Sundry Notes on Modular Arithmetic

Óscar Pereira

Abstract. Even back when I was first learning modular arithmetic, I thought most texts explained it in a way that is needlessly confusing. Later on, when I began teaching the subject, I re-encountered the same feeling—especially with non-math majors. This is my attempt at a less confusing explanation.

1 Introduction

The goal is to offer a self-contained explanation of modular arithmetic. The crux of modular arithmetic is that we can *loosen* the notion of equality of two integers, in the following sense: given a positive integer n, we consider the integers $0, n, 2n, 3n, \ldots$ to be all "equal", in some sense to be defined precisely below. Similarly, $1, 1+n, 1+2n, \ldots$ are also all equal, according to this new sense of equality—although their are *different from the previous class*! I.e., $1, 1+n, 1+2n, \ldots$ are all equivalent, but their are *not* equivalent to $0, n, \ldots$!

The main result is that we can do arithmetic with these "meta-integers" as if they were regular integers, even though they are not—with only a small caveat, which is the following. In regular integer arithmetic, we cannot have divisions by 0; here this is also true, but due to our extended notion of equality, this restriction has also to be suitably extended.

2 Congruences

The reader is assumed to be familiar with the integer division theorem (see the appendix for a proof, if needed):

Theorem 2.1 (Integer division). *Given integers a and $b > 0$, there exist unique integers q and r, with $0 \leq r < b$, such that $a = bq + r$.*

Computing the remainder of a number a, when divided by a number n, is called computing the remainder **modulo** n. We represent it as $a \mod n$.[1] Also related to integer division, is the notion of *divisibility*:

Definition 2.2. *Given two integers a and b, we say that a* **divides** *b, or equivalently, that b* **is a multiple of** *a, if there exists an integer k such that $b = ka$. In this case we write symbolically that $a \mid b$.*

We now come to the central notion of modular arithmetic: **congruence relations**.

Date: 27th January 2021. *Contact information:* {https://, oscar@}randomwalk.eu.

[1] In programming languages such as Python, this is computed as `a % n`.

Definition 2.3. *Given two integers a and b, and a positive integer $n \geq 2$, we say that a and b are* **congruent modulo** *n, if $n \mid (a-b)$.*[2] *This is denoted as $a \equiv b \pmod{n}$.*[3]

So congruences can be thought of as a "weaker" form of equality. In particular, for any modulus, $a = b$ implies $a \equiv b$—though the contrary is clearly false (e.g. $21 \equiv 5 \pmod{2}$, but $21 \neq 5$).

Congruence relations are a particular case of something that mathematicians call *equivalence relations*. By definition, they always have three properties, viz. reflexivity, symmetry, and transitivity. The next result illustrates what this means in the case of congruences.

Theorem 2.4. *For any modulus n, congruence relations are reflexive ($a \equiv a$), symmetric ($a \equiv b$ if and only if $b \equiv a$) and transitive ($a \equiv b$ and $b \equiv c$ imply $a \equiv c$).*

Proof. For any integer a, we have $a \equiv a \iff a - a = 0n$, which establishes reflexivity. Next, $a \equiv b$ means that $a - b = kn$, for some integer k. This is equivalent to $b - a = (-k)n$, which shows that $b \equiv a$. The same reasoning can be used to show that from $b \equiv a$, we obtain $a \equiv b$, thus settling symmetry. And finally, if $a \equiv b$ and $b \equiv c$, then this means that $b = a + kn$ and $c - b = k'n$. Replacing b in the latter equation yields $c - (a + kn) = k'n \iff c - a = n(k' + k) \iff a \equiv c$—which establishes transitivity. ∎

For an integer a and modulus n, the set of all integers that are congruent to a modulo n, are called the **equivalence class** of a modulo n. We represent it as $[a]_n$ (again, we can suppress the modulus if it is clear from context). When denoting an equivalence class by $[a]_n$, a is said to be the **representative** of the class. Recall the introductory paragraph. It should be clear by now that $0, n, 2n, \ldots$ are all in the same equivalence class. Similarly, $1, 1+n, 1+2n, \ldots$ are all in the same equivalence class—although it is a *different* equivalence class from that of $0, n, \ldots$. Using the notation introduced above, we would say that $1, 1+n, 1+2n, \ldots$ all belong to equivalence class $[1]_n$, which is the same as $[1+n]_n$, etc. And similarly, $0, n, \ldots$ all belong to equivalence class $[0]_n$, which is the same as equivalence class $[0+n]_n$, etc.

