Cambridge University Press
978-1-108-95823-3 — Algebra
Paolo Aluffi
Excerpt
More Information

Part I

Rings

1 The Integers

1.1 The Well-Ordering Principle and Induction

Our general goal is the generalization of common algebraic structures; we will try to capture their essence by determining axioms which are responsible for their properties. We start with the set of integers $\mathbb{Z}$, considered along with the two operations of addition $(+)$ and multiplication $(\cdot)$. We will spend some time trying to understand how $\mathbb{Z}$ is put together with respect to these two operations, and we will identify several key properties. We will then take a selection of those properties, the so-called *ring axioms,* and eventually aim at studying *all* structures that are defined by requiring a set R along with two operations (which will be called $+$ and $\cdot$ even if they may have nothing to do with the conventional $+$ and $\cdot$) to satisfy the ring axioms. These structures will be called *rings:* from this perspective, $\mathbb{Z}$ is a particular example of a ring. Other examples you are (hopefully) familiar with are $\mathbb{Q}$ ('rational' numbers), $\mathbb{R}$ ('real' numbers), $\mathbb{C}$ ('complex' numbers); but many more exist, and most of them have *nothing* to do with numbers. We will study 'all' of them at once in the sense that we will determine several features that every such structure (as opposed to specific examples like $\mathbb{Z}$ or $\mathbb{Q}$) must have. We will (implicitly) define a *category* of rings by specifying certain types of functions (which we will call 'homomorphisms') between rings.

In any case, we begin with $\mathbb{Z}$. We will start by recalling several known facts about $\mathbb{Z}$, from a perspective that will perhaps be a little more modern and rigorous than what may be seen in high-school math or Calculus. Everything we will see should sound familiar, but the viewpoint may seem unusual at first—the goal will be to single out the key facts that are responsible for 'the way $\mathbb{Z}$ works'. These facts will be useful in studying other examples, particularly coming from 'modular arithmetic', and the study of these examples will guide us in choosing the axioms that we will use in order to define what a ring is.

We do not really start with a blank slate: I will assume familiarity with the basic, elementary-school properties of addition and multiplication between integers. I will also assume familiarity with the notion of 'ordering' in $\mathbb{Z}$: if a and b are integers, we write $a \le b$ to mean that a is 'less than or equal to' b, in the ordinary sense. This ordering behaves in predictable ways with respect to the operations: for example, if $a \le b$ and $c \ge 0$, then $ac \le bc$; and similar statements you already know.

We can take these basic properties for granted, but it is helpful to introduce some terminology. We will begin by focusing on the fact that 'division' behaves rather peculiarly in $\mathbb{Z}$. For example, we can divide 18 by 3, getting a quotient of 6, which is an element of $\mathbb{Z}$. We can also divide 1 by 2, but the quotient is *not* an element of $\mathbb{Z}$: there is no integer c such that $1 = 2c$. We need some terminology to be able to deal with this distinction.

DEFINITION 1.1 Let $a, b \in \mathbb{Z}$. We say that 'b divides a', or 'b is a divisor of a', or 'a is a multiple of b', and write $b \mid a$, if there is an *integer* c such that $a = bc$.

Thus, 3 'divides' 18 since $18 = 3 \cdot 6$ and 6 is an integer; while 2 does *not* divide 1, because $1 = 2 \cdot \frac{1}{2}$ but $\frac{1}{2}$ is not an integer. The divisors of 12 are ±1, ±2, ±3, ±4, ±6, and ±12. Every integer divides 0, since $0 = b \cdot c$ for some integer c; as it happens, $c = 0$ works. On the other hand, the only integer that 0 divides is 0 itself.

With this understood, we can already record a useful (if completely elementary) fact.

