# Introduction to Abstract Algebra (In Progress)

Nihal Uppugunduri

May 21, 2025

# Contents

# Preface

## 1. Introduction

This text introduces a mathematical subject called *abstract algebra*.

Before we begin, I would like to thank Alex Tsun for his suggestions and feedback on parts of this material. Also, the inclusion of the first chapter is heavily influenced by Charles C. Pinter's *A Book of Abstract Algebra*, which is how I first learned this subject.

## 2. Prerequisites and Notation

In general, we assume a prerequisite of precalculus at the level of Art of Problem Solving's *Precalculus*. In particular, we assume familiarity with algebra (polynomials, functions, etc.), linear algebra (vectors, matrices, etc.), and complex numbers. We always assume comfortability with proofs and mathematical rigor. We use axiomatic set theory implicitly; if you do not know about this, then it doesn't matter, but if you do know about this, and if you care about equivalence to or dependence on the Axiom of Choice, then we will make skippable notes of that for the relevant theorems. We include an appendix where we cover the other relevant prerequisite background.

We assume familiarity with standard and common mathematical notation. For example:

1. $\mathbb{R}$ denotes the set of real numbers, $\mathbb{Z}$ the set of integers, and $\mathbb{Q}$ the set of rational numbers.

2. For either of these three sets, attaching the subscript + produces the subset of positive members, so for example $\mathbb{Z}_+$ is the set of positive integers.

3. Similarly, for either of these three sets, attaching the subscript 0 produces the subset of nonnegative members, so for example $\mathbb{R}_0$ is the set of nonnegative real numbers.

4. For any two sets $A$ and $B$, $A \times B$ denotes the set of ordered pairs where the first "coordinate" is from $A$ and the second is from $B$: $A \times B = \{(a, b) : a \in A, b \in B\}$. For example, if $A = \{1, 2\}$ and $B = \{2, 3\}$, then $A \times B = \{(1, 2), (1, 3), (2, 2), (2, 3)\}$.

5. In particular, for any set $S$, $S^2 = S \times S$ denotes the set of ordered pairs of elements of $S$: $S^2 = \{(a, b) : a, b \in S\}$. For example, if $S = \mathbb{R}$, then $S^2 = \mathbb{R}^2$ is the set of ordered pairs of real numbers (which geometrically is the set of two-dimensional points, aka two-dimensional space.)

6. For any three sets $A$, $B$, and $C$, $A \times B \times C$ denotes the set of ordered triples where the first "coordinate" is from $A$, the second from $B$, and the third from $C$: $A \times B \times C = \{(a, b, c) : a \in A, b \in B, c \in C\}$.

7. In particular, for any set $S$, $S^3 = S \times S \times S$ denotes the set of ordered triples of elements of $S$: $S^3 = \{(a, b, c) : a, b, c \in S\}$. For example, if $S = \mathbb{R}$, then $S^3 = \mathbb{R}^3$ is the set of ordered triples of real numbers (which geometrically is the set of three-dimensional points, aka three-dimensional space.)

8. For any $n$ sets $A_1, ..., A_n$, $A_1 \times ... \times A_n$ denotes the set of all ordered $n$-tuples where the $i$th "coordinate" is from $A_i$: $A_1 \times ... \times A_n = \{(a_1, ..., a_n) : a_i \in A_i\}$.

9. In particular, for any set $S$, $S^n$, or $S$ $\times$'d with itself $n$ times, is the set of ordered $n$-tuples of elements of $S$: $S^n = \{(a_1, ..., a_n) : a_i \in S\}$.

10. For $x \in A_1 \times ... \times A_n$, $x_i$ denotes the $i$th "coordinate" of $x$, which would be an element of $A_i$. For example, if $x = (4, 5, 6) \in \mathbb{R}^3$, then $x_1 = 4$, $x_2 = 5$, and $x_3 = 6$.

## 3. Differences From Other Treatments

The main new concepts in abstract algebra are what are called groups, rings, fields, and vector spaces. Many (if not most) abstract algebra textbooks follow this order, or at least start with something other than fields (for example, some start with rings.) However, here we start with fields. We think that the order should not affect the learning experience too strongly, and fields may be a more "familiar" concept to start with because of experience with rational, real, and complex numbers.

## 4. Practice Problems

If you want more practice problems in abstract algebra, we recommend that, in addition to standard "plug-and-chug" exercises that may involve just applying formulas in obvious ways, you also look at "brainteaser" problems that involve more creative combinations of theorems. The Putnam Mathematical Competition is a famous college-level extracurricular contest that features these kinds of "brainteaser" problems for typical undergraduate math subjects, including abstract algebra.

## 5. Feedback

I would appreciate any comments, feedback, or corrections. The best place for this is the comment section on my blog post announcing this text (`https://riemannzeta5.com/2024/09/13/starting-intro-to-abstract-algebra/`).

Now, without further ado, let's begin!

# Part I.

# Introduction

# 1. Primer: What is Abstract Algebra?

For this chapter, we only assume familiarity with real numbers, variables, and equations.

Let us look back briefly at our previous studies in algebra, to motivate and provide perspective for what is to come.

When we first learned math, we saw how to compute expressions with numbers, such as $2 + 3 = 5$. Then, when we learned algebra, we shifted to representing numbers with letters. This led to the concept of solving an equation for an unknown variable. For example, we would solve $2 + x = 5$ to get $x = 3$. We would then devote our attention to solving all types of equations – linear, quadratic, polynomial, etc. All the quantities involved in these equations were generally real numbers.