Congruences and integer division. Thanks to the division theorem, given any integer a, we can can always find another integer r, such that $a \equiv r \pmod{n}$ and $0 \leq r < n$—it should be easy to see that r is the remainder of the integer division of a by n. Such an r is said to be the **canonical representative** of the equivalence class $[a]_n$. This effectively partitions the integers $\mathbf{Z}$ into n equivalence classes, $[0]_n, [1]_n, \ldots, [n-1]_n$—the set of these classes is usually denoted $\mathbf{Z}_n$. (The integers $0, 1, \ldots, n-1$ are the canonical representatives of (the classes composing) $\mathbf{Z}_n$.) Any integer belongs to one, and only one, equivalence class—in fact, this is a property of all equivalence relations. Indeed, any given integer a naturally belongs to the class $[a]_n$. But furthermore, the equivalence classes are (pairwise) disjoint—and hence, if an integer a belongs to equivalence class $[a]_n$, it belongs *only* to that class. To see this, note that, if $[i]_n$ and $[j]_n$, with $i \neq j$, were not disjoint, i.e. they had least one common element, then thanks to transitivity, any element of $[i]_n$ would be equivalent—i.e., congruent—to any element of $[j]_n$, and vice-versa. In particular, we would have $i \equiv j$, which is the same as saying that $[i]_n$ and $[j]_n$ are the *same* equivalence class. This is a contradiction, because of the assumption that $i \neq j$—and hence, we conclude they are disjoint.

Modular arithmetic defines suitably adjusted counterparts to the usual operations of arithmetic, that apply to these sets of integers—$[0]_n, [1]_n, \ldots, [n-1]_n$—instead of to the integers themselves. We take on this task in the next section.

[2] There is not much point of having either $n = 1$ or $n = 0$. In the latter case, any integer is congruent only with itself, meaning that congruence ($\equiv$) reduces to ordinary equality ($=$). In the former case ($n = 1$), any integer is congruent with all other integers—the difference of any two integers is always a multiple of 1—which is not particularly interesting.

[3] The modulus can be omitted when clear from context.

3 Modular arithmetic

We want to define a binary operation on the set $\mathbf{Z}_n = \{[0]_n, [1]_n, \ldots, [n-1]_n\}$. We do so as follows.

Definition 3.1. *Given a modulus n, we define addition in $\mathbf{Z}_n$ as $[a]_n + [b]_n = [a+b]_n$. Multiplication in $\mathbf{Z}_n$ is defined similarly: $[a]_n[b]_n = [ab]_n$.*

This seems like a natural enough way of extending the non-modular versions of addition and multiplication. If nothing else, when $n = 0$, definition 3.1 coincides with regular addition and multiplication (cf. footnote 2 on page 2). However, we need to show that these operations are *well defined*: for example, $a+n$ also belongs to $[a]_n$, and $b+3n$ also belongs to $[b]_n$, and so for the operation of modular addition to be well defined, we must show that $a+n+b+3n$ belongs to $[a+b]_n$. In other words, we need to show that the result of the operation does *not* depend on the chosen representatives of the equivalence classes. This follows immediately from the next theorem.

Theorem 3.2. *With modulus n, suppose that $a \equiv a'$, and $b \equiv b'$. Then $a + b \equiv a' + b'$ and $ab \equiv a'b'$.*

Proof. **Addition.** We have $a' = a + k_1 n$ and $b' = b + k_2 n$, for some integers k_1 and k_2. Hence $a' + b' = (a + k_1 n) + (b + k_2 n) = (a + b) + n(k_1 + k_2)$ which shows that $a + b \equiv a' + b'$.