LEMMA 1.2 *If $b \mid a$ and $a \neq 0$, then $|b| \leq |a|$.*

Proof Indeed, by definition of divisibility we have $a = bc$ for some integer c; in particular, both b and c are nonzero since $a \neq 0$. Since $c \in \mathbb{Z}$ and $c \neq 0$, we have $|c| \geq 1$. And then

$$|a| = |b| \cdot |c| \geq |b| \cdot 1 = |b|$$

as claimed. □

'Divisibility' defines an order relation on the set of nonnegative integers (Exercise 1.2). What Lemma 1.2 says is that, to some extent, this 'new' order relation is compatible with the ordinary one $\leq$.

What if b does *not* divide a? We can still divide b into a (at least if $b \neq 0$), but we have to pay a price: we get a *remainder* as we do so. Even if this fact is familiar to you, we will look into why it works the way it does, since it is something special about $\mathbb{Z}$—we have no need for this complication when we divide *rational* numbers, for example. So there must be some special property of $\mathbb{Z}$ which is responsible for this fact.

This property is in fact a property of the ordering $\leq$ we reviewed a moment ago. It will be responsible for several subtle features of $\mathbb{Z}$; in fact, besides the basic high-school properties of addition and multiplication, it is essentially the *one* property of $\mathbb{Z}$ that makes it work the way it does. So we focus on it for a moment, before returning to the issue of divisibility.

As recalled above, $\mathbb{Z}$ comes endowed with an order relation $\leq$. In fact, this is what we call a 'total ordering'. By this terminology we mean that given any two integers a and b, one of three things must be true: $a < b$, or $a = b$, or $a > b$. The same can be said of other sets of numbers, such as the set of *rational* numbers $\mathbb{Q}$ and the set of *real* numbers $\mathbb{R}$; but there is something about the ordering relation on $\mathbb{Z}$ that makes it very special. Terminology: $\mathbb{Z}^{\geq 0}$ stands for the set of *nonnegative* integers:

$$\mathbb{Z}^{\geq 0} = \{a \in \mathbb{Z} \mid a \geq 0\} = \{0, 1, 2, 3, \dots\}.$$

(Similarly $\mathbb{Z}^{>0}$ stands for the set of positive integers, $\mathbb{Q}^{\leq 0}$ is the set of nonpositive rational numbers, and so on.) Another name for $\mathbb{Z}^{\geq 0}$ is $\mathbb{N}$, the set of 'natural numbers'. The fact we need has everything to do with $\mathbb{Z}^{\geq 0}$.

> FACT (Well-ordering principle) *Every nonempty set of nonnegative integers contains a least element.*

We can summarize this fact by saying that $\mathbb{Z}^{\geq 0}$ is 'well-ordered' by the relation $\leq$. A 'well-ordering' on a set S is simply an order relation such that every nonempty subset of S has a minimum.

The well-ordering principle should sound reasonable if not outright obvious: if we have many bags of potatoes, there will be one (or more) bags with the least number of potatoes. But whether this is obvious or not is a matter of perspective—if we were to attempt to *define* $\mathbb{Z}$ rigorously, the well-ordering principle for nonnegative integers would be one of the axioms we would explicitly adopt; there is (to my knowledge) no direct way to derive it from more basic properties of the set of integers.

Also note that $\leq$ is *not* a well-ordering on the set $\mathbb{Q}^{\geq 0}$ of nonnegative rationals, or on the set $\mathbb{R}^{\geq 0}$ of nonnegative reals. For example, the set of positive rationals is a nonempty subset of $\mathbb{Q}^{\geq 0}$, but it does not have a 'least' element. (If $q > 0$ were such an element, then $\frac{q}{2}$ would be even smaller and still rational and positive, giving a contradiction.) So the well-ordering principle is really a rather special feature of $\mathbb{Z}^{\geq 0}$. We will derive several key properties of $\mathbb{Z}$ from it, granting (as we already did above) simple facts about how the ordering $\leq$ behaves with respect to the operations $+, \cdot$ on $\mathbb{Z}$.