But we can in fact do "algebra" with other types of objects. For example, consider the set of all colors, and say we have an operation on colors, denoted by $+$, which is the result of mixing two colors. So yellow$+$blue = green, white$+$black = grey, and so on. We can now do "algebra with colors!" We can write down general properties – for instance, since mixing a color with itself produces just that color, we have, for any color $x$,

$$x + x = x.$$

Also, it does not matter which order we mix the colors (we define the color mixing operation so that proportions are always the same):

$$a + b = b + a.$$

We can write down (and solve) equations:

$$x + \text{blue} = \text{green} \to x = \text{yellow}.$$

If we compare the general properties satisfied by real numbers and colors, we see that there are some similarities (like $a + b = b + a$) but also some differences (like $x + x = x$.)

When abstract algebra was founded, mathematicians had been studying many other "algebras" like these – in fact, it is estimated that today over 200 algebras have been studied. Every time they wanted to formulate a new algebra, they would model some operations on objects and use those operations to build the main concepts of algebra: expressions, equations, and so forth. Each of these systems was both similar to and different from the others. As more exotic types of objects and operations were studied, the subject of algebra shifted from focusing on the real numbers to embracing the full landscape of all these algebras. Soon after, mathematicians started asking: instead of working in each of these algebras in isolation, could we prove theorems that could apply to multiple algebras at once?

Enter abstract algebra. We can see that many properties of an algebra can be derived from a few "fundamental" properties. For instance, consider the real numbers, which satisfy properties like:

1. $a + b = b + a$ (The *Commutativity of Addition*)
2. $ab = ba$ (The *Commutativity of Multiplication*)
3. $(a + b) + c = a + (b + c)$ (The *Associativity of Addition*)
4. $(ab)c = a(bc)$ (The *Associativity of Multiplication*)
5. The existence of a *neutral element* or *identity* 0 for addition, where $0 + r = r + 0 = r$
6. The existence of a similar identity 1 for multiplication, where $1r = r1 = r$
7. The existence, for each element $r$, of its *negative* or *additive inverse* $-r$, where $r + (-r) = (-r) + r = 0$

8. The existence, for each nonzero element $r$, of its *reciprocal* or *multiplicative inverse* $r^{-1}$, where $rr^{-1} = r^{-1}r = 1$ (The "nonzero" is important!)
9. $a(b + c) = ab + ac$ (The *Distributivity of Multiplication Over Addition*)
10. $0r = r0 = 0$

Now, say we were solving the linear equation $ax+b = 0$. Recall that the solution is $x = -b/a = -(ba^{-1})$ if $a \neq 0$. How did we arrive at that solution? Let us lay it out in terms of very small, basic steps:

| | | | |
|---|---|---|---|
| $ax + b$ | $=$ | $0$ | Start |
| $(ax + b) + (-b)$ | $=$ | $0 + (-b)$ | Add $-b$ to Both Sides |
| $(ax + b) + (-b)$ | $=$ | $-b$ | $0 + r = r + 0 = r$ for all $r$ |
| $ax + (b + (-b))$ | $=$ | $-b$ | Associativity of Addition |
| $ax + 0$ | $=$ | $-b$ | $r + (-r) = (-r) + r = 0$ for all $r$ |
| $ax$ | $=$ | $-b$ | $0 + r = r + 0 = r$ for all $r$ |
| $a^{-1}(ax)$ | $=$ | $a^{-1}(-b)$ | Multiply Both Sides by $a^{-1}$, $a \neq 0$ |
| $(a^{-1}a)x$ | $=$ | $a^{-1}(-b)$ | Associativity of Multiplication |
| $1x$ | $=$ | $a^{-1}(-b)$ | $rr^{-1} = r^{-1}r = 1$ for all $r$ |
| $x$ | $=$ | $a^{-1}(-b)$ | $1r = r1 = r$ for all $r$ |
| $x$ | $=$ | $(-b)a^{-1}$ | Commutativity of Multiplication |
| $x + ba^{-1}$ | $=$ | $(-b)a^{-1} + ba^{-1}$ | Add $ba^{-1}$ to Both Sides |
| $x + ba^{-1}$ | $=$ | $((-b) + b)a^{-1}$ | Distributivity |
| $x + ba^{-1}$ | $=$ | $0a^{-1}$ | $r + (-r) = (-r) + r = 0$ for all $r$ |
| $x + ba^{-1}$ | $=$ | $0$ | $0r = r0 = 0$ for all $r$ |
| $(x + ba^{-1}) + (-ba^{-1})$ | $=$ | $0 + (-ba^{-1})$ | Add $-ba^{-1}$ to Both Sides |
| $x + (ba^{-1} + (-ba^{-1}))$ | $=$ | $0 + (-ba^{-1})$ | Associativity of Addition |
| $x + 0$ | $=$ | $0 + (-ba^{-1})$ | $r + (-r) = (-r) + r = 0$ for all $r$ |
| $x$ | $=$ | $-(ba^{-1})$ | $0 + r = r + 0 = r$ for all $r$ |

Now, the fundamental idea that enables abstract algebra is this: we do not actually need $a, x$, or $b$ to be real numbers for this process to work. All we need

are those properties to be satisfied!

