in \mathbf{Z}$ such that $my = 1 - ax$, so

$$ax + my = 1.$$

This equation implies a and m are relatively prime since any common factor of a and m divides $ax + my$. Finally, if a and m are relatively prime then by Bezout's identity (a consequence of back-substituting in Euclid's algorithm) we can write $ax + my = 1$ for some x and y in $\mathbf{Z}$ and reducing both sides mod m implies $ax \equiv 1 \bmod m$. □

This explains Example 5.2, since 8 and 10 have a common factor of 2. Similarly, there is no solution to $3x \equiv 1 \bmod 15$ (common factor 3) or $35x \equiv 1 \bmod 77$ (common factor 7).

Example 5.5. Since $(8, 11) = 1$, 8 has a multiplicative inverse in $\mathbf{Z}/(11)$. We found it by an exhaustive search in Example 5.1, but now we can do it by a more systematic approach:

Euclid's algorithm

$$\begin{aligned} 11 &= 8 \cdot 1 + 3 \\ 8 &= 3 \cdot 2 + 2 \\ 3 &= 2 \cdot 1 + 1 \end{aligned}$$

Backwards substitution

$$\begin{aligned} 1 &= 3 - 2 \\ &= 3 - (8 - 3 \cdot 2) \\ &= 3 \cdot 3 - 8 \\ &= 3 \cdot (11 - 8) - 8 \\ &= \boxed{3 \cdot 11 - 8 \cdot 4} \end{aligned}$$

Reducing the equation $1 = 3 \cdot 11 - 8 \cdot 4$ modulo 11,

$$8(-4) \equiv 1 \bmod 11.$$

The inverse of 8 in $\mathbf{Z}/(11)$ is -4, or equivalently 7.

To summarize: solving for x in the congruence $ax \equiv 1 \bmod m$ is equivalent to solving for integers x and y in the equation $ax + my = 1$ (be sure you see why!), and the latter equation can be solved without any guesswork by reversing Euclid's algorithm on a and m when $(a, m) = 1$. If Euclid's algorithm shows $(a, m) \neq 1$, then there is no solution.

In the real numbers, every nonzero number has a multiplicative inverse. This is not generally true in modular arithmetic: if $a \not\equiv 0 \bmod m$ it need not follow that we can solve $ax \equiv 1 \bmod m$. (For instance, $4 \not\equiv 0 \bmod 6$ and $4 \bmod 6$ has no multiplicative inverse.) The correct test for invertibility in $\mathbf{Z}/(m)$ is $(a, m) = 1$, which is generally stronger than $a \not\equiv 0 \bmod m$. Although invertibility in $\mathbf{Z}/(m)$ is usually not the same as being nonzero in $\mathbf{Z}/(m)$, there is an important case when these two ideas agree: m is prime.

Corollary 5.6. *For prime p, an integer a is invertible in $\mathbf{Z}/(p)$ if and only if $a \not\equiv 0 \bmod p$.*

Proof. If $a \bmod p$ is invertible, then $(a, p) = 1$, so p does not divide a.

For the converse direction, suppose $a \not\equiv 0 \bmod p$. We show $(a, p) = 1$. Since (a, p) is a (positive) factor of p, and p is *prime*, (a, p) is either 1 or p. (The proof would break down here if p were not prime.) Since p does not divide a, $(a, p) \neq p$, so $(a, p) = 1$. Therefore the congruence $ax \equiv 1 \bmod p$ has a solution. □

The upshot of Corollary 5.6 is that our intuition from algebra over $\mathbf{R}$ carries over quite well to algebra over $\mathbf{Z}/(p)$: every nonzero number has a multiplicative inverse in the system. But $\mathbf{Z}/(m)$ for composite m is more delicate.

Why do we want to invert integers in $\mathbf{Z}/(m)$? (By "inverting" we always mean "inverting multiplicatively.") One reason is its connection to inverting matrices with entries in $\mathbf{Z}/(m)$. Your experience with linear algebra in $\mathbf{R}$ may suggest a square matrix with entries in $\mathbf{Z}/(m)$ is invertible whenever its determinant is nonzero in $\mathbf{Z}/(m)$, but that is *false*.