Even before we see the well-ordering principle in action in 'algebra' proper, it is useful to observe that it already plays a role in a specific logical tool with which you are already familiar: the process of *induction* depends on it. Every proof by induction can be converted into an argument appealing to the well-ordering principle. (In fact, my preference is often to do so.) Why?

As you likely know, induction works as follows. We want to prove a certain property $P(n)$ for all integers $n \geq 0$. Suppose we manage to prove that

(i) $P(0)$ holds: the property is true for $n = 0$; and

(ii) the implication $P(n) \implies P(n+1)$ holds for all $n \geq 0$.

Then induction tells us that indeed, our property $P(n)$ holds for all $n \geq 0$. This smacks of magic, particularly the first few times you see it in action, but it is in a sense 'intuitively clear': the 'seed' $P(0)$ holds because you have proved it by hand in (i); and then $P(1)$ holds since $P(0)$ holds *and* you have proved in (ii) that $P(0) \implies P(1)$; and then $P(2)$ holds since $P(1)$ holds *and* you have proved that $P(1) \implies P(2)$; and then $P(3)$ holds since $P(2)$ holds *and* you have proved that $P(2) \implies P(3)$; and so on forever.

The problem with this argument is that 'so on forever' is not mathematics. There is an alternative argument which *is* rigorous, *once you grant the truth of the well-ordering principle.* Here it is.

Induction from the well-ordering principle Assume we have established (i) and (ii). We have to prove that $P(n)$ holds for all $n \geq 0$.

Let $F \subseteq \mathbb{Z}^{\geq 0}$ be the set of nonnegative integers n such that $P(n)$ does *not* hold; then we have to prove that $F = \emptyset$. We will prove that this is necessarily the case by showing that $F \neq \emptyset$ leads to a contradiction.

Assume then that our set F *is a nonempty set of nonnegative integers.* By the well-ordering principle, F has a least element $\ell \in \mathbb{Z}^{\geq 0}$: that is, $P(\ell)$ *does not hold,* and ℓ is the least nonnegative integer with this property.

By (i) we know that $P(0)$ holds, and therefore $\ell > 0$. Then $n = \ell - 1$ is a nonnegative integer, and $n < \ell$, therefore $P(n)$ holds, since ℓ is the *least* nonnegative integer for which the property does not hold. By (ii), $P(n) \implies P(n+1)$; so $P(n+1)$ holds. But $n + 1 = \ell$, so this shows that $P(\ell)$ *does* hold.

We have reached a contradiction: $P(\ell)$ would both hold and not hold. Therefore, the assumption that $F \neq \emptyset$ leads to a contradiction, and we must conclude that $F = \emptyset$, as we needed. □

Several proofs in what follows could be handled by induction or interchangeably by an appeal to the well-ordering principle. Which to use is essentially a matter of taste. I will often insist on using the well-ordering principle, to stress that we are really using a specific feature of $\mathbb{Z}$. Also, I often find it somewhat easier to write a lean, rigorous argument by directly invoking the well-ordering principle rather than induction.

1.2 'Division with Remainder' in $\mathbb{Z}$

The first substantial application of the well-ordering principle is the fact that in $\mathbb{Z}$ we can perform a 'division with remainder', as mentioned above. For example, $13 = 4 \cdot 3 + 1$: 13 divided by 4 gives 3 with a remainder of 1. The official statement is the following.

> THEOREM 1.3 (Division with remainder) *Let a, b be integers, with $b \neq 0$. Then there exist a unique 'quotient' $q \in \mathbb{Z}$ and a unique 'remainder' $r \in \mathbb{Z}$ such that*
>
> $$a = bq + r \qquad \text{with } 0 \leq r < |b|.$$

Remark 1.4 The 'uniqueness' part is important. It is clear that we can write a division-with-remainder in many ways: $13 = 4 \cdot 4 - 3 = 4 \cdot 3 + 1 = 4 \cdot 2 + 5 = \cdots$; but the theorem claims that *one and only one* of these ways will satisfy the condition that the remainder is ≥ 0 and < 4. ⌟

Proof of the theorem We can assume that $b > 0$. Indeed, if $b < 0$, we can apply the statement to $-b > 0$ and then flip the sign of q after the fact. Switching the sign of b does not change $|b|$, so the condition on r is unchanged.