In other words, if we assume that we generally had some set of mathematical objects – they might be real numbers, or they might not be – together with some operations on those objects, that satisfied the properties above, then we can solve linear equations $ax + b = 0$ for those objects, the same way we solve linear equations for real numbers. In fact, anything we do in this general setting can immediately carry over to any algebra that shares these properties with the real numbers. We *abstract* the fact that the objects and operations satisfy the properties away from what the objects or operations actually are, hence the term, *abstract algebra*.

To give an example of this process, consider the generalization of the algebra of real numbers to this concept:

**Definition 1.0.1.** A *field* is a set $F$ together with operations $+, \times$ on the elements of $F$. These operations must be *binary*: they both must take in two elements and output one element. They must also satisfy the following properties:

1. *Commutativity*: For all $a, b$ in $F$, $a + b = b + a$ and $a \times b = b \times a$
2. *Associativity*: For all $a, b, c$ in $F$, $(a + b) + c = a + (b + c)$ and $(a \times b) \times c = a \times (b \times c)$
3. *Identity*: There exist *distinct* elements of $F$ called 0 and 1, such that, for all $a$ in $F$, $0 + a = a + 0 = a$ and $1 \times a = a \times 1 = a$
4. *Inverse*: For all $a$ in $F$, there exists an element $-a$ such that $a + (-a) = (-a) + a = 0$, and if $a$ is *nonzero*, there exists an element $a^{-1}$ such that $a \times a^{-1} = a^{-1} \times a = 1$
5. *Distributivity*: For all $a, b, c$ in $F$, $(a + b) \times c = (a \times c) + (b \times c)$ and $a \times (b + c) = (a \times b) + (a \times c)$

From these defining properties of a field, we can actually prove $0x = x0 = 0$ as a consequence. Thus, linear equation solutions for general fields are the same as linear equation solutions for real numbers. We can go on to ask how much

real number algebra can carry over – and indeed we can discuss polynomials over fields and theorems for fields that heavily generalize theorems about real numbers.

We can then identify many sets of objects which these theorems can apply to. For example, consider the set of just 0 and 1 with the following "addition" and "multiplication" operations:

$$\begin{aligned} 0+0 &= 0, \\ 0+1 &= 1, \\ 1+0 &= 1, \\ 1+1 &= 0, \\ 0\times 0 &= 0, \\ 0\times 1 &= 0, \\ 1\times 0 &= 0, \\ 1\times 1 &= 1. \end{aligned}$$

This is actually a field! (It is called $\mathbb{Z}_2$.) In fact, this is a special case of a more general kind of construction. Consider say the set of numbers from 0 to 9, with the operations of usual addition and multiplication but "wrapped around" to stay in 0 through 9 (so for example $9+9=18 \to 18-10=8$.) In other words, the operations are the remainders when divided by 10 of usual addition and multiplication. We can do this for any number $n > 1$: we define $\mathbb{Z}_n$ to be the set of numbers from 0 to $n-1$ (including both) with the operations as remainders when divided by $n$ of usual addition and multiplication. Then, it turns out that if $n$ is prime, $\mathbb{Z}_n$ is a field! We will prove this in general later on. In fact, these "finite fields" turn out to have lots of interesting applications, including being foundational to modern computer security.

We can alternatively study a different set of defining properties. For example, another concept in abstract algebra is a *group*:

**Definition 1.0.2.** A *group* is a set $G$ together with a binary operation $\times$ satisfying:

1. *Associativity*: For all $a, b, c$ in $G$, $(a \times b) \times c = a \times (b \times c)$
2. *Identity*: There exists an element of $G$, which we usually call $e$, such that, for all $g$ in $G$, $g \times e = e \times g = g$
3. *Inverse*: For all $g$ in $G$, there exists an element $g^{-1}$ such that $g \times g^{-1} = g^{-1} \times g = e$

This is just a subset of the defining properties of fields. This automatically means that all fields are groups[1], but so are many other algebras that aren't fields – and all of group theory applies to all of these cases. As an example, it turns out that the set of transformations you can apply to a Rubik's cube is a group – in fact, group theory is an important tool for discovering algorithms to solve Rubik's cubes quickly.

These abstracted concepts – fields and groups – are called *algebraic structures*. An algebraic structure is essentially any defining list of properties that a set of operations are required to satisfy.

The strength of abstract algebra is many-fold. First, theorems on algebraic structures extend results on familiar algebras like the real numbers to other sets of objects. But in the other direction, studying abstract algebra has actually resulted in new theorems even for familiar algebras like the real numbers! We will see more examples of this phenomenon later on.

[1] Did we mean "all groups are fields"? It may seem wonky, but think about it for a moment and you'll see that it's actually the other way around. A smaller set of defining properties means fewer requirements, which means more possibilities, so groups are more general than fields, not less. Groups include fields, so all fields are groups, but groups include other algebras too.