Example 5.7. We work with matrices having entries in $\mathbf{Z}/(10)$. Let $A = \left(\begin{smallmatrix} 1 & 3 \\ 1 & 1 \end{smallmatrix}\right)$. The determinant of A is $-2 \equiv 8 \bmod 10$, so $\det A \not\equiv 0 \bmod 10$. However, there is no inverse for A as a mod 10 matrix. We can see why by contradiction. Suppose there is an inverse matrix, and call it B. Then $AB \equiv \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \bmod 10$. (Congruence of matrices means congruence of corresponding matrix entries on both sides.) Writing $B = \left(\begin{smallmatrix} x & y \\ z & t \end{smallmatrix}\right)$, we compute AB to get $\left(\begin{smallmatrix} x+3z & y+3t \\ x+z & y+t \end{smallmatrix}\right) \equiv \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \bmod 10$. Then $x + 3z \equiv 1 \bmod 10$ and $x + z \equiv 0 \bmod 10$. The second congruence says $x \equiv -z \bmod 10$, and replacing x with $-z$ in the first congruence yields $2z \equiv 1 \bmod 10$. But that's absurd: $2z$ is even and 1 is odd, so $2z \not\equiv 1 \bmod 10$. (Said differently, if $2z \equiv 1 \bmod 10$ then $2z = 1 + 10y$ for some integer y, so $2z - 10y = 1$, but the left side is even and 1 is not even.)

As a real matrix, A is invertible and $A^{-1} = \left(\begin{smallmatrix} -1/2 & 3/2 \\ 1/2 & -1/2 \end{smallmatrix}\right)$. This inverse makes no sense if we try to reduce it modulo 10 (what is $1/2 \bmod 10$?!?), and that suggests there should be a problem if we try to invert A as a mod 10 matrix.

Let's look at determinants in modular arithmetic. Suppose $n \times n$ matrices A and B satisfy $AB \equiv I_n \bmod m$. Taking determinants of both sides tells us (by Theorem 3.2) that

$$(\det A)(\det B) \equiv 1 \bmod m,$$

so $\det A$ is *invertible* in $\mathbf{Z}/(m)$. Invertibility of $\det A$ in $\mathbf{Z}/(m)$ is usually stronger than $\det A \not\equiv 0 \bmod m$. For instance, the 2×2 matrix A in Example 5.7 has determinant $8 \bmod 10$, which is not invertible. Thus A is not invertible mod 10. That is an easier way to see A is not invertible than the calculations in Example 5.7!

Example 5.8. Let $m = 14$ and $A = \left(\begin{smallmatrix} 1 & 4 \\ 3 & 2 \end{smallmatrix}\right)$ as a matrix with entries in $\mathbf{Z}/(14)$. The determinant of A is $2 - 12 = -10 \equiv 4 \bmod 14$, which is not invertible. Even though A has a nonzero determinant, there is no matrix inverse for A over $\mathbf{Z}/(14)$.

We now see that we have to be able to recognize invertible elements of $\mathbf{Z}/(m)$ before we can recognize invertible matrices over $\mathbf{Z}/(m)$, because an invertible matrix will have an invertible determinant. If we want to do linear algebra over $\mathbf{Z}/(m)$ then we need Euclid's algorithm (and Bezout's identity) to invert determinants.

6. Block ciphers

This last section gives a cryptographic application of linear algebra over $\mathbf{Z}/(m)$. In cryptography, we are interested in various methods of encoding and decoding text.

The simplest idea, going back to Julius Caesar, is called a *shift cipher*. Simply pick an integer c and encode a letter by replacing it with the letter c places further along in the alphabet (wrap around to the start of the alphabet from the end if necessary). For instance, if we shift by 3, then we have

$$\mathrm{A} \mapsto \mathrm{D}, \mathrm{B} \mapsto \mathrm{E}, \mathrm{C} \mapsto \mathrm{F}, \ldots, \mathrm{Z} \mapsto \mathrm{C}.$$

The message

DOODLE

is encoded as

GRRGOH.

The encoded message

XKFJXI

is decoded by shifting *back* by 3 letters to

ANIMAL.

The Caesar cipher has two critical flaws. First of all, since there are only 26 shifts, an adversary (who knows or suspects that the message was encoded with the Caesar cipher) can simply decode the message with each of the 26 possible shifts until a meaningful decryption is found. The second problem, which is illustrated in the examples above, is that a letter gets encoded to the same letter every time. A cipher with this feature (called a simple substitution cipher, whether or not the substitutions come from a cyclic shift of the alphabet) is easily cracked by taking into account the frequency of letters in ordinary English text. For instance, the most common letter in ordinary English text is E, so the most common letter in a text (of reasonable length) encoded by a simple substitution cipher is probably the encrypted form of the letter E.