Assume then that $b > 0$, and consider all integer linear combinations of a and b of the form $a - bx$ with $x \in \mathbb{Z}$. Note that there are nonnegative integers of this type: if a is itself nonnegative, then $a - b \cdot 0 = a$ is nonnegative; and if a is negative, then $a - ba = a(1 - b)$ is nonnegative because $1 - b \leq 0$. Therefore, the set

$$S = \{a - bx \text{ with } x \in \mathbb{Z}, \text{ such that } a - bx \geq 0\}$$

is a nonempty set of nonnegative integers. By the well-ordering principle, it contains a least element r: that is, there is some $x = q$ such that $r = a - bq \geq 0$ is smaller than any other nonnegative number of the form $a - bx$. I claim that these q and r are the unique numbers whose existence is claimed in the statement.

Indeed: by construction we have that $a = bq + r$ and that $r \geq 0$; we have to verify that (1) $r < b$ (note that $|b| = b$ since we are assuming $b > 0$) and (2) the numbers q and r are unique with these properties.

For (1), argue by contradiction. If we had $r \geq b$, then we would have $r - b \geq 0$; and

$$0 \leq r - b = (a - bq) - b = a - b(q + 1)\,.$$

This would show that $r - b$ is an element of S, since it is a nonnegative integer linear combination of a and b. But that is a contradiction, since $r - b < r$ while r was chosen to be the least element of S. Since $r \geq b$ leads to a contradiction, we can conclude that $r < b$ as we claimed.

For (2), assume that q', r' are also integers satisfying the requirement spelled out in the theorem: that is, $a = bq' + r'$, and $0 \leq r' < b$. We will show that necessarily $q' = q$ and $r' = r$ (this is what we mean by stating that q and r are 'unique').

Since $a = bq + r = bq' + r'$, we have

$$b\,(q - q') = r' - r\,. \tag{1.1}$$

Now, since both r and r' are in the interval $[0, b - 1]$, their difference cannot exceed $b - 1$. In other words, $|r' - r| \leq b - 1$, and in particular $|r' - r| < |b|$. But (1.1) shows that b is a divisor of $r' - r$; *if* $r' - r \neq 0$, then by Lemma 1.2 we would necessarily have $|b| \leq |r' - r|$, contradicting the fact that $|b| > |r' - r|$ we observed a moment ago. The only possibility then is that $r' - r = 0$. This shows that $r' = r$, and further $b\,(q - q') = 0$. Since $b \neq 0$, this implies that $q - q' = 0$, and it follows that $q' = q$. We are done. □

1.3 Greatest Common Divisors

The next item in our agenda is an important notion, with which again you are likely to be familiar, but possibly in a slightly different form.

DEFINITION 1.5 Let $a, b \in \mathbb{Z}$. We say that a nonnegative integer d is the 'greatest common divisor' of a and b, denoted $\gcd(a, b)$ or simply (a, b), if

- $d \mid a$ and $d \mid b$; and
- if $c \mid a$ and $c \mid b$, then $c \mid d$.

If at least one of a and b is not 0, then $d = \gcd(a, b) \neq 0$. Indeed, d must divide both a and b according to the first requirement in the definition (this makes d a 'common divisor'), so d cannot be 0 unless both a and b are 0—keep in mind that 0 only divides 0. The second requirement in the definition says that if c is also a common divisor, then it must divide d. By Lemma 1.2, this implies that $|c| \le |d|$ if $d \neq 0$, and hence $c \le d$ since d is nonnegative. Therefore d is the 'greatest' (as in 'largest') common divisor of a and b *if* a and b are not both 0.