Multiplication. $a'b' = (a + k_1 n)(b + k_2 n) = ab + ak_2 n + k_1 nb + k_1 nk_2 n = ab + n(ak_2 + k_1 b + k_1 nk_2)$, and thus $ab \equiv a'b'$. ∎

This result shows that when computing $[a]_n + [b]_n$, we can replace a and b with any elements in their respective equivalence classes, because the result will still belong to the equivalence class of $[a+b]_n$. Similar remarks apply to multiplication.

Suppose you have a congruence like $ax \equiv b \pmod{n}$. If you can find a' such that $aa' \equiv 1 \pmod{n}$, then the above theorem says you can multiply both sides of the former congruence by a', thus obtaining $x \equiv ba' \pmod{n}$. Furthermore, any integer is congruent with itself (reflexivity), and so given a congruence like $c \equiv d \pmod{n}$, you can add or multiply both sides by any integer, and the congruence will still hold. You can also *subtract*: subtracting a is the same as adding $-a$, and $-a \equiv n - a \pmod{n}$, so it is just as adding $n - a$ to both sides. However, **modular division is not always possible**—this will be the topic of §4.

Before going there, however, there is another property that is of immense utility in practice. From integer division, we have $a = nq + r$, and hence $a \mod n = r = a - nq$. Thus:

$$\big[a \equiv (a \bmod n)\big] \pmod{n} \tag{1}$$

This allows to compute the remainder of very large numbers much more efficiently. Indeed, we have that:

$$(a+b) \pmod{n} \equiv a + b \equiv (a \bmod n) + (b \bmod n) \equiv \\ [(a \bmod n) + (b \bmod n)] \pmod{n}$$

And similarly,

$$(ab) \pmod{n} \equiv ab \equiv (a \bmod n)(b \bmod n) \equiv \\ [(a \bmod n)(b \bmod n)] \pmod{n}$$

These properties also show that when there are more than two factors, we can do things "piecewise". I.e., take the modulus as we go along the sum:

$$(a+b+c) \pmod n \equiv (a+b)+c \equiv \big[(a+b) \bmod n\big]+c \equiv \Big\{\big[(a+b) \bmod n\big]+c\Big\} \pmod n$$

$$(abc) \pmod n \equiv (ab)c \equiv \big[(ab) \bmod n\big]c \equiv \Big\{\big[(ab) \bmod n\big]c\Big\} \pmod n$$

So, for example, whenever we have (to compute the remainder of) an expression that consists of sums of products, we can just compute the remainder of all parcels, and then the remainder of the full expression. A typical example is the rule to "cast out nines": as any integer can be written in the form $\sum d_i 10^i, 0 \leq d_i \leq 9$, and as $10 \equiv 1 \pmod 9$, to compute the remainder of the division of that integer by 9, we just sum the digits, casting out nines wherever possible—which is a lot simpler than remaindering over the whole integer.

Modular vs. non-modular operations. I finish this section with an important remark. If $[a]_n + [b]_n = [c]_n$, then any integer in $[c]_n$ can be written as the *non-modular* sum of an integer in $[a]_n$ and an integer in $[b]_n$ (exercise: show this). However, *this is not a requirement for the modular addition operation to well-defined!* And indeed, this does not happen with modular multiplication: for example $[4]_7 \times [4]_7 = [2]_7$, and $23 \in [2]_7$. But 23 is prime, and hence it cannot be written as the non-modular product of an integer in $[4]_7$ by another integer in $[4]_7$.

For modular addition, to be well-defined only means that the non-modular addition of any integer in $[a]_n$ with any other integer in $[b]_n$, will yield an integer that belongs to $[c]_n$—indeed, this is what it means to say that modular addition is independent of the chosen class representatives. The same also applies, *mutatis mutandis*, to modular multiplication.