Before moving on to other cryptographic methods, we describe the Casesar cipher using modular arithmetic. Identify the letters of the alphabet with two-digit strings:

$$\mathrm{A} = 01, \mathrm{B} = 02, \ldots, \mathrm{Z} = 26. \tag{6.1}$$

View the letters as elements of $\mathbf{Z}/(26)$. The shift by 3, say, is the encryption function $E(x) \equiv x + 3 \bmod 26$. It is decrypted with $D(y) \equiv y - 3 \bmod 26$.

We can get something a bit more mathematically interesting than a shift cipher by allowing scaling before translating: $E(x) = ax + b$, where $(a, 26) = 1$. A plain shift has $a = 1$. Suppose $a = 5$ and $b = 3$. Then $E(x) \equiv 5x+3 \bmod 26$, so letting $x = 1, 2, 3, \ldots, 25, 0$ gives the encoding

$$\mathrm{A} \mapsto \mathrm{H}, \mathrm{B} \mapsto \mathrm{M}, \ldots, \mathrm{Z} \mapsto \mathrm{C}.$$

The decryption is based on inverting a linear function: If $y = 5x+3$, then $x = (1/5)(y-3)$. In $\mathbf{Z}/(26)$, the inverse of 5 is 21 (check!), so the decryption function is $D(y) = 21(y-3) = 21y + 15$ (because $12(-3) = -36 \equiv 15 \bmod 26$).

Using $E(x) = 5x + 3$ on $\mathbf{Z}/(26)$, check the encrypted version of DOODLE is now

WZZWKB.

While the introduction of the scaling before we shift enlarges the number of possible ciphers, we are still encoding one letter at a time, and each letter gets encoded in the same way no matter where it appears in the text, so a frequency analysis on a moderate amount of text would let us determine the scaling and shift parameters.

We now use linear algebra over $\mathbf{Z}/(26)$ to encode blocks of text. This is called a *block cipher*. We will describe the idea using two-letter blocks. For instance,

DOODLE

becomes

DO OD LE.

Each block can be viewed as a column vector over $\mathbf{Z}/(26)$:

$$\begin{pmatrix}\text{D}\\ \text{O}\end{pmatrix} = \begin{pmatrix}4\\15\end{pmatrix}, \quad \begin{pmatrix}\text{O}\\ \text{D}\end{pmatrix} = \begin{pmatrix}15\\4\end{pmatrix}, \quad \begin{pmatrix}\text{L}\\ \text{E}\end{pmatrix} = \begin{pmatrix}12\\5\end{pmatrix}.$$

Pick a 2×2 matrix over $\mathbf{Z}/(26)$. For concreteness, take

$$\mathcal{A} = \begin{pmatrix}3 & 2\\ 1 & 11\end{pmatrix}.$$

Our encoding algorithm will be the product of $\mathcal{A}$ with each vector, calculations being made modulo 26:

$$\mathcal{A}\begin{pmatrix}4\\15\end{pmatrix} = \begin{pmatrix}16\\13\end{pmatrix}, \quad \mathcal{A}\begin{pmatrix}15\\4\end{pmatrix} = \begin{pmatrix}1\\7\end{pmatrix}, \quad \mathcal{A}\begin{pmatrix}12\\5\end{pmatrix} = \begin{pmatrix}20\\15\end{pmatrix}.$$

Turning each coordinate back into a letter, the encoded form of DOODLE is

PMAGTO.

Notice that the two D's in DOODLE, as well as the two O's, have been encoded by different letters. (Why did this happen?) Clearly this is a desirable feature from a security standpoint.

Suppose we received the message

UPFOOU

and we know it was encoded by multiplication by the above matrix $\mathcal{A} = \left(\begin{smallmatrix}3 & 2\\ 1 & 11\end{smallmatrix}\right)$. Since encoding is $E(\mathbf{v}) = \mathcal{A}\mathbf{v}$, decoding is achieved by multiplication by the inverse of $\mathcal{A}$: $D(\mathbf{v}) = \mathcal{A}^{-1}\mathbf{v}$. When is a 2×2 matrix invertible? In a first linear algebra course you learn that a (square) matrix with real entries is invertible precisely when its determinant is nonzero, and in the 2×2 case $\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right)$ has inverse

$$\frac{1}{ad-bc}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}.$$

For matrices with entries in $\mathbf{Z}/(m)$, the rule for invertibility of a matrix is not that the determinant is nonzero. For example, taking $m = 26$, the matrix $\left(\begin{smallmatrix}1 & 2\\ 3 & 4\end{smallmatrix}\right)$ has determinant $-2 \equiv 24 \bmod 26$, which is not 0 mod 26, but this matrix is not invertible mod 26: for no 2×2 matrix M is $\left(\begin{smallmatrix}1 & 2\\ 3 & 4\end{smallmatrix}\right)M \equiv \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right) \bmod 26$.