# Part II.

# The Main Algebraic Structures

# 2. Fields

## 2.1. Definition and Basic Properties

The first algebraic structure we study is the field. Fields abstract the properties of real numbers, as we just saw in the previous chapter:

**Definition 2.1.1.** A *field* is a set $F$ together with operations $+, \times$ on the elements of $F$. These operations must be *binary*: they both must take in two elements and output one element. They must also satisfy the following properties:

1. *Commutativity*: For all $a, b$ in $F$, $a + b = b + a$ and $a \times b = b \times a$
2. *Associativity*: For all $a, b, c$ in $F$, $(a+b)+c = a+(b+c)$ and $(a \times b) \times c = a \times (b \times c)$
3. *Identity*: There exist *distinct* elements of $F$ called 0 and 1, such that, for all $a$ in $F$, $0 + a = a + 0 = a$ and $1 \times a = a \times 1 = a$
4. *Inverse*: For all $a$ in $F$, there exists an element $-a$ such that $a + (-a) = (-a) + a = 0$, and if $a$ is *nonzero*, there exists an element $a^{-1}$ such that $a \times a^{-1} = a^{-1} \times a = 1$
5. *Distributivity*: For all $a, b, c$ in $F$, $(a + b) \times c = (a \times c) + (b \times c)$ and $a \times (b + c) = (a \times b) + (a \times c)$

For $a, b \in F$, we write $ab = a \times b$, just like for numbers. We also adopt the standard order of operations conventions when writing expressions, so that for example $a + bc = a + (b \times c)$.

Note: in abstract algebra we typically use the same symbol to refer to both a set and a set with some operations we have defined on it. The usage is dependent on context.

To remember the terminology of commutativity and associativity, we can visualize them as follows: commutativity looks like "$a$ and $b$ commuting to each others' positions," while associativity looks like "$b$ associating first with $a$ and then with $c$."

Before we dive further into the theory of fields, we mention some examples. Our "prototypal" one, which we based the very definition of a field off of, is $\mathbb{R}$, the set of real numbers, with the usual addition and multiplication. But there are others too: as discussed previously, $\mathbb{Z}_2$ is a field. (As a reminder, we define $\mathbb{Z}_n$, for any positive integer $n$, to be $\{0, 1, ..., n-1\}$ with addition and multiplication modulo $n$.) Additionally, $\mathbb{Z}_3$ is also a field. (If you're unsure, check all the properties explicitly to see that these are in fact fields.)

Now, often times, we'd like to consider properties that hold for specific algebras and ask whether they generalize to the abstract setting – that is a large part of the motivation of abstract algebra. But we also want to balance this aspiration with some skepticism: not all properties will carry over to the general case. Here, to gain some intuition for what works, we can take our inspiration from $\mathbb{R}$, which has very rich algebraic properties, but if we are ever in doubt, we can verify whether the property in question is true for something like $\mathbb{Z}_2$ or $\mathbb{Z}_3$. These "look very different" from $\mathbb{R}$ but they are still fields. If the property fails for these cases, then it can't carry over to general fields.

So let's start now by looking at the properties of $\mathbb{R}$. As mentioned earlier:

**Theorem 2.1.2.** We have $x0 = 0x = 0$.

Think for a moment about how we may prove this from the defining field properties. The only thing we know about 0 is that it is an additive identity, and our desired result involves multiplication. There's actually only one field property that relates addition and multiplication to each other, and that is distributivity. (Notice that all the other properties talk about either addition or multiplication,

not both.) Thus, we expect that we will need to involve $a + 0 = 0 + a = a$ and distributivity at some point.

Let's first just try to show that $x0 = 0$. We have $0 + a = a$; since we want $x0$ to appear at some point, it seems like a good idea to multiply both sides on the left by $x$. Distributivity further helps our hunch, since $x(0 + a) = x0 + xa$, so we've made our desired term appear. This yields $x0 + xa = xa$. Now, we'd like to cancel $xa$ from both sides like we do with numbers; then, we'd have our desired result. But remember, these are field elements, not numbers. Is cancellation still possible? Yes, it is: that's exactly what the additive inverse allows us to do. Putting this together, we have our proof:

*Proof.* We have $0 + a = a$. Multiplying both sides by $x$, $x(0 + a) = xa$. By distributivity, $x(0 + a) = x0 + xa$, so $x0 + xa = xa$. Now, $xa$ has an additive inverse $-(xa)$, so $(x0 + xa) + (-(xa)) = xa + (-(xa))$. The RHS is 0, and we can use associativity of addition on the LHS: $(x0 + xa) + (-(xa)) = x0 + (xa + (-(xa)))$. But this is just $x0 + 0 = x0$. This yields $x0 = 0$, as desired.

See if you can come up with the proof that $0x = 0$. □

An important point that results from this is the fact that, in any field, 0 can't have a multiplicative inverse. How can we prove this? Well, say it did have an inverse, call it $0^{-1}$, then we'd have $00^{-1} = 1$, but also what we've just shown implies that $00^{-1} = 0$, and 0 and 1 are required to be different, which is a contradiction. (If we had a set that satisfied all the field properties except that it had $0 = 1$, then it would necessarily be "trivial": for any element $x$ of that set, we'd have $1 = 0 \to x1 = x0 \to x = 0$, so the set could only contain one element, namely 0, or equivalently 1.) This is why the *nonzero* part in the multiplicative inverse property is necessary, and consequently why there is some asymmetry in the additive and multiplicative inverse properties.

However, just as we were able to cancel added terms by using additive inverses, we can cancel *nonzero* multiplied terms by using multiplicative inverses *of nonzero elements*. (Write this out and prove it if you're unsure.) So while

we can't cancel out 0 in multiplication, we can cancel out anything else. This is familiar from real-number algebra, and it holds true for general fields too.