4 Linear congruences, and multiplicative inverses

Consider again a congruence of the form: $ax \equiv b \pmod n$, with a, b and n integers, and x an unknown integer variable. This is called a *linear congruence*, due to its similarity with linear equations like $ax = b$. It was mentioned in the previous section that if we can find a' such that $aa' \equiv 1 \pmod n$, then we can re-write the first congruence in the form $x \equiv ba' \pmod n$—in fact, this can be seen as *solving* the linear congruence, because we now know to which equivalence class x belongs to, namely $[ba']_n$. However, such an a' does not always exist—and this is the reason why, in the previous section, it was stated that modular division, here understood as multiplication by the modular inverse, is not always possible. In fact, the modular inverse of a modulo n exists if and only if the *greatest common divisor* (gcd) of a and n is 1. To be able to understand why this is so, however, we need some extra theory.

Definition 4.1. *Given two integers, not simultaneously zero, a and b, their* **greatest common divisor** *(gcd) is the greatest non-negative integer d such that it divides both a and b. If $a = b = 0$, then we define $gcd(0,0) = 0$.*

From this way of defining the gcd we get that $gcd(a,0) = gcd(0,a) = |a|$ holds for any integer a.

Theorem 4.2. *Let a and b be two integers, and let $d = gcd(a, b)$. Then $d = xa + yb$, for some $x, y \in \mathbf{Z}$. Furthermore, every other common divisor of both a and b, also divides d.*

Proof: see the appendix.

Remark 4.3. Given integers a, b, both their gcd d, as well as the integers x, y that allow us to write $d = ax + by$, are found via the *Extended Euclidean Algorithm*.[4] △

Corollary 4.4. *An integer r can be written as $r = as + bt$, if and only if $gcd(a, b) \mid r$.*

Proof. As $gcd(a, b) = as + bt$ for some s, t, it is obvious that any multiple of the gcd can also be written as a linear combination of a and b. Conversely, any linear combination of a and b is divisible by any common divisor of a and b, and in particular by the gcd. ■

Remark 4.5. If a and b are two integers not simultaneously 0, then $gcd(a, b)$ is the smallest positive integer that can be written as a linear combination of a and b. Indeed this follows immediately from corollary 4.4. (If $a = b = 0$, then by definition 4.1, $gcd(a, b) = 0$.) △

Theorem 4.6. *If a and b are integers, then $gcd(a, b) = 1$ if and only if $xa + yb = 1$, for some integers x and y.*

Proof. If $gcd(a, b) = 1$, then by theorem 4.2, $xa + yb = 1$, for some $x, y \in \mathbf{Z}$. Conversely if $xa + yb = 1$, then any common divisor of a and b must divide 1, which implies that the only non-negative common divisor of a and b is 1—and so $gcd(a, b) = 1$. ■

The case where the gcd of two integers is 1 is so important, it has its own name. It is of fundamental importance in algebra and number theory.

Definition 4.7. *Two integers a, b such that $gcd(a, b) = 1$ are said to be* **relatively prime**.

Note that according to this definition, the only integers that are relatively prime to 0 are 1 and -1.

Back to modular arithmetic. Theorem 4.6 makes it clear why the modular inverse of a modulo n exists if and only if $gcd(a, n) = 1$: if $gcd(a, n) \neq 1$, then there is no integer a' that verifies a linear combination of the form $aa' + ny = 1$ (for any y), and so there is no a' such that $aa' \equiv 1 \pmod{n}$. Also note that $[0]_n$ never has a multiplicative inverse.[5]

So returning to our linear congruence, $ax \equiv b \pmod{n}$, if $gcd(a, n) = 1$, then it has a solution. However, even if $gcd(a, n) \neq 1$, it might be possible to solve the linear congruence—the next result makes this clear.

Theorem 4.8. *Let $ax \equiv b \pmod{n}$ be a linear congruence. It has a solution if and only if $d \mid b$, where $d = gcd(a, n)$.*

Proof. Suppose there is a solution, that is, there exists an equivalence class $[z]_n$ for which it holds that $az \equiv b \pmod{n}$. From the definition of congruence, we can write $az + nk = b$, for some integer k. By corollary 4.4, this is equivalent to having $d \mid b$. ■

Note that if $d = 1$, then $d \mid b$ is true for any b.

[4] In the computer algebra system Sage (https://www.sagemath.org), this is implemented by the function `xgcd`.