It is clear that the gcd *exists:* if you have two integers a and b, not both equal to 0, then you can simply list the divisors of a, lists the divisors of b, and then $\gcd(a, b)$ is simply the largest integer in the intersection of these two sets.

Example 1.6 To find that $\gcd(30, -42) = 6$, we can list all divisors of 30:

$$-30, -15, -10, -6, -5, -3, -2, -1, 1, 2, 3, 5, \underline{\mathbf{6}}, 10, 15, 30$$

and all divisors of -42:

$$-42, -21, -14, -7, -6, -3, -2, -1, 1, 2, 3, \underline{\mathbf{6}}, 7, 14, 21, 42$$

and then note that 6 is the largest integer that occurs in both lists.

Beware, however, that this method to find a gcd is extremely inefficient: finding the divisors of an integer is a time-consuming business, and if the integer is largish (say, a few thousand digits) then I suspect that its factorization may take current algorithms longer than the age of the universe to be carried out on the fastest imaginable digital computers. Can you think of a *faster* way to find the gcd of two integers? We will soon encounter a very efficient alternative to the 'inspection' method given above (Theorem 1.14).

Remark 1.7 You may wonder why I have not replaced the second requirement with something like

- if $c \mid a$ and $c \mid b$, then $c \le d$,

which would seem to justify the terminology 'greatest common divisor' in a more direct way. (This may be the definition you have run into in previous encounters with this notion.) The main reason is that when we recast this notion for more general 'rings', the relation of divisibility will be available, while the (perhaps) simpler relation $\le$ will *not.* Also, Definition 1.5 makes good sense for all possible integers a, b, including the case $a = b = 0$: in this case, since *every* number divides 0, the second requirement says that *every integer should divide* $\gcd(0, 0)$. Since 0 is the only number divisible by every integer, this tells us that $\gcd(0, 0) = 0$ according to Definition 1.5. Even in this case, $\gcd(a, b)$ is 'greatest'; but with respect to the divisibility relation from Exercise 1.2 rather than the ordinary order relation $\le$. ⌟

The following fact is another easy consequence of the well-ordering principle.

> THEOREM 1.8 *Let a, b be integers. Then the greatest common divisor $d = \gcd(a, b)$ is an integer linear combination of a and b. That is, there exist integers m and n such that $d = ma + nb$.*

> *In fact, if a and b are not both* 0*, then* $\gcd(a,b)$ *is the smallest positive linear combination of a and b.*

For example, I pointed out above that $(30, -42) = 6$. The theorem then tells me that there must be integers m and n such that $30m - 42n = 6$. Does this seem 'obvious' to you? It does not look obvious to me, but a couple of attempts quickly yield $m = 3$, $n = 2$, which give $3 \cdot 30 - 2 \cdot 42 = 90 - 84 = 6$. (By the end of the section we will see a 'systematic' way to find such integers, cf. Remark 1.16.)

Notice that it is clear that the integers m and n are *not* going to be unique, since if $d = ma + nb$, then also $d = (m - b)a + (n + a)b$, $d = (m + b)a + (n - a)b$, and so on. For instance,

$$(3 + 42) \cdot 30 - (2 + 30) \cdot 42 = 45 \cdot 30 - 32 \cdot 42 = 1350 - 1344 = 6\,,$$

therefore $m = 45$, $n = 32$ also work in our example. Doing this type of experimentation is excellent practice, but it is kind of clear that mindless arithmetic will not prove the theorem in general. The missing ingredient is the well-ordering principle, and the beginning of the proof will remind you of the beginning of the proof of Theorem 1.3.

Proof of Theorem 1.8 If $a = b = 0$, then $\gcd(a,b) = \gcd(0,0) = 0$, which is a linear combination of a and b (because $0 = 1 \cdot 0 + 1 \cdot 0$). Therefore, we may assume that a and b are not both 0. Consider all linear combinations $ma + nb$ of a and b. I claim that some of them are positive. Indeed, take $m = a$, $n = b$; then $ma + nb = a^2 + b^2$, and this number is positive since a and b are not both 0.