Now, another point that we haven't yet elaborated on is that here, 0 and 1 are identities (for addition and multiplication respectively), but *not necessarily the only identities*. If we don't have uniqueness, then we can't really speak of "the" additive or multiplicative identity. Will this be an issue when trying to generalize more of real number algebra to fields? It turns out that we don't need to worry:

**Theorem 2.1.3.** The additive and multiplicative identities are unique.

Again, think for a moment about how you would prove this. Consider addition first. Another way of saying this is that if we had two additive identities, say $0_1$ and $0_2$, then both of them would have to be equal. How do we show this? Well, we want to involve $0_1$, $0_2$, and addition at some point. It seems like a good idea to consider $0_1 + 0_2$:

*Proof.* If we had two additive identities $0_1$ and $0_2$, then consider $0_1 + 0_2$. On the one hand, $0_1 + 0_2 = 0_1$ since $0_2$ is an identity; on the other hand, $0_1 + 0_2 = 0_2$ since $0_1$ is an identity. Thus, $0_1 = 0_2$, and the desired result follows.

See if you can come up with the proof that multiplicative identities are unique. □

Notice that we didn't actually use any other field properties here: we just used the fact that $0_1$ and $0_2$ are identities. This shows us that even beyond fields, whenever we have identities in abstract algebra, they will be unique. This will apply later on when we discuss for example group and ring theory.

Let's go back to $\mathbb{R}$. A common way to think of subtraction is as the "inverse" of addition, and similarly of division as the "inverse" of multiplication. Indeed, we can say for numbers that $a - b = a + (-b)$ and $a/b = ab^{-1}$. But we have inverses for any field! Does that mean we can extend subtraction and division to general fields as well?

The main issue in doing that stems again from uniqueness: on a field, we're given that $-b$ is *an* additive inverse for $b$, not necessarily *the*. So if we wanted to say subtract $b$ from $a$, via $a-b=a+(-b)$, which value of $-b$ would we pick? Furthermore, different choices of $-b$ would necessarily yield different values of $a+(-b)$, because of cancellation. (Write this out and explicitly prove it if you're unsure.) However, it turns out that we don't need to worry here either:

**Theorem 2.1.4.** Additive and multiplicative inverses are unique. (For $x \in F$, there is exactly one element $-x$ that adds with $x$ to 0, and similarly, if $x \neq 0$, exactly one element $x^{-1}$ that multiplies with $x$ to 1.)

As before, think about how we could prove this. Start with addition. A safe bet would be to replicate part of our previous method for uniqueness of identities, by assuming there were two inverses and showing they must be equal. So take $x \in F$, and say that $y$ and $z$ are both additive inverses of $x$. Then, we'd have $x+y=0$ and $x+z=0$. Now, try to complete the proof from here. And similarly, try to establish uniqueness of multiplicative inverses.

Uniqueness of inverses now allows subtraction $a-b=a+(-b)$ and division $a/b=ab^{-1}$, which we will henceforth use throughout regardless of whether we are on $\mathbb{R}$ or a general field.

Let's look at some other properties that can carry over from $\mathbb{R}$; in particular, focus on consequences of the associative property. Take addition first. We know that addition is only required to be a binary operation (takes in two elements), so if we have an expression like $a+b+c$, what does that mean? What order do we evaluate things in? Is it $(a+b)+c$ or $a+(b+c)$? Well, associativity implies in this case that it doesn't matter, since these have to be the same. What about $a+b+c+d$ – is it $((a+b)+c)+d$, or $(a+b)+(c+d)$, or $a+((b+c)+d)$, or what? Well, we can just repeatedly apply associativity to get that these are all equivalent too! For example, $((a+b)+c)+d=(a+(b+c))+d=a+((b+c)+d)=a+(b+(c+d))$.

In a similar vein, with repeated application we can show that associativity implies that we never need parentheses when adding any number of elements: all the possible orders of evaluation will be the same. (Try establishing this formally – as a hint, induction helps a lot for this sort of argument.) Thus, we don't need

to worry about writing out expressions like $a_1 + a_2 + ... + a_n$ when we have associativity. And the same is true for multiplication, since that is associative as well.

Now, focus on consequences of the commutative property. Again, take addition first. Off the bat, this tells us that it doesn't matter which order we add two elements in: whether we put $a$ or $b$ first, we have $a + b = b + a$. But we just saw that we can add as many elements as we want, without specifying parentheses. So it's natural to ask: if we re-order more than two elements and then add them, would the result be the same? For example, do we have that $a + b + c = c + b + a = a + c + b = b + a + c$? (Remember that these are field elements, not necessarily numbers.) It turns out that yes, we do! For example, $a + b + c = a + (b + c) = a + (c + b) = (c + b) + a = c + b + a$.

Similarly, by repeated application of commutativity and associativity, we can show that in general we can reorder as many elements as we'd like without changing the result. (Again, try to establish this formally; as before, induction can help.)

These two consequences yield that a lot of the algebra we do for real numbers carries over to general fields too. For example, just like in real-number algebra, we can collect like terms: we have $a+b+a+c+b = a+a+b+b+c$, $abacb = aabbc$, and so on. Furthermore, just like for numbers, we can take multiples and powers of general field elements: we can write down something like $10x + 3y$ or $x^{10}y^3$ and know that this is equal to any re-ordering of 10 $x$'s and 3 $y$'s added (or multiplied) together, in any order of evaluation. Thus, for any $x \in F$ and any positive integer $n$, it makes sense to write $nx$ and $x^n$.