[5] Except if $n = 1$, in which case every integer a verifies $0a \equiv 1 \pmod{1}$. In fact, $0 \equiv 1 \pmod{1}$, so one could say that it is not even possible to define the notion of a multiplicative inverse... Which is another good reason to rule out this case.

5 Modular cancellation laws

From theorem 3.2, we know that if $b \equiv c \pmod{n}$, then $a+b \equiv a+c \pmod{n}$—after all, for any integer a, and any modulus n, $a \equiv a \pmod{n}$ is always true. Conversely, if $a+b \equiv a+c \pmod{n}$, then by the definition of congruence, $(a+b)-(a+c) = nk \iff b-c = nk \iff b \equiv c \pmod{n}$. So we conclude that:

$$\left[a+b \equiv a+c \pmod{n}\right] \quad \text{if and only if} \quad \left[b \equiv c \pmod{n}\right] \tag{2}$$

Now for multiplication, we reason as for addition: also from theorem 3.2, we know that if $b \equiv c \pmod{n}$, then $ab \equiv ac \pmod{n}$. For the converse however, we need to recall that, as explained a couple of paragraphs above, a has an inverse modulo n if and only if $gcd(a,n) = 1$. And hence, if $ab \equiv ac \pmod{n}$, we can "cancel" a only if it is co-prime to n. Assuming this is the case, we obtain $a(b-c) = nk$ and multiplying both sides by the modular inverse of a, denote it a', we obtain $aa'(b-c) = nka'$. And as $aa' \equiv 1 \pmod{n}$ (this is what being a modular inverse means, after all), again applying theorem 3.2 comes $b-c \equiv nk'a \Rightarrow b \equiv c \pmod{n}$. So we have shown two things. First, that for any a:

$$\text{if} \quad \left[b \equiv c \pmod{n}\right] \quad \text{then} \quad \left[ab \equiv ac \pmod{n}\right] \tag{3}$$

And second, **only for integers a co-prime to n**, we have:

$$\text{if} \quad \left[ab \equiv ac \pmod{n}\right] \quad \text{then} \quad \left[b \equiv c \pmod{n}\right] \tag{4}$$

From modular congruences to remaindering. The two operations are related as follows (see proof in the appendix):

Theorem 5.1. *$a \bmod n = b \bmod n$ if and only if $a \equiv b \pmod{n}$.*

With this result, property (2) yields:

$$\left[(a+b) \mod n = (a+c) \mod n\right] \quad \text{if and only if} \quad \left[b \mod n = c \mod n\right] \tag{5}$$

And properties (3) and (4), together with the restriction **that a is co-prime to n**, yield:

$$\left(ab \mod n = ac \mod n\right) \quad \text{if and only if} \quad \left(b \mod n = c \mod n\right) \tag{6}$$

So to sum up, "cancellation" applies the same way to equality and remainder ($a \mod n = b \mod n$), and to congruences ($a \equiv b \pmod{n}$)—which intuitively is what one would except, seeming as theorem 5.1 basically says that both things are equivalent.

And of course, cancelling is the basic technique of equation solving—so one can solve equations involving mod and equality ($=$) by forgetting the mod, and replacing the equality with an equivalence ($\equiv$).

6 Conclusion

Hopefully the information presented in this paper will help the reader/student find his way when reading about modular arithmetic in more advanced books. As a way to help "wheat the appetite" about said more advanced topics, I will end by providing one more bit of information: we can take the set $\mathbf{Z}_n = \{[0]_n, \ldots, [n-1]_n\}$, and endow it with the sum and multiplication operations given in definition 3.1, to form algebraic structures known as groups, rings or fields. Then we have a choice, of either working with integers and congruences, or with those algebraic structures. Ultimately, it boils down to a matter of taste, preference, and convenience. Happy readings!