Therefore, the set

$$S = \{ma + nb \mid m \in \mathbb{Z}, n \in \mathbb{Z}, \text{and } ma + nb > 0\}$$

is nonempty. This is the standard setting to apply the well-ordering principle: since S is a nonempty set of nonnegative integers, it must have a least element. Let d be this element, and let m and n be integers such that $d = ma + nb$. That is, d is the smallest positive linear combination of a and b. We are going to verify that d is the gcd of a and b, and this will prove the theorem.

Since $d \in S$, then $d > 0$: so d is nonnegative and not equal to 0. In order to prove that d is the gcd of a and b, we have to prove that

(i) $d \mid a$ and $d \mid b$; and
(ii) if $c \mid a$ and $c \mid b$, then $c \mid d$.

If $a = 0$, $d \mid a$ is automatic. If $a \neq 0$, we can use division with remainder: by Theorem 1.3, we know that there exist integers q and r such that $a = dq + r$, with $0 \le r < d$. What can we say about r? Note that

$$r = a - dq = a - (ma + nb)\,q = a\,(1 - m) + b\,(-nq) :$$

this shows that r is a linear combination of a and b. Can it be an element of S? No! Because $r < d$, and d has been chosen to be the smallest element of S. But then r cannot be positive, since r is a linear combination of a and b, and S contains all *positive* linear combinations of a and b. Since $r \geq 0$ and r is not positive, it follows that $r = 0$. This proves that $a = dq$, showing that $d \mid a$.

By essentially the same argument, we can deduce that $d \mid b$. (The roles of a and b are intechangeable.) This takes care of (i).

We still have to prove (ii). Suppose we have a common divisor c of a and b: $c \mid a$ and $c \mid b$. Then we have $a = uc$, $b = vc$ for some integers u and v. This gives

$$d = ma + nb = m\,(uc) + n\,(vc) = (mu + nv)\,c$$

and proves that $c \mid d$, as we needed. □

Theorem 1.8 has nice, if somewhat technical, applications. Here is one.

COROLLARY 1.9 *Let a, b be integers. Then* $\gcd(a, b) = 1$ *if and only if* 1 *may be expressed as a linear combination of a and b.*

Proof If $\gcd(a, b) = 1$, then the number 1 may be expressed as a linear combination of a and b by Theorem 1.8. On the other hand, if 1 may be expressed as a linear combination of a and b, then 1 is necessarily the *smallest* positive linear combination of a and b, because 1 is the smallest positive integer. It follows that $\gcd(a, b) = 1$, again as a consequence of Theorem 1.8. □

DEFINITION 1.10 We say that a and b are *relatively prime* if $\gcd(a, b) = 1$.

For example, 3 and 7 are relatively prime. It follows that *every* integer is a linear combination of 3 and 7: indeed, 1 is a linear combination, and a multiple of a linear combination is a linear combination. For example, if you give me a random integer, say 173238384731, then I can do a small computation and tell you that

$$173238384731 = 866191923655 \cdot 3 + (-346476769462) \cdot 7\,.$$

(Exercise: How did I do that?)

And here is another application of Theorem 1.8. This will become handy in an important proof later on.

COROLLARY 1.11 *Let a, b, c be integers. If $a \mid bc$ and* $\gcd(a, b) = 1$*, then $a \mid c$.*

Proof By Theorem 1.8, $1 = ma + nb$ for some integers m and n. Multiply by c:

$$c = 1 \cdot c = (ma + nb)\,c = mac + nbc = (mc)\,a + n\,(bc)\,.$$

Now the hypothesis tells us that $a \mid bc$, so $bc = ra$ for some integer r. Then

$$c = (mc)\,a + n\,(ra) = (mc + nr)\,a\,,$$

and this shows that $a \mid c$ as needed. □