A natural question then is: what about negative multiples and powers? Let's focus on the multiples first. If an integer is negative, so of the form $-n$ for $n$ a positive integer, then there are multiple "reasonable" ways we could define $(-n)x$: it could be $-(nx)$ (the additive inverse of $nx$), or $n(-x)$ (the $n$th multiple of $-x$.) Which of these should we pick? Hopefully, it would be nice if these turned to be equal to each other, then we wouldn't have to pick between two "reasonable-sounding" choices. Also, it seems that we've had some luck so far with such "reasonable" equivalencies holding true for general fields, so maybe the same

would be true here as well.

So let's try to prove that they must be equal. Take $n = 2$ as a "guinea pig." We want to show that $(-x)+(-x) = -(x+x)$. Well, rearranging, this is equivalent to $(-x) + (-x) + x + x = 0$, and since we can reorder terms, this is the same as $(-x) + x + (-x) + x = 0$. But we know that $-x + x = 0$, so we then have $(-x) + x + (-x) + x = 0 + 0 = 0$, as desired. Similarly, we can show in general that $-(nx) = n(-x)$. (As an exercise, prove this formally.)

In a similar way, for multiplication, we have for any positive integer $n$ that $(x^{-1})^n = (x^n)^{-1}$ (again, as an exercise, prove this formally), so we can define $x^{-n}$ unambiguously to be either of these two.

If we have positive integer and negative integer multiples/powers, what about the zero multiple/power? Again, look at multiples first. What would $0x$ be? We have to be careful: we already have an element of the field that we're calling 0 as well. To avoid confusion, let's label these $0_{\mathbb{Z}}$ and $0_F$. We know that $0_F x = 0_F$, and we are asking what $0_{\mathbb{Z}} x$ should be – i.e., what to define it as. We can get rid of the possible confusion around 0 by just defining these to be equal: we define $0_{\mathbb{Z}} x = 0_F$ for all $x$. Then, we can unambiguously write $0x = 0$.

What about the zeroth power? Looking at numbers, since $x^0 = 1$ for any nonzero real number $x$, we can analogously define $x^0 = 1$ for any nonzero $x$ on a field.

So now that we have multiples and powers, do standard results about multiples and powers of numbers carry over to fields? For example, is it true that for integers $m, n$ and field elements $x, y$, we have $mx + nx = (m + n)x$, or $x^{m+n} = x^m x^n$? As an exercise, prove those two as well as the following also-familiar properties:

$$(mn)x = m(nx),$$

$$x^{mn} = (x^m)^n,$$

$$(nx)y = n(xy).$$

Remember the impact of our theory: though it looks familiar since we've already seen it for numbers, these results apply to *any* field. That includes many fields that look different from $\mathbb{R}$, like $\mathbb{Z}_2$, $\mathbb{Z}_3$, and many others that we will encounter later on.

What are some other things that we typically do with numbers? Well, one thing is that we often expand products of sums, like

$$(a+b)(c+d) = ac + ad + bc + bd,$$

$$(a+b)(c+d+f) = ac + ad + af + bc + bd + bf,$$

and so on. Are these true for fields as well?

Well, how did we derive them for numbers? It was really just repeatedly applying distributivity! For example, we have

$$\begin{aligned}(a+b)(c+d) &= a(c+d) + b(c+d) \\ &= (ac+ad) + (bc+bd) \\ &= ac + ad + bc + bd.\end{aligned}$$

Well, mainly distributivity, but in conjunction with some other properties like associativity. But these properties hold for fields too (by definition), so these expansions must hold as well for fields. In general, we have that if we multiply two sums $A$ and $B$, then the result will be the product of each term in $A$ with each term in $B$, all added together:

$$\sum_{i=1}^{m} a_i \sum_{j=1}^{n} b_j = \sum_{1 \le i \le m, 1 \le j \le n} a_i b_j.$$

(If you are unsure, prove this formally.)

We can obtain similar expansions for multiplying three, or four, or any number of sums together, on a general field.

A particular consequence of all this is the *binomial theorem*, which hopefully is familiar to you from $\mathbb{R}$. As a recap, this is a result that specifies how to expand $(a+b)^2$, $(a+b)^3$, and so on – in general, $(a+b)^n$. It states that

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}.$$

By what we have just discussed, this must hold generally on all fields.

Remember again that our work here is applicable beyond $\mathbb{R}$. As an example, we know that $\mathbb{Z}_3$ is a field; thus, the binomial theorem must hold on $\mathbb{Z}_3$. Furthermore, each binomial coefficient $\binom{n}{i}$ is divisible by $n$ except for $i = 0$ and $i = n$. Thus, if $n$ is a multiple of 3, then on $\mathbb{Z}_3$, we have that each term corresponding to $i$ other than 0 or $n$ vanishes ($3x = 0$ for all $x \in \mathbb{Z}_3$), yielding just

$$(a+b)^n = a^n + b^n.$$

This is a very "simple" and "nice" expansion that holds on $\mathbb{Z}_3$, but not on say $\mathbb{R}$. (This is sometimes jokingly called the "freshman's dream.")

Constantly keep in mind the generality and power of the work we are doing here, especially as we progress and continue to take intuition from familiar algebras like $\mathbb{R}$.