Theorem 6.1. *A 2×2 matrix with entries in $\mathbf{Z}/(m)$ has an inverse precisely when its determinant is invertible in $\mathbf{Z}/(m)$, in which case its inverse is given by the same formula as in the real case.*

Proof. Let A be the matrix. If A has an inverse matrix B, so $AB \equiv \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \bmod m$, then taking determinants of both sides implies $(\det A)(\det B) \equiv 1 \bmod m$. Therefore $\det A$ is invertible in $\mathbf{Z}/(m)$. Conversely, if $\det A$ is invertible in $\mathbf{Z}/(m)$ then the familiar inverse matrix formula from (real) linear algebra actually makes sense if we apply it to the mod m matrix A and a direct calculation shows the formula works as an inverse. □

The matrix $\mathcal{A} = \left(\begin{smallmatrix} 3 & 2 \\ 1 & 11 \end{smallmatrix}\right)$ over $\mathbf{Z}/(26)$ has determinant $31 \equiv 5 \bmod 26$. The inverse of this determinant (in $\mathbf{Z}/(26)$) is 21. Therefore

$$\mathcal{A}^{-1} = 21\begin{pmatrix} 11 & -2 \\ -1 & 3 \end{pmatrix} \equiv \begin{pmatrix} 23 & 10 \\ 5 & 11 \end{pmatrix}.$$

Call this matrix $\mathcal{B}$. (Check $\mathcal{AB} = I_2$ and $\mathcal{BA} = I_2$.) The endoded message UPFOOU has blocks

$$\begin{pmatrix} \mathrm{U} \\ \mathrm{P} \end{pmatrix} = \begin{pmatrix} 21 \\ 16 \end{pmatrix}, \quad \begin{pmatrix} \mathrm{F} \\ \mathrm{O} \end{pmatrix} = \begin{pmatrix} 6 \\ 15 \end{pmatrix}, \quad \begin{pmatrix} \mathrm{O} \\ \mathrm{U} \end{pmatrix} = \begin{pmatrix} 15 \\ 21 \end{pmatrix},$$

so we decode by multiplying each vector by $\mathcal{B}$ on the left:

$$\mathcal{B}\begin{pmatrix} 21 \\ 16 \end{pmatrix} = \begin{pmatrix} 19 \\ 21 \end{pmatrix}, \quad \mathcal{B}\begin{pmatrix} 6 \\ 15 \end{pmatrix} = \begin{pmatrix} 2 \\ 13 \end{pmatrix}, \quad \mathcal{B}\begin{pmatrix} 15 \\ 21 \end{pmatrix} = \begin{pmatrix} 9 \\ 20 \end{pmatrix}.$$

Putting the output blocks back together and converting them into letters gives the decoded message:

SUBMIT.

Of course, 2-block ciphers can be cracked using a frequency analysis on pairs of letters rather than on individual letters. (As an example, Q is almost always followed by U in English.) The idea of a 2-block cipher extends to blocks of greater length using larger matrices. A difficulty with a block cipher is that knowledge of the encryption and decryption matrices are essentially equivalent: anyone who discovers the encryption matrix can compute its inverse (the decryption matrix) using row reduction or other methods of computing matrix inverses. This means the encryption matrix must be kept secret from your enemies. Until the 1970s, everyone (outside US and British intelligence) thought encryption and decryption had to be more or less symmetric processes. But this is false: there are cryptographic algorithms where the encryption process can be made public to the world and decryption is still secure. Number theory is a source of such algorithms used in e-commerce, ATM transactions, cell phone communications, *etc.*

We end with some exercises.

1. Explain why the matrix

$$\begin{pmatrix} 5 & 3 \\ 7 & 1 \end{pmatrix}$$

can't be used in a 2-block cipher. (Hint: Try to invert the matrix over $\mathbf{Z}/(26)$.)

2. Messages are encoded in blocks of length 2 using

$$\begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$$

You receive the encrypted message

AAOGIXZC.

What is the original message, in plain English?