A Proofs

Theorem 2.1. *Given integers a and $b > 0$, there exist unique integers q and r, with $0 \leq r < b$, such that $a = bq + r$.*

Proof. Consider the set $S = \{ a - bt \mid t \in \mathbb{Z} \text{ and } a - bt \geq 0 \}$. S is nonempty: if $a \geq 0$, set $t = 0$, otherwise set $t = a$, to obtain $a - ba = a(1 - b)$ which is non-negative, because the first multiplicative factor is negative, and the other is either zero, or negative. From the well-ordering principle, it follows that S must have a smallest element; let that element be $r = a - bq$. We need to show that $r < b$ and that r and q are unique.

To show that $r < b$, suppose that that was *not* the case; i.e. suppose that $r \geq b$. Then $r - b \geq 0$, and as also $r - b = a - bq - b = a - b(q + 1)$, we conclude that $r - b \in S$. But $r - b < r$, and r was supposed to be S's smallest element—which shows our supposition that $r \geq b$ cannot be true. Hence, $r < b$.

To show the uniqueness of q and r, let q', r' be such that $a = bq + r = bq' + r'$ (and $0 \leq r' < b$). We can, without loss of generality, assume that $r' \geq r$.[6] Then, rearranging terms, we obtain $r' - r = b(q - q')$. Thus $b \mid r' - r$, but as $0 \leq r' - r \leq r' < b$, this can only be if $r' - r = 0$—which immediately gives $r' = r$ and $q = q'$. ■

Theorem 4.2. *Let a and b be two integers, and let $d = gcd(a, b)$. Then $d = xa + yb$, for some $x, y \in \mathbf{Z}$. Furthermore, every other common divisor of both a and b, also divides d.*

Proof. From the way we have defined the gcd, the result is obvious when either a or b, or both, are 0. So let us assume that neither is 0.

Consider the set $S = \{ r, s \in \mathbb{Z} \mid ar + bs \geq 1 \}$. This set is not empty (e.g. make $r = a$ and $s = b$); thus by the well-ordering principle, it contains a smallest element. Let $d = xa + by$ be that element. Dividing a by d, we get $a = dq + r \iff a = (xa + by)q + r \iff r = a(1 - xq) - byq$. Thus the remainder is also a linear combination of a and b—which means that if $r > 0$, then $r \in S$. But r must be smaller than d, and d is, by assumption, supposed to be the smallest element of S—so r cannot belong to S. Hence we conclude that $r = 0$ (i.e. $d \mid a$). With b a similar reasoning shows that $d \mid b$—and thus d is a common divisor of both a and b.

Given that any number that divides a and b must divide any linear combination of theirs, we conclude that any common divisor of a and b must also divide d. In particular this also shows that d must be the *greatest* common divisor—indeed if d' were a common divisor that was greater than d, then we would have $d' \mid d$, which is a contradiction.[7] ■

Theorem 5.1. *$a \bmod n = b \bmod n$ if and only if $a \equiv b \pmod{n}$.*

Proof. ($\rightarrow$) $a \bmod n = b \bmod n \iff a - nq_1 = b - nq_2 \iff a - b = n(q_1 - q_2) \iff n \mid (a - b) \iff a \equiv b \pmod{n}$.

($\leftarrow$) $a \equiv b \pmod{n} \iff (a \bmod n + nq_1) - (b \bmod n + nq_2) = kn \iff a \bmod n - b \bmod n = n(k - q_1 + q_2)$. This shows $a \bmod n \equiv b \bmod n \pmod{n}$. But as both $a \bmod n$ and $b \bmod n$ belong to $\{0, \ldots, n-1\}$, their difference belongs to $\{-(n-1), \ldots, -1, 0, 1, \ldots, n-1\}$—and the only multiple of n in this range is 0. Hence $a \bmod n = b \bmod n$. ■

[6] What this means is that, if they are different, then one of them must be greater than the other. But *which one* of them plays that role does not matter: if I were to assume that $r \geq r'$, I could redo the same reasoning and obtain the same conclusion.

[7] In my view this already shows the gcd to be *unique*, for no set of integers can contain two distinct greatest elements. However, the gcd's uniqueness can also be shown explicitly: let d' now be another gcd. We would necessarily have $d \mid d'$ and $d' \mid d$, and as the gcd is always non-negative by definition, we conclude that $d = d'$.