## 2.2. Polynomials Over Fields

As we previously discussed, the commutative and associative properties allow us to collect like terms and like factors, so that for example $x + y + x = 2x + y$ and $xyx = x^2y$. Now, consider any expression involving say just addition, like $x + y + x$. Say it can include both variables, like in this case $x$ and $y$, and

constants, like for example specific elements $c$ and $d$ of the field. Take for example $x + y + c + x + d$. This can be rewritten as $2x + y + c + d$. In fact, any such expression will eventually reduce to a sum of a multiple of $x$, a multiple of $y$, and a constant: something of the form $mx + ny + c$, where $m, n \in \mathbb{Z}_+$.

What if the expression involves just multiplication? Similarly, it will reduce to a product of a power of $x$, a power of $y$, and a constant: something of the form $x^m y^n c$, where $m, n \in \mathbb{Z}_+$. Analogously to real algebra, we typically write this as $cx^m y^n$ instead.

Now, what if the expression involves both addition and multiplication? Well, if we expand all the products of sums, it will reduce to just a sum of terms of the previous form, $cx^m y^n$. (If you're unsure, try out some examples to see this. As an exercise, write out a formal proof.) A similar result holds for any number of variables: any expression in variables $x_1, ..., x_n$ involving just addition and multiplication will reduce to a sum of terms of the form $cx_1^{e_1}...x_n^{e_n}$.

But in real number algebra, this is exactly what a polynomial is!

If we define polynomials on fields analogously to real number algebra (as a sum of terms of that form), then we have that any expression involving just addition and multiplication reduces to a polynomial. Polynomials hold a lot of importance in field theory, and in this section we'll begin studying them for general fields.

As with $\mathbb{R}$, we'll concentrate on polynomials in a single variable, for simplicity. Thus, we consider expressions of the form

$$a_n x^n + ... + a_0,$$

where $a_i \in F$ for $i \in [0, n]$.

We can recover much of our work with polynomials over $\mathbb{R}$ for polynomials over fields. However, we first must discuss an important subtlety.

Consider the field $\mathbb{Z}_2$. The polynomial expressions $0x^2 + 1x + 0 = x$ and $1x^2 + 0x + 0 = x^2$ have distinct coefficients. However, the corresponding *functions*, $p(x) = x, q(x) = x^2$, produce the same outputs given an input:

$p(0) = q(0) = 0, p(1) = q(1) = 1$. Thus, these functions are equivalent: the polynomial expression is just a mechanism to compute an output value, and functions are just mappings of inputs to outputs, irrespective of what the "journey" from input to output is. Thus, in general we need to distinguish between a polynomial defined by the coefficients and a polynomial function defined by the resulting input-output mapping. The polynomial expressions $x$ and $x^2$ are distinct, but the corresponding polynomial functions are equivalent.

We often denote polynomials – not polynomial functions – with an expression using a capital $X$, like $p = X, q = X^2$. The $X$ is not a variable for a function here, but rather a "notational placeholder" so that we can write down the polynomial expression. Then, the corresponding function would use a different letter for the function variable. The $X$ is called an *indeterminate*.

(Really, all the "data" for the polynomial is contained in the coefficients. The indeterminate doesn't "add any more data" – it is just a notational device.)

We define *degree* and other polynomial-related terms for fields analogously to real polynomials. For a field $F$, we denote the set of polynomials over $F$ by $F[X]$. To elucidate results for $F[X]$ that are similar to those for the set of real polynomials $\mathbb{R}[X]$, we can add, subtract, and multiply polynomials in $F[X]$ just like we do polynomials in $\mathbb{R}[X]$. For example, if the polynomials have the same degree, $p, q \in \mathbb{F}[X]$ with $p = p_0 + p_1X... + p_nX^n, q = q_0 + q_1X... + q_nX^n$, then we have

$$p + q = \sum_{k=0}^{n} (p_k + q_k)X^k,$$

$$pq = \sum_{k=0}^{2n} \left( \sum_{i=0}^{k} p_i q_{k-i} \right) X^k.$$

If the degrees of $p$ and $q$ are different, then we use these formulas with 0's in place of the missing corresponding coefficients.

With these operations, we can do much of the same polynomial algebra that we do with $\mathbb{R}[X]$. For example, we have:

**Theorem 2.2.1.** The operations on $F[X]$ satisfy:

1. Commutativity (of both addition and multiplication)
2. Associativity (of both addition and multiplication)
3. Existence and uniqueness of identity (for both addition and multiplication)
4. Existence and uniqueness of *additive* inverse

You can write out the proofs yourself to check these properties.

This shows that $F[X]$ itself is "almost" a field. For example, in $F[X]$, we can rearrange terms in expressions and cancel common terms in equations. However, note that multiplicative inverses do not necessarily exist for polynomials. Indeed, the multiplicative inverse of the real polynomial function $p(x) = x$ is $f(x) = 1/x$, which does not correspond to a real polynomial. Thus, in general, $F[X]$ is not a field.

Despite this, $F[X]$ has an important additional property. Throughout the rest of this section, assume the polynomial $p$ is nonzero. We have:

**Theorem 2.2.2.** $F[X]$ supports polynomial "long division," with a version of the Euclidean division algorithm (see the appendix.) Specifically, for any $p, b \in F[X]$ with $b \neq 0$, there exist $q, r \in F[X]$ such that

$$p = qb + r$$

with $\deg r < \deg b$ if $\deg b > 0$ and $r = 0$ if $b$ is constant. Furthermore, given $p$ and $b$, $q$ and $r$ are unique.

*Proof.* First, we show existence. We are motivated by the procedure of polynomial "long" division we use for $\mathbb{R}[X]$. The first step in such a procedure is to compare the leading coefficients of $p$ and $b$; we then subtract a multiple of $b$ from $p$ so that the degree is smaller, and we repeat the process. We can turn this into a formal proof by contradiction, by the technique of "infinite descent": we

assume for the sake of contradiction that we have a minimal counterexample to our hypothesis, then we come up with an even smaller counterexample, which contradicts the original minimality. (Thus, the counterexamples "keep descending infinitely," which would be impossible if the "size" of the example is always say a nonnegative integer.) Here, the degree will keep decreasing, so that is what we should use for minimality. Let's write out the details:

Fix a nonzero polynomial $b$. Assume for the sake of contradiction that there existed a polynomial $p$ that didn't satisfy the property with $b$. We may choose $p$ so that its degree is smallest. Denote the degrees of $p, b$ by $m, n$, respectively, and let

$$p' = p - \left(\frac{p_m}{b_n}\right) b.$$

Now, if $p'$ satisfied the property (with $b$), then there would exist $q, r$ with $\deg r < \deg b$ such that

$$p' = qb + r.$$

But then

$$p = p' + \left(\frac{p_m}{b_n}\right) b = \left(\frac{p_m}{b_n} + q\right) b + r,$$

even though $p$ doesn't satisfy the property, contradiction. Thus, $p'$ can't satisfy the property either, but $\deg p' < \deg p$, yielding our desired contradiction.

This establishes existence. The next part is uniqueness. As a challenge, write out the details of this yourself, before reading further. (As a hint, try proof by contradiction again.)

To show uniqueness, assume for the sake of contradiction that there were distinct $(q, r), (q', r')$ with $p = qb + r = q'b + r'$, so that

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

This polynomial cannot be nonzero, since if it were nonzero, $\deg(r' - r) \leq \max(\deg r, \deg r') < \deg b \leq \deg((q - q')b)$. Thus, $r' = r$. Since $b \neq 0$, it follows that $q = q'$. □

This support for Euclidean division in $F[X]$ is immensely useful. For example, we can prove a generalization of the Factor Theorem, which is again familiar from $\mathbb{R}[X]$. Recall that this relates a polynomial's factors to its roots (also called zeroes) – specifically, for a real polynomial $p(x)$, it has a real number $a$ as a root if and only if it has $x - a$ as a factor. Now, for general fields, we need to keep in mind that the polynomial function is what has zeroes, while the polynomial is what has factors. The generalized Factor Theorem then relates information about the polynomial to information about its corresponding function, allowing us a way to move between these concepts. Specifically, we have the following definition and theorem:

**Definition 2.2.3.** Let $p, b \in F[X]$. We say $b$ is a *factor* of $p$ if there exists $d \in F[X]$ such that $p = db$.

**Theorem 2.2.4.** Let $p \in F[X]$, and let $p(x) : F \to F$ be the corresponding function. The polynomial $X - k$ is a factor of $p$ if and only if $p(k) = 0$. (In this case, $k$ is called a zero of $p$.)

As an exercise, show this yourself using Euclidean division. (As a hint, divide by $X - k$.)

A consequence is:

**Corollary 2.2.5.** If $p \in F[X]$ has degree $n$, then the corresponding function $p(x)$ has at most $n$ zeros.

*Proof.* If $p$ had $n + 1$ zeros $a_1, ..., a_{n+1}$, then each $X - a_i$ would be a factor of $p$. This means that

$$\prod_{i=1}^{n+1} (X - a_i)$$

is a factor of $p$, so $p$ has degree at least $n + 1$, contradiction. □

(In the future, we will automatically use function concepts for polynomials directly, with the implicit convention that these apply to the corresponding function. For example, we may speak about the zeros of a polynomial $p \in F[X]$.)

Next, in $\mathbb{R}$, you may recall that given two values of a linear polynomial we can completely determine its form, and given three values of a quadratic polynomial we can completely determine its form. Similarly, for any polynomial of degree $n$, it is uniquely determined by $n+1$ values. We can even write out an expression for what the polynomial must be, given the values we know. The same is true for general fields:

**Theorem 2.2.6.** If $p \in F[X]$ and $a_1, ..., a_{n+1}, b_1, ..., b_{n+1} \in F$ with the $a_i$ distinct and $p(a_i) = b_i$, then

$$p = \sum_{i=1}^{n+1} b_i \prod_{j=1, j\neq i}^{n+1} \frac{X - a_j}{a_i - a_j}.$$

*Proof.* Let the right-hand side be the polynomial $q$, and let $d = p - q$. If $d$ is nonzero, then each $a_i$ is a zero of $d$. Thus, $\deg d \geq n + 1$, but since $\deg p = \deg q = n$, $\deg d \leq n$, contradiction. □

(This formula is called *Lagrange interpolation.*)

We were careful to distinguish in general between polynomials and polynomial functions. However, the fact that polynomials are uniquely determined by $n+1$ values implies that these notions coincide for infinite fields: if $F$ is infinite, then if two polynomial functions are equal their corresponding polynomials are as well.